HANDLING MISSING DATA WITH SUPERPOPULATION MODELS, DESIGN BASED APPROACH AND MACHINE LEARNING.

Brigitte Gelein

N° d'ordre : 2017-20
N° de série : G-16

THESE / AGROCAMPUS OUEST

Sous le label de l'Université Européenne de Bretagne

pour obtenir le diplôme de :

DOCTEUR DE L'INSTITUT SUPERIEUR DES SCIENCES AGRONOMIQUES, AGRO-ALIMENTAIRES, HORTICOLES ET DU PAYSAGE

Spécialité : Mathématiques et leurs interactions

Ecole Doctorale : MathSTIC.

IRMAR UMR CNRS 6625.

présentée par :

Brigitte GELEIN

HANDLING MISSING DATA WITH SUPERPOPULATION MODELS, DESIGN BASED APPROACH AND MACHINE LEARNING.

soutenue le **24 octobre 2017** devant la commission d'Examen :

Composition du jury :

Hervé CARDOT - *Professeur, Université de Bourgogne*	Rapporteur
Nikos TZAVIDIS - *Professeur, Université de Southampton*	Rapporteur
Lise BELLANGER - *Maître de conférence, HDR, Université de Nantes*	Membre
Olivier SAUTORY - *Administrateur, méthodologue, Insee*	Membre
David CAUSEUR - *Professeur, Agrocampus Ouest*	Directeur de thèse
David HAZIZA - *Professeur, Université de Montréal*	Directeur de thèse

Remerciements

First of all, I would like to express my great thanks to Hervé Cardot and Nikos Tzavidis for accepting to report on my thesis, as well as Lise Belanger and Olivier Sautory for their kind participation as examiners in the PhD defense, despite their overburden work schedule.

J'adresse également de profonds remerciements à vous, "mes deux Davids", David Causeur et David Haziza. Votre soutien sans faille, la richesse de nos échanges, votre patience et votre amitié ont évidemment été des éléments cruciaux dans l'aboutissement de cette thèse. Ces six ans de thèse à temps partiel (voire très partiel sur certaines périodes) ont été comme une longue course d'endurance.

Je tiens à exprimer toute ma reconnaissance à mes collègues des services administratifs, logistiques, informatiques et de la scolarité de l'Ensai. Leur implication et leur dévouement m'ont permis de bénéficier d'un contexte favorable à l'accomplissement de mon travail d'enseignement et de ma responsabilité de la filière de troisième année *Statistique pour les Sciences de la vie*. Evoquer cette filière ne peut se faire sans une pensée amicale et un peu nostalgique que j'adresse avec beaucoup d'affection à tous ses élèves passionnés et passionnants. "Biostat un jour, biostat toujours" pour reprendre un de leurs slogans. Spéciale dédicace à Déléguée, EmmaCarena, Hiboux, Marie Bofbof, Rafiki, Valou le Bon Goût... désolée je n'ai pas retenu tous les surnoms mais je pense à tous les autres aussi :) ! ! !

Merci aussi pour leur soutien à Aurélie et Hervé qui nous ont fait découvrir le MölkkyPong, Laurence et Thierry pour leur générosité et leur hospitalité, Magalie pour sa très précieuse amitié, François (C) pour nos échanges statistiques et cinématographiques, François et Laurence (P) pour leur superbe investissement en faveur des orphelinats au Burkina, Yann pour sa bonne humeur et ses talents informatiques et artistiques, Marcel pour nos premiers pas en apiculture, Carole, Farah et Sandra pour les sorties entre filles, à Jocelyn Le Grand JJ champion de ping pong (pas si bon au billard... peut-être à cause de Denys), à mes anciens collègues Denys, Pierre(s), ... merci à tous ceux qui m'ont accordé du temps, de l'amitié et ces sourires du quotidien qui font du bien.

Un grand merci aussi à tous ceux qui nous ont apporté leur soutien dans les

moments très difficiles : Jacques et Véronique ainsi qu'Antoine, Floran, Guillaume, Mathieu, Pierre-Edouard, Yannick et tous les autres - et ils ont été nombreux :) ! Merci aussi aux syndicats CGT-SUD et FSU qui ont essayé de nous aider.

Pour finir de façon plus intime, j'adresse de tendres pensées à Nora et Guillaume, à ma famille en Poitou-Charentes et Rhône-Alpes. Je vous aime.

Contents

List of Figures

List of Tables

Chapitre 0

Introduction (version française)

Les instituts nationaux de statistique tels que l'Insee en France, Statistique Canada ou encore Eurostat au niveau international ont pour but la mise à disposition d'informations fiables pour une aide à la décision, avec des bases solides et destinée aux représentants élus, aux entreprises, aux syndicats, aux associations ainsi qu'aux citoyens. Afin de mieux comprendre la démographie, la sociologie et l'économie, les analystes et les chercheurs mettent en œuvre des méthodes statistiques pour analyser les données. Ces dernières peuvent être fournies par des recensements, des enquêtes ou encore des sources administratives. Peu importe l'origine de ces données, elles sont toutes susceptibles de présenter des données manquantes.

La théorie des sondages rencontre de nouveaux champs d'application en relation avec les méthodes d'apprentissage ou Machine Learning, ainsi que les données massives ou Big Data. Le sujet principal de cette thèse est le *traitement des données manquantes y compris par les méthodes d'apprentissage*. Le traitement de la non-réponse est d'un intérêt pratique très important étant donnée la baisse constante du taux de réponse aux enquêtes depuis plusieurs décennies.

Dans le domaine des enquêtes, les données collectées sont utilisées pour estimer des paramètres dits de population finie, pour décrire certains aspects de la population étudiée (comme un total, un coefficient de corrélation ou encore une fonction de distribution). Un certain nombre de procédures d'estimation peuvent

être mises en œuvre pour estimer ces paramètres d'intérêts. Certaines d'entre elles recourent à de l'*information auxiliaire*, contenue dans un ensemble de variables disponibles pour toutes les unités de l'échantillon interrogé et dont les totaux sur la population sont disponibles grâce à d'autres sources telles que les recensements ou les sources administratives. A condition d'être disponible pour toute la population avant échantillonnage, l'information auxiliaire peut également être utilisée à l'étape de l'élaboration du *plan de sondage*, dans le but d'obtenir des estimateurs plus efficaces. Enfin, l'information auxiliaire peut également fournir des éléments pour réduire des erreurs liées au problème de *couverture* et de non réponse.

En statistique d'enquête, on distingue la *non réponse totale* de la *non réponse partielle*. La première a lieu lorsqu'aucune information n'est utilisable pour une unité de l'échantillon alors que la seconde correspond au cas où seules quelques variables d'intérêt sont renseignées. La non réponse peut affecter la qualité des estimateurs quand répondants et non répondants présentent des caractéristiques différentes au regard des variables d'intérêt. Les trois principaux effets de la non réponse sont : (i) biais des estimateurs ponctuels, (ii) augmentation de la variance de ces estimateurs (en raison de la diminution de la taille de l'échantillon par rapport à la taille initialement prévue), et biais des estimateurs de variance sur les cas complets (Haziza, 2009). La non réponse totale est généralement traitée par des procédures d'ajustement de *poids* (Groves et al., 2001, Särndal et Lundström, 2005). En revanche, la non réponse partielle est plutôt traitée par des méthodes d'*imputation*(Brick et Kalton 1996). Ces deux approches (pondération et imputation) partagent le même objectif : réduire le biais de non réponse et, si possible, limiter la variance de non réponse.

Dans un contexte de non réponse partielle et donc d'imputation, on distingue l'*imputation simple* de l'*imputation multiple*. L'imputation simple consiste à remplacer une valeur manquante par une seule valeur artificielle. Un grand nombre de méthodes d'imputation sont basées sur ce principe : notamment l'imputation par la régression (dont l'imputation par le ratio et l'imputation par la moyenne consti-

tuent des cas particuliers), l'imputation par les plus proches voisins, l'imputation aléatoire par hot-deck et l'imputation historique (Haziza, 2009). Avec l'imputation simple, un seul tableau de données (tableau imputé) est produit. C'est sur ce tableau que les chercheurs et chargés d'études pourront appliquer des procédures classiques d'estimation sur données complètes pour calculer des estimateurs ponctuels, sans recourir aux indicateurs de réponse. Bien que l'imputation multiple soit utilisée dans un grand nombre d'applications, elle peut conduire à des conclusions erronées en termes d'inférence (Kim et al., 2006).

En ce qui concerne les procédures d'ajustement des poids pour traiter le problème de la non réponse totale, on distingue deux catégories (Särndal, 2007, et Haziza and Lesage, 2016). Dans la première, les poids de base sont multipliés par l'inverse de la probabilité de réponse estimée. Dans la seconde catégorie de méthodes, on ajuste les poids de base par une forme de calage dont la post-stratification et l'estimateur par le ratio constituent des cas particuliers. Dans ce travail de thèse, on utilise la première approche. De façon à se prémunir contre une éventuelle mauvaise spécification du modèle de prédiction des probabilités de réponse, il est courant de construire des classes de pondération (appelées groupes homogènes de réponse). Au sein de chacune des classes d'unités de l'échantillon, on affecte la même probabilité de réponse estimée (Little, 1986, Eltinge et Yansaneh, 1997, Haziza et Beaumont, 2007).

Ce travail de thèse est organisé de la façon suivante. Le chapitre 2 est consacré aux éléments de base de théorie des sondages nécessaires à la compréhension des chapitres suivants.

Dans le chapitre 3, nous nous plaçons dans un contexte de non réponse partielle. Nous proposons une nouvelle méthode d'imputation préservant la corrélation entre les variables d'intérêt. En effet, l'imputation marginale qui consiste à traiter séparément chaque variable nécessitant de l'imputation, conduit généralement à des estimateurs biaisés pour les paramètres mesurant les relations entre va-

riables. C'est le cas par exemple du coefficient de corrélation linéaire de Pearson. De façon à résoudre ce problème, deux approches principales ont été explorées dans la littérature. La première consiste à mettre en œuvre une procédure d'imputation marginale suivie d'une procédure de correction du biais à l'étape de l'estimation. Cette approche a été étudiée notamment par Skinner et Rao (2002) ainsi que Chauvet et Haziza (2012). Dans la seconde approche, les valeurs manquantes sont imputées conjointement, avec prise en compte des relations entre les variables d'intérêt. Shao et Wang (2002) ont proposé une procédure d'imputation conjointe par régression aléatoire. Ils ont montré que cette procédure conduit à des estimateurs asymptotiquement non biaisés pour les coefficients de corrélation. On peut également se référer à Chauvet et Haziza (2012) pour une version pleinement efficace (fully efficient) de la procédure de Shao et Wang. Dans le chapitre 3 de ce travail de thèse, on propose une méthode d'imputation en deux étapes. La première étape consiste à obtenir des valeurs imputées initiales par la méthode de Shao et Wang. Ensuite les valeurs initiales sont modifiées de façon à respecter des contraintes de calages. Les valeurs utilisées pour le calage correspondent aux estimateurs MIVQUE des paramètres de modèle (Causeur, 2006). Lorsque la distribution bivariée des variables à imputer est symétrique ou faiblement asymétrique, la procédure que nous proposons s'avère significativement plus efficace que la procédure de Shao et Wang en termes d'erreur quadratique moyenne. Les résultats par simulations confirment ces résultats. Ce travail a été publié dans Journal of MultiVariate Analysis (Gelein et al., 2014).

Dans le chapitre 4, nous considérons le problème d'imputation de variables d'intérêt qui présentent un grand nombre de valeurs nulles. Basées sur un modèle de régression sur données comportant beaucoup de valeurs nulles, Haziza et al. (2014) ont proposé des procédures d'imputation conduisant à des estimateurs doublement robustes de la moyenne de la population finie. En effet, ils obtiennent un estimateur imputé de la moyenne consistant si l'une ou l'autre des deux conditions suivantes est respectée : soit la variable d'intérêt, soit le mécanisme de non réponse

est correctement modélisé. Cependant, ces méthodes ne sont pas nécessairement appropriées quand on souhaite estimer des paramètres plus complexes tels que la fonction de répartition en population finie. Dans ce chapitre, nous proposons donc deux procédures d'imputation qui préservent la fonction de répartition contrairement aux méthodes présentées par Haziza et al. (2014). Les résultats d'une étude par simulation illustrent les bonnes performances des méthodes que nous proposons en termes de biais et d'erreur quadratique moyenne. Ce travail a été soumis à une revue avec comité de relecture.

Au chapitre 5, nous considérons le problème de l'estimation des probabilités de réponse dans un contexte de pondération pour correction de la non réponse totale. Les probabilités de réponse peuvent être estimées par des méthodes paramétriques ou non paramétriques. La classe des modèles paramétriques inclut la régression logistique comme cas particulier. Les méthodes paramétriques présentent cependant plusieurs inconvénients : (i) elles ne sont pas robustes par rapport à une mauvaise spécification de la forme du modèle, (ii) elles ne sont pas non plus robustes à la non prise en compte d'éventuelles interactions entre prédicteurs ou de termes quadratiques, (iii) elles peuvent conduire à des probabilités estimées très proches de zéro, conduisant à des estimateurs potentiellement instables (Little et Vartivarian, 2005, et Beaumont 2005). En pratique, les méthodes non paramétriques sont généralement préférées car, contrairement aux méthodes paramétriques, elles protègent des risques de mauvaise spécification du modèle de non réponse. La classe des méthodes non paramétriques comprend la régression par noyaux (Giommi, 1984, Da Silva et Opsomer, 2006), la régression par polynômes locaux (Da Silva et Opsomer, 2009), la pondération de classes formées sur la base d'une estimation préliminaire des probabilités de réponse (Little, 1986, Eltinge et Yansaneh, 1997, Haziza et Beaumont, 2007), l'algorithme CHi square Automatic Interaction Detection (CHAID de Kass, 1980), Classification and Regression Trees (CART Breiman et al., 1984, Phipps et Toth, 2012), Conditional inference trees (Ctree) pour des cibles simples ou multiples (Hothorn et al. 2006).

Dans ce chapitre, nous faisons une vaste étude par simulation pour comparer un grand nombre de méthodes d'estimation des probabilités de réponse par apprentissage supervisé, dans un cadre de population finie. Dans ces simulations, nous couvrons un large champ de méthodes paramétriques ou non, avec des règles de décisions simples ou agrégées telles que Bagging, Random Forests (Breiman, 1996), Boosting (Freund et Shapire, 1996, Friedman et al. 2000) ; voir également Hastie et al. (2009) pour une revue très complète des méthodes d'apprentissage. Pour chaque méthode, ce sont les performances de l'estimateur par expansion et de l'estimateurs de Hajek d'un total qui sont mesurées en termes de biais relatif et d'efficacité relative.

Chapter 1

Introduction (english version)

National statistical offices like Insee in France, Statistics Canada or Eurostat at an international level, aim at providing solid foundations for good informed decisions by elected representatives, firms, unions, non-profit organizations, as well as individual citizens. In order to better understand demography, society and economy, analysts and researchers implement statistic methods to analyse data. The latter can be provided by censuses, surveys and administrative sources. Regardless of the type of data, it is virtually certain one will face the problem of missing values. Survey sampling theory meets new fields of research in association with machine learning and big data handling. The main topic of this PhD work is *how to deal with missing values*. This is an important practical topic given that response rates in surveys have been steadily declining in the past decades.

In surveys, the collected data are typically used to estimate finite population parameters, which are those describing some aspects of the finite population under study (e.g., population totals and population means). A number of estimation procedures can be used to estimate finite population parameters. Some procedures make use of auxiliary information, which is a set of variables available for all the sample units and whose population totals (e.g, census counts) is available from an external source (e.g., census or administrative data). Provided it is available for all the population units prior to sampling, auxiliary information can also be used at the design stage to improve the efficiency of the sampling designs, leading

to more efficient estimators. Finally, auxiliary information may be used to reduce nonsampling errors such as nonresponse and coverage errors.

Surveys statisticians distinguish unit nonresponse from item nonresponse. The former occurs when no usable information is available on a sample unit, whereas the latter occurs when some variables (but not all) are recorded. Nonresponse may affect the quality of the estimates when the respondents and the nonrespondents exhibit different characteristics with respect to the survey variables. The main effects of nonresponse consist in: (i) bias of point estimators, (ii) increase of the variance of point estimators (due to the fact that the observed sample has a smaller size than the one initially planned), and (iii) bias of the complete data variance estimators (Haziza, 2009). Unit nonresponse is usually handled through weight adjustment procedures (Groves et al. 2001, and Särndal and Lundström 2005), whereas item nonresponse is treated by some form of imputation (Brick and Kalton 1996). These approaches (weight adjustment or imputation) share the same goals: reduce the nonresponse bias and, possibly, control the nonresponse variance.

In the context of imputation for item nonresponse, it is customary to distinguish single from multiple imputation (Rubin, 1987; Little and Rubin, 2002). Single imputation consists of replacing a missing value with a single artificial value. A number of imputation procedures are used in practice: Regression imputation (that includes ratio imputation and mean imputation as special cases), nearest-neighbour imputation, random hot-deck imputation and historical imputation, among others (Haziza, 2009). With single imputation, a single completed data set (also called an imputed data set) is produced, making it possible for secondary analysts to apply complete data estimation procedures for computing point estimates. That is, the latter can be readily obtained using complete data estimation procedures without requiring the response indicators (or response flags). Although multiple imputation is widely used in a number of fields for handling

missing data, it may lead to invalid inferences in finite population sampling; see Kim et al. (2006).

Turning to weighting adjustment procedures for handling unit non-response, two types of weighting procedures are commonly used (e.g., Särndal, 2007 and Haziza and Lesage, 2016): in the first, the basic weights are multiplied by the inverse of the estimated response probabilities, whereas the second uses some form of calibration, that includes post-stratification and raking as special cases, for adjusting the basic weights. In this PhD work, we focus on weight adjustment by the inverse of the estimated response probabilities. To protect against a possible model misspecification, it is customary to form weighting classes (also called response homogeneous groups) so that within a class the sample units have similar response probabilities (Little, 1986, Eltinge and Yansaneh, 1997 and Haziza and Beaumont, 2007).

This PhD thesis is organised as follows. In Chapter 2, we describe the theoretical set-up used in the following chapters and define several concepts that will prove useful in this thesis.

In chapter 3, we propose a new imputation method for preserving correlations between survey variables. Marginal imputation, which consists of treating separately each variable requiring imputation, generally leads to biased estimators of parameters (e.g., coefficients of correlation) measuring relationships between variables. To overcome this problem, two main approaches have been studied in the literature: the first consists of using a marginal imputation procedure followed by a bias-adjustment procedure at the estimation stage. This approach was investigated by Skinner and Rao (2002) and Chauvet and Haziza (2012), among others. In the second approach, the missing values are imputed using a joint procedure, which accounts for the relationships between items. Shao and Wang proposed a joint random regression imputation procedure and showed that it leads to asymptotically unbiased estimators of coefficients of correlation; see also Chau-

vet and Haziza (2012) for a fully efficient version of the Shao-Wang procedure. Shao and Wang (2002) proposed a joint imputation procedure and showed that it leads to asymptotically unbiased estimators of coefficients of correlation. In this chapter, we propose a two-step imputation procedure: first, initial imputed values are obtained using the Shao-Wang procedures. Then, the initial values are modified so as to satisfy calibration constraints, which corresponds to MIVQUE estimators of model parameters (Causeur, 2006). When the bivariate distribution of the variables being imputed is symmetric or exhibits a low degree of asymmetry, the proposed procedure is shown to be significantly more efficient than the Shao-Wang procedure in terms of mean square error. Results from a simulation study supports our findings. This work was published in Journal of Multivariate Analysis (Gelein et al., 2014).

In chapter 4, we consider the problem of imputing survey variables exhibiting a large number of zero-valued observations. Based on a zero-inflated regression model, Haziza et al. (2014) proposed imputation procedures that leads to doubly robust estimators of the population mean, in the sense that the imputed estimator of the mean is consistent whether the variable of interest or the non-response mechanism is adequately modeled. However, these methods are not necessarily appropriate when estimating more complex parameters such as the population distribution function. In this chapter, the interest lies in estimating a finite population distribution function. We propose two imputation procedures and show that they preserve the distribution function, unlike the procedures considered in Haziza et al. (2014). Results of a simulation study illustrate the good performance of the proposed methods in terms of bias and mean square error. This work has been submitted to a peer-reviewed journal.

In chapter 5, we consider the problem of estimating the response probabilities in the context of weighting for unit nonresponse. The response probabilities may be estimated using either parametric or nonparametric methods. The

class of parametric models includes logistic regression as a special case. There are several issues associated with the use of a parametric model: (i) they are not robust to the misspecification of the form of the model ; (ii) they are not robust to the non-inclusion of interactions or predictors that account for curvature (e.g., quadratic terms), both of which may not have been detected during model selection; (iii) they may yield very small estimated response probabilities, resulting in very large nonresponse adjustment factors, ultimately leading to potentially unstable estimates; e.g., Little and Vartivarian (2005) and Beaumont (2005). In practice, nonparametric methods are usually preferred because, unlike parametric methods, they protect against the misspecification of the nonresponse model. The class of nonparametric methods include kernel regression (Giommi, 1984, Giommi 1987, Da Silva and Opsomer, 2006), local polynomial regression (Da Silva and Opsomer, 2009), weighting classes formed on the basis of preliminary estimated response probabilities (Little, 1986, Eltinge and Yansaneh, 1997, Haziza and Beaumont, 2007), the CHi square Automatic Interaction Detection (CHAID) algorithm (Kass, 1980), Classification and regression trees (Breiman et al., 1984, Phipps and Toth, 2012), Conditional inference trees (Ctree) for simple and multiple targets trees (Hothorn et al. 2006). In this chapter, we conduct an extensive simulation study to compare methods for estimating the response probabilities in a finite population setting. In our study, we attempted to cover a wide range of (parametric and nonparametric) "simple" methods as well as aggregation methods like Bagging, Random Forests (Breiman, 1996), Boosting (Freund and Shapire, 1996 and Friedman et al. 2000); see also Hastie et al. (2009) for a comprehensive overview of machine learning methods. For each method, we assessed the performance of the propensity score estimator and the Hajek estimator in terms of relative bias and relative efficiency.

References

Beaumont, J. F. (2005), Calibrated imputation in surveys under a quasi-model-assisted approach, Journal of the Royal Statistical Society Series B (Statistical Methodology), 67(3), 445–458.

Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone (1984). Classification And Regression Trees. New York: Chapman and Hall.

Brick, J.M. and Kalton, G. (1996) Handling missing data in survey research. Statistical Methods in Medical Research, 5, 215–238.

Causeur, D. (2006), MIVQUE and Maximum Likelihood Estimation for Multivariate Linear Models with Incomplete Observations, Sankhya Ser A, 68(3), 409–435

Chauvet, G., Haziza, D. (2012), Fully efficient estimation of coefficients of correlation in the presence of imputed survey data, Canadian Journal of Statistics, 40(1).

Da Silva, D.N., J.D. Opsomer (2006). A kernel smoothing method of adjusting for unit non-response in sample surveys. The Canadian Journal of Statistics, 34, 563–579.

Da Silva, D.N., J.D. Opsomer (2009). Nonparametric propensity weighting for survey nonresponse through local polynomial regression. Survey Methodology, 35, 165–176

Eltinge, J.L., Yansaneh, I.S. (1997). Diagnostics for formation of nonresponse adjustment cells, with an application to income nonresponse in the U.S. consumer expenditure survey. Survey Methodology, 23, 33—40

Gelein, B., Haziza, D. and Causeur, D. (2014) Preserving relationships between

variables with MIVQUE based imputation for missing survey data. Journal of Multivariate Analysis. 131, 197-–208.

Giommi, A.(1984). On the estimation of the probability of response in finite population sampling (Italian, Societa Italiana di Statistica, Atti della Riunione Scientifica della Societa Italiana, 32.

Giommi, A. (1987). Nonparametric methods for estimating individual response probabilities. Survey Methodology, 13, 127–134.

Groves, R. M., Dillman, D., Eltinge, J. L., Little,R. J. A. (2001). Survey Nonresponse. Wiley, New York.

Hajek, J. (1971) Comment on An essay on the logical foundations of survey sampling by Basu, D. in Foundations of Statistical Inference (Godambe, V.P. and Sprott, D.A. eds.), p. 236. Holt, Rinehart and Winston.

Hastie, T., Tibshirani, R., Friedman, J. H. (2009). The elements of statistical learning: Data mining, inference, and prediction. New York: Springer.

Haziza, D., Beaumont, J.-F. (2007). On the construction of imputation classes in surveys. International Statistical Review, 75, 25—43.

Haziza, D. (2009). Imputation and inference in the presence of missing data. Handbook of statistics, 29, 215–246.

Haziza, D., Kuromi G. (2007). Handling Item Non Response in Surveys.CS-BIGS 1(2), 102–118.

Haziza, D., Nambeu, C.-O., Chauvet, G. (2014). Doubly robust imputation procedures for finite population means in the presence of a large number of zeroes. Canadian Journal of Statistics, 42, 650–669.

Haziza, D., Lesage, E. (2016). A discussion of weighting procedures for unit nonresponse. Journal of Official Statistics, 32, 129-145.

Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. Applied Statistics 29 (2), 119–127.

Kim, J.K., Brick, J.M., Fuller, W.A. (2006). On the bias of the multiple-imputation variance estimator in survey sampling. Journal of the Royal Statistical Society: Series B, 68, 509—521.

Little, R.J.A. (1986). Survey Nonresponse Adjustments for Estimates of Means. International Statistical Review, 54, 139—157.

Little, R.J.A., Rubin, D.B. (2002). Statistical analysis with missing data. (Second Edition.) New York: Wiley.

Little R. J. A., Vartivarian S. (2005). Does weighting for nonresponse increase the variance of survey means ? , Survey Methodology, 31, 161–168.

Rubin, D.E. (1987). Multiple imputation for nonresponse in surveys. New York: Wiley.

Phipps, P., Toth, D. (2012). Analyzing establishment nonresponse using an interpretable regression tree model with linked administrative data. Annals of Applied Statistics, 6, 772–794.

S*ä*rndal, C.-E. (2007). The calibration approach in survey theory and practice, Survey Methodology, 33 (2), 99–119

S*ä*rndal, C.-E., and Lundstr*ö*m, S. (2005). Estimation in Surveys with Nonresponse. Wiley, New York.

Shao, J., Wang, H. (2002). Sample Correlation Coefficients Based on Survey Data Under Regression Imputation, Journal of the American Statistical Association, 97, 544–552.

Skinner, C.J., Rao, J. N. K. (2002), Jackknife variance estimation for multivariate statistics under hotdeck imputation from common donors, Journal of Statistical Planning and Inference, pp 149-167,

Chapter 2

Theoretical set up

This chapter provides a brief presentation of some concepts that will prove useful in the subsequent chapters. This chapter is based on the following sources: Haziza and Kuromi (2007), Kim (2014), Montaquila (2010), De Leeuw et al. (2008), Pfefferman and Rao (2009), Thompson (2012), Ardilly (2006), Favre-Martinoz (2015) and Tillé (2001). We start by quoting Mahalanobis (1965): "Large scale sample surveys, when conducted in the proper way with a satisfactory survey design, can supply with great speed and at low cost information of sufficient accuracy of practical and with the possibility of ascertainment of the margin of uncertainty on an objective bias". Thus, estimating characteristics of a population requires "satisfactory survey design" and, before that, the definition of the population itself.

2.1 Which population ?

In survey research and other applications, the main purpose is often to estimate the parameters of a finite population rather than the parameters of a statistical model.

2.1.1 Finite population

In order to construct reliable estimates, the finite population must be defined precisely before the implementation of a well designed sampling procedure. The

basic theory and methods of probability sampling from finite populations were significantly developed during the first half of the twentieth century. For instance, the seminal contribution of Neyman (1934) spells out the advantages of probability sampling in comparison to purposive selection.

Definition 2.1. *The units of the finite population U are said identifiable if they can be referred to with a number or a label $U = \{1, ..., j, ..., N\}$.*

An example of finite population is that for the British Columbia Smoking Survey (BCSS), conducted in 2006 to gather information related to the smoking history, mobility history and risk propensity of British Columbia residents. The target population consisted of residents aged 18 and over, living in private occupied dwellings at the time of the Canadian Community Health Survey (CCHS). Some individuals were excluded from the scope of the survey, including those living on Indian Reserves and on Crown Lands, institutional residents, full-time members of the Canadian Armed Forces, as well as residents of remote regions.

Another example is Insee's monthly business outlook survey in the building industry, that records the opinion of entrepreneurs in the sector on recent activity and on their future activity, so as to assess the current situation and forecast activity both at national and European levels. The finite population with a firm as statistical unit covers companies working on the construction of individual houses and miscellaneous buildings, general building work, roofing, framing, fitting and finishing work (and more precisely, the sectors defined by the following NAF Rev. 2 codes: 41.2, 43.2, 43.3, 43.9). Since 2008, the survey takes place every month.

2.1.2 Sources of errors

The concept of total survey error is important for studying the properties of an estimation procedure. The total error of an estimate is the difference between an estimate and the true value. It can be expressed as the sum of four components

(Groves et al., 2004):

Total survey error	= coverage errors	+ sampling errors
	+ measurement errors	+ nonresponse errors

Coverage represents the percentage of the population of interest that is included in the sampling frame. Undercoverage for instance, occurs when segments of the population are missing from the sampling frame. As quoted by Lohr (2008), both undercoverage and nonresponse lead to missing data. This may result in biased estimates if the missing units (either non-respondents or non-covered units) exhibit different characteristics of interest than those which are in the sampling frame and/or respond to the survey. Coverage errors may also be caused by erroneous inclusions in the frame (for instance, a firm erroneously included while its activity has stopped). Lastly, overcoverage corresponds to the case where a unit from the target population appears more than once in the sampling frame, or where some unit which does not belong to the target population appears in the sampling frame.

Measurement errors result from wrong responses to questions or incorrect measurements. For example, in a survey on AIDS, persons with AIDS may say they do not suffer from AIDS while they do, but fear that their illness would be revealed.

Sampling error occurs because measures are taken on a sample instead of the entire population.

As for **nonresponse**, we distinguish unit nonresponse that exists when no usable information is available on a sample unit, from item nonresponse that occurs when some variables (but not all) are recorded (see subsection 2.6).

2.1.3 Superpopulation model

Let $y_1, \cdots, y_q$ denote q survey variables and let $\mathbf{y}_i = (y_{i1}, y_{i2}, ..., y_{iq})^\top$ be the vector of survey variables associated with unit $i \in U$. The vector $\mathbf{y}_i$ may be treated as deterministic, or considered as the realization of a random variable Y whose distribution is specified by some superpopulation model. This means that

the finite population can be seen as coming from a theoretical infinite population. Therefore, we distinguish between two sources of randomness. One of them comes from the random selection of the sample, and is the only source of randomness under the so-called design-based approach (see Section 2.3 for more details). But if we appeal to some superpopulation model, another source of randomness is due to the (random) generation of the values for the variables of interest.

Let $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{ip})^\top$ be the p-vector of *auxiliary variables* attached to unit $i \in U$. In the superpopulation model approach, observed vectors $(\mathbf{x}_i^\top, \mathbf{y}_i^\top)^\top$, $i \in U$, are realizations of i.i.d. random vectors $(\mathbf{X}_i^\top, \mathbf{Y}_i^\top)^\top$, where

$$\mathbf{X}_i = (\mathbf{X}_{i1}, \mathbf{X}_{i2}, \ldots, \mathbf{X}_{ip})^\top \text{ and } \mathbf{Y}_i = (\mathbf{Y}_{i1}, \mathbf{Y}_{i2}, ..., \mathbf{Y}_{iq})^\top. \tag{2.1.1}$$

Expectation and variance under the superpopulation model m are denoted as $E_m(.)$ and $V_m(.)$.

2.2 Finite population parameters

A parameter of interest θ, can be defined as an unknown number (or a vector) that describes the finite population. Since it is unknown, we want to estimate the so called finite population parameter $\theta = \theta(\mathbf{y}_i,\ i \in U)$, which is a function of $\mathbf{y} = (\mathbf{y}_1, ..., \mathbf{y}_i, ..., \mathbf{y}_N)^\top$. This function can be linear in the values of an variable of interest y, such as for the total $t_y = \sum_{i \in U} y_i$ or for the mean $\overline{y} = t_y/N$. If N is unknown, estimating $\overline{y}$ is achieved by estimating separately the numerator t_y and the denominator N. In this PhD work, we also are interested in more complex parameters such as the finite population coefficient of correlation between two variables (see chapter 3) and the finite population distribution function of a survey variable (see chapter 4).

The finite population coefficient of correlation between the variables y_1 and y_2 is defined as:

$$R_{12} = \frac{t^{11} - t^{10}t^{01}/N}{(t^{20} - (t^{10})^2/N)^{1/2}(t^{02} - (t^{01})^2/N)^{1/2}},$$

where $t^{ab} = \sum_{i\in U}(y_{1i})^a(y_{2i})^b$ with $(a,b) \in \{(1,0),(2,0),(1,1),(0,1),(0,2)\}$. For example, $t^{10} = \sum_{i\in U} y_{1i}$ and $t^{11} = \sum_{i\in U} y_{1i}y_{2i}$.

The finite population distribution function of a survey variable y is defined as:

$$F_N(t) = \frac{1}{N}\sum_{i\in U} 1(y_i \leq t) \tag{2.2.1}$$

where $1(\cdot)$ is the usual indicator function.

Totals, coefficients of correlation and distribution functions are some parameters of interest we considered in Chapters 3 to 5, while handling missing values with different methods and approaches.

2.3 Sampling design and inclusion probabilities

Probability sampling methods are widely used to select units which appear in the sample. The randomness coming from the probability design reduces investigator discretion in units' selection. The variance of estimators may be estimated under the sole randomization associated to the sampling design. If the sample selection is not probabilistic, the variance may also be estimated but model assumptions are required. The validity of the variance estimation depends on the validity of the model assumptions.

2.3.1 Sampling design

Given a finite population U, a sampling design specifies for every possible sample its probability of being drawn.

Definition 2.2. *A sampling design without replacement p(.) is a probability distribution on all the non-empty subsets $s \subset U$, such that $\sum_{s\subset U} p(s) = 1$ and $p(s) \geq 0$.*

We note Ω the set of all the subsets $s \subset U$.

2.3.2 Inclusion probabilities

Let $E_p(.)$, $V_p(.)$ and $Cov_p(.)$ denote respectively the expectation, variance and covariance with respect to the sampling design $p(.)$. Let $\boldsymbol{\delta}= (\delta_1, \ldots, \delta_N)^\top$, be the N-vector of selection indicators, such that $\delta_i = 1$ if unit $i \in S$ and $\delta_i = 0$, otherwise. In the design-based approach, the δ_i's are random variables since they depend on the random sample S. On the other hand, the δ_i's are treated as fixed in the model-based approach.

Definition 2.3. *Each unit $i \in U$ is selected according to a first-order inclusion probability in the sample*

$$\pi_i = E_p(\delta_i) = P(i \in S) = \sum_{s \ni i} p(s).$$

The joint inclusion probability of units i and k in S is:

$$\pi_{ik} = E_p(\delta_i \delta_k) = P(i, k \in S) = \sum_{s \ni i,k} p(s) \text{ for all } i, k \in U \ .$$

By convention, we note $\pi_{ii} = \pi_i$.

Definition 2.4. *We define the variance-covariance matrix of inclusion indicators δ_i as $\Delta = (\Delta_{ik})$, with $i = 1, ..., N$ and $k = 1, ..., N$, where*

$$\Delta_{ik} \quad = \quad \begin{cases} Cov_p(\delta_i, \delta_k) = E_p(\delta_i \delta_k) - E_p(\delta_i)E_p(\delta_k) = \pi_{ik} - \pi_i \pi_k & if \ i \neq k, \\ V_p(\delta_i) = E_p(\delta_i^2) - E_p(\delta_i)^2 = \pi_i(1 - \pi_i) & if \ i = k. \end{cases}$$

2.3.3 Examples of sampling designs and related inclusion probabilities

Commonly used probability sampling designs include simple random sampling, systematic sampling, stratified sampling, cluster sampling, probability proportional-to-size sampling and stratified multistage sampling. For all of them, each unit of U has a known nonzero probability of being sampled. Next, we describe simple random sampling without replacement and stratified random Sampling.

Definition 2.5. *Simple random sampling without replacement (SRSWOR) is defined by*

$$p(s) \quad = \quad \begin{cases} \binom{N}{n}^{-1} & \text{if } card(s) = n, \\ 0 & \text{otherwise.} \end{cases}$$

For simple random sampling, the first-order inclusion probability π_i is given by

$$\pi_i = \sum_{s \ni i} p(s) = \sum_{s \ni i} \binom{N}{n}^{-1} = \binom{N-1}{n-1}\binom{N}{n}^{-1} = \frac{n}{N}, \text{ for all } i \in U.$$

The joint inclusion probabilities are given by

$$\pi_{ik} = \sum_{s \ni i,k} p(s) = \sum_{s \ni i,k} \binom{N}{n}^{-1} = \binom{N-2}{n-2}\binom{N}{n}^{-1} = \frac{n(n-1)}{N(N-1)}, \text{ for all } i \neq k \in U.$$

The variance-covariance matrix of $\boldsymbol{\delta} = (\delta_1, ...\delta_N)^\top$ is $\Delta = (\Delta_{ik})$, with $i = 1, ..., N$ and $k = 1, ..., N$, where

$$\Delta_{ik} \quad = \quad \begin{cases} Cov_p(\delta_i\delta_k) = \pi_{ik} - \pi_i\pi_k = -\frac{n}{N(N-1)}(1 - \frac{n}{N}) & \text{if } i \neq k, \\ V_p(\delta_i) = \pi_i(1-\pi_i) = \frac{n}{N}(1 - \frac{n}{N}) & \text{otherwise.} \end{cases}$$

Simple random sampling without replacement is among the simplest probability sampling designs and serves as the basis of more complex sampling designs such as stratified random sampling. According to Lohr (2008), a simple random sample is a good choice for a design if little is known about the population being studied, which is the case if the sampling frame is a mere list of addresses with no additional information, for instance. However, if we can obtain additional information (such as the gender for individuals), it can be used to *stratify the population* to improve the efficiency of survey estimates.

Definition 2.6. In stratified simple random sampling, the finite population U is partitioned into H strata U_h of size N_h such that $U = \bigcup_{h=1}^{H} U_h$ and $U_h \cap U_g = \emptyset$ for $h \neq g$. A SRSWOR of size n_h is selected in each stratum h, $h = 1, \cdots, H$. The selection in a given stratum is independent of the selection in any other stratum.

The inclusion probabilities are given by

$$\pi_i = \frac{n_h}{N_h},\ i \in U_h.$$

The joint inclusion probabilities are given by

$$\pi_{ik} = \begin{cases} \frac{n_h(n_h-1)}{N_h(N_h-1)} & if\ i \in U_h,\ k \in U_h,\ i \neq k, \\ \frac{n_h n_g}{N_h N_g} & if\ i \in U_h,\ k \in U_g,\ h \neq g. \end{cases}$$

The matrix $\Delta = (\Delta_{ik})$, is given by

$$\Delta_{ik} = \begin{cases} \frac{n_h(N_h-n_h)}{N_h^2} & if\ i = k,\ i \in U_h, \\ -\frac{n_h(N_h-n_h)}{N_h^2(N_h-1)} & if\ i \neq k,\ i \text{ and } k \in U_h, \\ 0 & if\ i \in U_h,\ k \in U_g \text{ and } h \neq g. \end{cases}$$

Stratified random sampling exhibits several advantages over simple random sampling provided that auxiliary information is available: (i) it ensures that population subgroups of interest are represented in the sample; (ii) it allows choosing the sample size for each stratum; (iii) it is more efficient than SRSWOR if the strata are homogeneous with respect to the survey variables (e.g., Cochran 1977, Särndal et al. 1992).

As an example of stratified sampling, some description elements of the whole sale trade tendency survey conducted by Insee in France are provided below (table 2.3.1). The Consumer Price Index (CPI) is the instrument used to measure inflation. It allows the estimation of the average variation between two given periods in the prices of products consumed by households.

Table 2.3.1: The consumer price index: an example of stratified sampling

Statistical unit	Retail outlet
Reference period	Monthly
Sampling plan	The sampling plan is stratified according to three types of criteria: *Geographical criterion:* surveys are carried out in 99 conurbations of over 2,000 inhabitants situated throughout metropolitan France and of any size and 10 conurbations in four Overseas Departments (Guadeloupe, Martinique, Guyane, Réunion). *Type of product:* a sample of just over 1,100 product families, called "varieties" is defined in order to take account of the heterogeneity of products within items. The variety is the basic level for tracking products and calculation of the index. The list of varieties remains confidential and the CPI is not disseminated at this level. *Type of retail outlet:* a sample of 30,000 retail outlets, stratified according to the form of sale, has been created in order to represent the diversity of products and purchasing methods used by consumers, and to take account of the price variations, which are differentiated according to the forms of sale. By cross-referencing these different criteria, just over 200,000 series (specific products in a given retail outlet) can be monitored. To these figures can be added approximately 190,000 "pricing" - type series collected in a centralised manner.
Other specifications	The CPI covers all market goods and services consumed throughout the country by resident and non-resident (e.g. tourists) households. Its theoretical scope is defined as the actual final monetary consumption of households. Following major extensions carried out primarily in services, the CPI's *rate of coverage* was 97% in 2016 (2015 base). The main shortcomings concerning coverage still relate to private hospital services and life insurance.

Source: Insee, France

2.4 Properties of estimators in survey sampling theory

Definition 2.7. *Let $\hat{\theta}$ be an estimator of a parameter of interest θ. The design expectation of $\hat{\theta}$ is defined as:*

$$E_p(\hat{\theta}) = \sum_{s \in \Omega} p(s)\hat{\theta}_s,$$

where $\hat{\theta}_s$ is the estimator $\hat{\theta}$ computed from the sample s.

Definition 2.8. *The design bias of an estimator $\hat{\theta}$ is defined as:*

$$B_p(\hat{\theta}) = E_p(\hat{\theta}) - \theta.$$

An estimator $\hat{\theta}$ is design unbiased for θ if and only if

$$E_p(\hat{\theta}) = \theta.$$

Definition 2.9. *The design variance of an estimator $\hat{\theta}$ is defined as:*

$$V_p(\hat{\theta}) = E_p\left[\{\hat{\theta} - E_p(\hat{\theta})\}^2\right].$$

Definition 2.10. *The design mean square error (MSE) of an estimator $\hat{\theta}$ is defined as:*

$$MSE_p(\hat{\theta}) = E_p\left\{(\hat{\theta} - \theta)^2\right\} = V_p(\hat{\theta}) + B_p(\hat{\theta})^2.$$

2.5 Weighted estimators in the absence of non-response

In this section the properties of estimators are examined with respect to the sampling design, whereby the y-values and the $\mathbf{x}$-values are treated as fixed.

2.5.1 Weighting Survey Data

As mentioned by Biemer and Christ (2008), after the survey data have been collected, they must be appropriately weighted before any analysis. The weighting

process leads to the creation of a new variable w_i for each sample unit. If the weight is $w_i = \frac{1}{\pi_i}$, it can then be interpreted as the number of individuals in the target population represented by unit $i \in s$. Ignoring the weights and treating the data as if they were coming from a simple random sample, is equivalent to setting equal weights. It usually results in biased estimates in the case of unequal probability sampling and post-survey weight adjustments.

2.5.2 The Horvitz-Thompson estimator

Basic sampling weights are defined as the inverse of the inclusion probabilities (Horvitz and Thompson, 1952); that is, $w_i = \frac{1}{\pi_i}$, for all $i \in S$.

Definition 2.11. The Horvitz-Thompson estimator of the total t_y is given by:

$$\hat{t}_{y_\pi} = \sum_{i \in S} w_i y_i.$$

This estimator is also called π-estimator or expansion estimator since the values y_i are expanded by the inverse of the inclusion probabilities.

Remark 2.1. In the case of a stratified random sampling without replacement, the Horvitz-Thompson estimator of the total t_y is:

$$\hat{t}_{y_\pi} = \sum_{h=1}^{H} \hat{t}_{y_{\pi h}},$$

where

$$\hat{t}_{y_{\pi h}} = \frac{N_h}{n_h} \sum_{i \in S_h} y_i.$$

Note that the Horvitz-Thompson estimator is generally not location-scale invariant since we have:

$$\frac{1}{N} \sum_{i \in U} \frac{a + b y_i}{\pi_i} \neq a + b \frac{1}{N} \sum_{i \in U} \frac{y_i}{\pi_i}.$$

As explained by Cassel and Wretman (1976), having a probability sampling design with $\pi_i > 0$ for all $i \in U$ is a necessary and sufficient condition for the existence of a design-unbiased estimator of the population total.

Theorem 2.1. *If $\pi_i > 0$, for all $i \in U$, then $\hat{t}_{y_\pi}$ is a design-unbiased estimator of t_y.*

Remark 2.2. If at least one $\pi_i = 0$ for $i \in U$, we are in a situation of *undercoverage* and $\hat{t}_{y_\pi}$ is biased under the sampling design. The bias equals:

$$E_p(\hat{t}_{y_\pi}) - t_y = \sum_{i \in U | \pi_i = 0} y_i.$$

Definition 2.12. *If the finite population size N is known, the Horvitz-Thompson estimator of the mean $\bar{y}$ in U is*

$$\hat{\bar{y}}_\pi = \hat{t}_{y_\pi}/N.$$

The variance of the Horvitz-Thompson estimator of t_y is given by the following theorem:

Theorem 2.2. *The variance of the Horvitz-Thompson estimator of t_y is*

$$V_p(\hat{t}_{y_\pi}) = \sum_{i,k \in U} \frac{y_i y_k}{\pi_i \pi_k} \Delta_{ik}.$$

Furthermore, for a fixed-size sampling design, the variance can be written in the so-called Sen-Yates-Grundy form

$$V_p(\hat{t}_{y_\pi}) = -\frac{1}{2} \sum_{i,k \in U} (\pi_{ik} - \pi_i \pi_k) \left(\frac{y_i}{\pi_i} - \frac{y_k}{\pi_k} \right)^2.$$

Corollary 2.1. *With simple random sampling without replacement, we have*

$$V_p(\hat{t}_{y_\pi}) = N^2 \frac{(1 - \frac{n}{N})}{n} S_y^2,$$

where $S_y^2 = (N-1)^2 \sum_{i \in U} (y_i - \bar{y})^2$ is the dispersion of the population y-values.

With a fixed-size sampling design, the variance of $\hat{t}_{y_\pi}$ equals zero if the inclusion probabilities π_i are proportional to the variable of interest y_i. In practice, we do not know the y-values prior to sampling. If an auxiliary variable x is closely related with y, then a sampling design with $\pi_i \propto x_i$ can lead to very efficient

sampling design. This is the rationale behind *probability proportional to size* (π-ps) sampling designs.

A sampling design is said measurable if $\pi_i > 0$, for all $i \in U$ and $\pi_{ik} > 0$, for all $(i,k) \in U^2$ (Särndal et al. 1992). The measurable property allows the calculations of valid variance estimates.

Theorem 2.3. *For a measurable sampling design, an unbiased estimator of the variance of* $\hat{t}_{y_\pi}$ *is given by:*

$$\hat{V}_{HT}(\hat{t}_{y_\pi}) = \sum_{i \in S} \frac{y_i^2}{\pi_i^2}(1-\pi_i) + \sum_{i,k \in S, i \neq k} \frac{y_i y_k}{\pi_i \pi_k \pi_{ik}} \Delta_{ik}.$$

For fixed-sized designs, an alternative variance estimator is the Sen-Yates-Grundy estimator:

$$\hat{V}_{SYG}(\hat{t}_{y_\pi}) = -\frac{1}{2} \sum_{i,k \in S} \frac{(\pi_{ik} - \pi_i \pi_k)}{\pi_{ik}} (\frac{y_i}{\pi_i} - \frac{y_k}{\pi_k})^2.$$

Corollary 2.2. *With a simple random sample without replacement, the Sen (1953) and Yates-Grundy (1953) variance estimator reduces to*

$$\hat{V}_{SYG}(\hat{t}_{y_\pi}) = N^2 \frac{(1 - \frac{n}{N})}{n} S_{y_S}^2,$$

where $S_{y_S}^2 = \frac{1}{n-1} \sum_{i \in S} (y_i - \bar{y}_S)^2$ *is the dispersion of the y-values in the sample and* $\bar{y}_S = \frac{1}{n} \sum_{i \in S} y_i$ *is the sample mean.*

The Horvitz-Thompson estimator does not necessarily achieve small variance if the π_i's are not proportional to y_i. Thus, to improve the efficiency of the resulting estimator, auxiliary information is often incorporated at the estimation stage: for instance if the population total of an auxiliary variable is known from external sources. That is why we consider the ratio estimator and its special case the Hajek estimator (1971) in the following subsection.

2.5.3 Calibrated estimators

The Horvitz-Thompson estimator $\hat{t}_{y_\pi}$ has the interesting property of being design-unbiased for the total t_y. However, the variance of $\hat{t}_{y_\pi}$ may be large, in particular

if the auxiliary variables used in the sampling design do not explain the y-variable well. On the other hand, it is possible that at the estimation stage some set $\mathbf{x}_k$ of auxiliary variables is available, and that their totals $t_{\mathbf{x}} = \sum_{i \in U} \mathbf{x}_i$ are known from an external source such as census or administrative data. The purpose of calibration (Deville and Särndal, 1992) is to adjust the estimators on this auxiliary information to produce more precise estimates than can be obtained from the y data alone.

The Horvitz-Thompson estimator applied to the set of auxiliary variables $\mathbf{x}_i$, namely

$$\hat{t}_{x_\pi} = \sum_{i \in S} \pi_i^{-1} x_i,$$

is not necessarily equal to $t_{\mathbf{x}}$. The purpose of calibration is to modify the design weights $w_i = 1/\pi_i$ in order to produce calibrated weights w_{ci}, a) which are close to the original weights w_i, and b) which enable to match exactly the known totals $t_{\mathbf{x}}$. The purpose of constraint a) is that the new weights remain close to the design weights, so that the new estimator remains approximately unbiased. The purpose of constraint b) to have coherence between estimations computed from S and some known auxiliary totals. Also we may obtain a variance reduction if the variable y is well explained by the auxiliary variables.

More precisely, Deville and Särndal (1992) propose to solve the optimization problem

$$\min_{\{w_{ci}\}} \sum_{i \in s} \frac{w_i}{q_i} G\left(\frac{w_{ci}}{w_i}\right) \quad \text{such that} \quad \sum_{i \in s} w_{ci} \mathbf{x}_i = t_{\mathbf{x}}, \tag{2.5.1}$$

with $G(\cdot)$ some distance function, and q_i some weight attributed to unit i in the sample, see Deville and Särndal (1992). This leads to the calibrated estimator

$$\hat{t}_{y_{cal}} = \sum_{i \in S} w_{ci} y_i.$$

Under some conditions on the sampling design, the distance function, the variable of interest and the calibration variables, Deville and Särndal (1992) prove that the variance of the calibrated estimator is approximately given by

$$V_p\left(\hat{t}_{y_{cal}}\right) \simeq V_p\left(\hat{t}_{E_\pi}\right), \tag{2.5.2}$$

where

$$E_i = y_i - \mathbf{B}^\top \mathbf{x}_i \quad \text{and} \quad \mathbf{B} = \left(\sum_{i\in U} \frac{\mathbf{x}_i \mathbf{x}_i^\top}{q_i}\right)^{-1} \sum_{i\in U} \frac{\mathbf{x}_i y_i}{q_i}. \tag{2.5.3}$$

It follows from (2.5.2) that the variance of the calibrated estimator may be significantly reduced, as compared to that of the Horvitz-Thompson estimator, if there exists a strong (linear) relationship between the auxiliary variables and the variable of interest.

In the particular case when the Euclidean distance is used, we have

$$G(x) = \frac{1}{2}(x-1)^2.$$

In this case, the calibrated estimator $\hat{t}_{y_{cal}}$ is called the Generalized REGression (GREG) estimator, and it simplifies as

$$\hat{t}_{y_{greg}} = \hat{t}_{y_\pi} + \hat{\mathbf{B}}^\top \left(t_{\mathbf{x}} - \hat{t}_{\mathbf{x}_\pi}\right),$$

where

$$\hat{\mathbf{B}} = \left(\sum_{i\in S} w_i \frac{\mathbf{x}_i \mathbf{x}_i^\top}{q_i}\right)^{-1} \sum_{i\in S} w_i \frac{\mathbf{x}_i y_i}{q_i}.$$

A particular important case occurs when one auxiliary variable only is used, say $\mathbf{x}_i = x_i$, and when $q_i = x_i$. In such case, the GREG estimator simplifies to give the so-called ratio estimator

$$\hat{t}_{y_r} = t_x \frac{\hat{t}_{y_\pi}}{\hat{t}_{x_\pi}}.$$

From (2.5.2), the variance of the ratio estimator can be approximated by

$$V_p(\hat{t}_{y_r}) \simeq \sum_{i\in U}\sum_{k\in U}(\pi_{ik}-\pi_i\pi_k)\frac{y_i - Rx_i}{\pi_i}\frac{y_k - Rx_k}{\pi_k},$$
$$\text{where} \quad R = \frac{t_y}{t_x}.$$

A possible variance estimator is given by

$$\hat{V}_{HT}(\hat{t}_{y_r}) = \sum_{i\in S}\sum_{k\in S}\frac{(\pi_{ik}-\pi_i\pi_k)}{\pi_{ik}}\frac{(y_i - \hat{R}x_i)}{\pi_i}\frac{(y_k - \hat{R}x_k)}{\pi_k},$$
$$\text{where} \quad \hat{R} = \frac{\hat{t}_{y_\pi}}{\hat{t}_{x_\pi}}.$$

The ratio estimator is useful when there is strong positive correlation between x_i and y_i (e.g., Särndal et al., 1992). On the contrary, if the correlation is negative, the ratio estimator is actually worse than the Horvitz-Thompson estimator.

The Hajek estimator of t_y is a special case of the ratio estimator, obtained with $x_i = 1$.

Definition 2.13. *The Hajek estimator of the finite population total t_y is*

$$\hat{t}_{y_H} = N\left(\sum_{i\in S}\frac{1}{\pi_i}\right)^{-1}\sum_{i\in S}\frac{y_i}{\pi_i}.$$

Definition 2.14. The Hajek estimator of the finite population mean $\bar{y}$ is

$$\hat{\bar{y}}_H = \left(\sum_{i\in S}\frac{1}{\pi_i}\right)^{-1}\sum_{i\in S}\frac{y_i}{\pi_i}$$

2.6 Non-response

Design-based theory, as presented in the previous sections, is applicable when the survey has complete response. As quoted by Rässler, Rubin and Schenker (2009), survey data can be imperfect in various ways. For example, errors due to noncoverage, problems with interviewers or missing values may affect data

quality. In particular, surveys typically suffer from missing-data problems due to nonresponse. Haziza and Kuromi (2007) listed the main effects of nonresponse: (i) bias of point estimators; (ii) increase in the variance of point estimators and (iii) bias of complete data variance estimators.

Little and Rubin (2002) distinguish four main groups of methods in incomplete data analysis. They gather in the first group simple procedures such as complete-case analysis, which discard the units with incomplete data and analyze only the units with complete data. The second group of methods includes weighting procedures which introduce a factor into the survey weight for each responding unit equal to the inverse of the estimated probability of response for that unit. The third group is composed of methods for handling item nonresponse such as single and multiple imputation. Single imputation methods fill in values that are missing and the completed data are then analyzed as if they were fully observed data. Multiple imputation is designed for reflecting the added uncertainty due to the fact that imputed values are usually not the real values. The last group of methods includes direct analyses using model based procedures, in which models are built for the observed data, with inferences based on likelihood or Bayesian analyses.

We treat separately item response from unit response in the two following subsections and give some corresponding methods to deal with missingness.

2.6.1 Item nonresponse

The following subsection is inspired from Haziza (2009) and Haziza and Kuromi (2007). A so called nonresponse bias occurs if respondents and nonrespondents are different in expectation with respect to the survey variables. Moreover, non-response reduces the observed sample size in comparison with the sample size initially planned. Thus, it induces an increase in the variance of estimators, which is called the nonresponse variance. Imputation methods aim both at reducing the nonresponse bias and controling the nonresponse variance as much as

possible. Since imputation is essentially a modeling exercise, auxiliary variables available for both respondents and nonrespondents are necessary. The quality of the imputed estimates will thus depend on the availability and judicious use of good auxiliary information at the imputation stage.

Let r_i be the response indicator, such that $r_i = 1$ if unit i responded to item y, and $r_i = 0$ otherwise. We note $\mathbf{r} = (r_1, \ldots, r_N)^\top$ for the vector of the response indicators. Let $p_i = P(r_i = 1)$ denote the response probability to item y for unit i. Assuming that the individuals respond independently of one another, we have $p_{ik} = P(r_i = 1, r_k = 1) = p_i p_k$, for all $i, k \in S$.

Definition 2.15. *The unknown distribution of the response indicators, $P(r_i \mid S)$ is called the nonresponse mechanism.*

In the presence of nonresponse to item y, it is not possible to compute the Horvitz Thompson estimator for t_y since some y-values are missing. An imputation mechanism is used to replace the missing values. That is, an artificial value y_i^* is used to replace the missing y_i.

Definition 2.16. An imputed estimator for t_y based on observed and imputed values is

$$\hat{t}_{yI} = \sum_{i \in S} w_i r_i y_i + \sum_{i \in S} w_i (1 - r_i) y_i^*. \tag{2.6.1}$$

Remark 2.3. The imputed estimator (2.6.1) is the weighted mean of the observed and the imputed values which depend on the imputation method used to replace the missing data.

The imputation mechanism is motivated by an underlying imputation model, which corresponds to a set of assumptions on the variable y subject to missingness. An imputation model is a set of assumptions about the distribution of the variable requiring imputation. Regression imputation is motivated by the following linear regression model m.

Definition 2.17. *The so called regression imputation model is:*

$$m: \ y_i = \mathbf{x}^\top \beta + \epsilon_i \tag{2.6.2}$$

with $\mathbf{x}_i$ *a vector of auxiliary variables, which is assumed to be known on the whole sample including non-respondents and*

$$E_m(\epsilon_i) = 0, \ E_m(\epsilon_i \epsilon_k) = 0 \ if i \neq k, E_m(\epsilon_i^2) = \sigma^2$$

where $E_m(.)$ *denotes the expectation with respect to the model (2.6.2).*

In *deterministic regression imputation*, a missing value y_i is replaced by its predicted value, $\hat{y}_i$ obtained by fitting the imputation model (2.6.2) using the respondents y-values only:

$$y_i^* = \hat{y}_i = \mathbf{x}_i^\top \hat{\beta}_r \tag{2.6.3}$$

where

$$\hat{\beta}_r = \left(\sum_{i \in S} w_i r_i \mathbf{x}_i \mathbf{x}_i^\top \right)^{-1} \left(\sum_{i \in S} w_i r_i \mathbf{x}_i y_i \right)$$

which is the weighted least square estimator of β.

We now turn to *random regression imputation*. Let us denote S_r the set of respondents for the variable of interest y. The imputed value used for a missing y_i is:

$$y_i^* = \hat{y}_i + \epsilon_i^* \tag{2.6.4}$$

where $\hat{y}_i$ is provided by (2.6.3), and ϵ_i^* is a residual randomly drawn from the observed estimated residuals $e_k = y_k - \mathbf{x}_k^\top \hat{\beta}_r$, with

$$P(\epsilon_i^* = e_k) \ = \ \frac{w_k}{\sum_{l \in S} w_l r_l}.$$

One drawback of deterministic regression imputation is the distortion of the distribution of the variables of interest being imputed. This distortion grows along with the nonresponse rate and the lack of adequacy of the model. By contrast, random regression imputation tends to preserve the distribution of the variables

of interest. However, it suffers from an additional component of variance coming from the randomness of the imputation mechanism.

The properties of the imputed estimator can be studied by using the decomposition of its total error:

$$\hat{t}_{y_I} - t_y = (\hat{t}_{y_\pi} - t_y) + (\hat{t}_{y_I} - \hat{t}_{y_\pi}) \tag{2.6.5}$$

The first term $(\hat{t}_{y_\pi} - t_y)$ on the right-hand side of (2.6.5) is called the *sampling error* of $\hat{t}_{y_I}$ whereas the second term $(\hat{t}_{y_I} - \hat{t}_{y_\pi})$ is called the *nonresponse error* of $\hat{t}_{y_I}$.

Let $E_q(.)$ and $V_q(.)$ denote the expectation and variance under the non-response mechanism, conditionally on the vector $\mathbf{y}_U$ of population values and on the vector $\boldsymbol{\delta}_U$ of sample membership indicators.

Definition 2.18. The bias of the imputed estimator is defined as:

$$\begin{aligned} B(\hat{t}_{y_I}) &= E(\hat{t}_{y_I} - t_y) \\ &= E_p E_q(\hat{t}_{y_I} - t_y \mid S) \\ &= E_p(\hat{t}_{y_\pi} - t_y) + E_p E_q(\hat{t}_{y_I} - \hat{t}_{y_\pi} \mid S) \\ &= E_p(B_q) \end{aligned}$$

where $B_q = E_q(\hat{t}_{y_I} - \hat{t}_{y_\pi} \mid S)$ denotes the conditional nonresponse bias.

We now turn to the general definitions of two types of missingness given by Rubin (1996).

Definition 2.19. Assuming that the true probability of response associated with unit i is related to a certain vector of variables $\mathbf{x}_i$:

i) if the vector $\mathbf{x}_i$ contains fully observed variables only, then the data are said to be Missing At Random (MAR) and the response mechanism is *ignorable*,

ii)if the vector $\mathbf{x}_i$ includes variables that are subject to missingness, then the data are Not Missing At Random (NMAR) and the response mechanism is *not ignorable*.

The imputed estimator is unbiased if $B_q = 0$. The nonresponse bias B_q will be negligible if the vector of auxiliary variables is correctly specified and the nonresponse mechanism is *ignorable*.

In the following chapter of this thesis, we assume that nonresponse mechanism is *ignorable*.

Definition 2.20. *Assuming that the imputed estimator $\hat{t}_{y_I}$ is conditionally unbiased for $\hat{t}_{y_\pi}$, the variance of the imputed estimator is defined as:*

$$\begin{aligned} V(\hat{t}_{y_I}) &= E(\hat{t}_{y_I} - t_y)^2 \\ &= E_p E_q(\hat{t}_{y_I} - \hat{t}_{y_\pi} \mid S)^2 + E_p E_q(\hat{t}_{y_\pi} - t_y)^2 \\ &+ 2E_p E_q\left[(\hat{t}_{y_I} - \hat{t}_{y_\pi})(\hat{t}_{y_\pi} - t_y) \mid S\right] \\ &= E_p(\hat{t}_{y_\pi} - t_y)^2 + E_p E_q(\hat{t}_{y_I} - \hat{t}_{y_\pi} \mid S)^2 \\ &= V_p(\hat{t}_{y_\pi}) + E_p V_q(\hat{t}_{y_I} - \hat{t}_{y_\pi} \mid S) \end{aligned}$$

where $V_p(\hat{t}_{y_\pi})$ represents the sampling variance and $E_p V_q(\hat{t}_{y_I} - \hat{t}_{y_\pi} \mid S)$ the nonresponse variance.

The sampling variance depends on the sampling procedure, the selected sample size and the population being sampled. The nonresponse variance tends to be lower with a high response rate and a good predictive power of the imputation model.

There are some risks in case of using imputation, as underlined by Haziza and Kuromi (2007), for example. First, imputation leads to a complete data file but inference will be valid only if the assumptions made on the response mechanism and/or the imputation model are truly satisfied. Also, some imputation methods distort the distribution of the imputed variables. For example, marginal imputation treating each item separately distorts the relationships between variables (see Chapters 3 and 4). Finally, in terms of variance estimation, treating the imputed values as if they were observed may lead to substantially negatively biased

variance estimators, and more particularly if the nonresponse rate is appreciable.

2.6.2 Unit nonresponse

Weighting adjustment is often used to handle unit nonresponse in sample surveys. Groves et al. (2002) and Särndal & Lundström (2005) provided comprehensive overviews of nonresponse weighting adjustment (NWA) methods in survey sampling.

Each element of the population is considered as having its own individual probability of responding. The sampling weight of the respondent is increased using the information observed in the sample. It allows respondents to properly represent the original population. This procedure comes from the theory for two-phase sampling, according to which the set of respondents is treated as a second phase sample from the original sample: it leads to multiply the inverse of the response probability by the sampling weight of each respondent. However, unlike the sampling phase, the response phase is beyond the control of the statistician since unit nonresponse occurs with unknown probabilities. The estimation theory built around the idea that each unit i is equipped with a known individual inclusion probability, π_i, and an unknown individual response probability, p_i is called "quasi-randomization theory" (Oh and Scheuren, 1983).

In practice, two types of weighting procedures are commonly used (Haziza and Lesage, 2016): in the first, the basic weights are multiplied by the inverse of the estimated response probabilities, whereas the second uses some form of calibration for adjusting the basic weights, which includes post-stratification and ratio estimation as special cases. In this thesis, we focus on weighting adjustments by the inverse of the estimated response probabilities.

In the presence of unit nonresponse, the survey variables are recorded for a subset S_r of the original sample S. This subset is often referred to as the set of respondents. Let r_i be a response indicator such that $r_i = 1$ if unit i is a

respondent and $r_i = 0$, otherwise. We assume that the true probability of response associated with unit i is related to a certain vector of variables $\mathbf{x}_i$; that is, $p_i = P(r_i = 1 \mid S, \mathbf{v}_i)$. We assume that $0 < p_i \leq 1$ and that the response indicators are mutually independent. The latter assumption is generally not realistic in the context of multistage sampling designs because sample units within the same cluster (e.g., household) may not respond independently of one another; see Skinner and D'Arrigo (2011) and Kim et al. (2016) for a discussion of estimation procedures accounting for the possible intra-cluster correlation.

Little and Rubin (2002) distinguish three missing-data mechanisms: missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR).

Definition 2.21. *Missing Completely At Random (MCAR). Data are missing completely at random if the missingness is unrelated to the (unknown) missing values of that variable as well as unrelated to the values of other variables. We have*

$$P(r_i = 1 \mid y_i; \boldsymbol{x}_i) = P(r_i = 1), \ \forall i \in S.$$

In the MCAR situation, the only impact of nonresponse is an increase in estimators' variance, since the sample size is reduced.

Definition 2.22. *Missing At Random (MAR). Data are missing at random if the missingness is possibly related to the observed data in the data set, but, conditionally on these data, is not related to any unknown value. We have*

$$P(r_i = 1 \mid y_i; \boldsymbol{x}_i) = Pr(r_i = 1 \mid \boldsymbol{x}_i), \ \forall i \in S.$$

where y_i is the variable of interest, and $\mathbf{x}_i$ a vector of auxiliary variables known for each unit $i \in S$.

It means that the missing values are a random sample of all values within classes defined by observed values (i.e., conditional on the observed data, the missingness is completely at random).

Definition 2.23. *Not Missing At Random (NMAR). The missingness depends on some unobserved (missing) values, even after conditioning on all observed values. We have*

$$P(r_i = 1 \mid y_i; \boldsymbol{x}_i) \neq Pr(r_i = 1 \mid \boldsymbol{x}_i) \ \forall i \in S$$

where y_i is the variable of interest, and $\boldsymbol{x}_i$ a vector of auxiliary variables known for each unit i in S.

MCAR can be unrealistically restrictive and in practice, it is not possible to determine whether or not the MAR assumption holds. However, the MAR assumption can be made more plausible by conditioning on fully observed variables that are related to both the probability of response and the survey variables; e.g., Little and Vartivarian (2005). Thus, in chapter 5, we assume that response mechanisms under study are MAR.

If the response probabilities p_i were known, an unbiased estimator of t_y would be the double expansion estimator (Särndal et al., 1992):

$$\widehat{t}_{y,DE} = \sum_{i \in S_r} \frac{w_i}{p_i} y_i. \tag{2.6.6}$$

In practice, the response probabilities p_i are not known and need to be estimated. To that end, a model for the response indicators r_i, called a nonresponse model, is assumed and the estimated probabilities $\widehat{p}_i$ are obtained using the postulated model (e.g., Särndal and Swensson, 1987; Ekholm and Laaksonen, 1991). This leads to the Propensity Score Adjusted (PSA) estimator:

$$\widehat{t}_{y,PSA} = \sum_{i \in S_r} \frac{w_i}{\widehat{p}_i} y_i, \tag{2.6.7}$$

where $\widehat{p}_i$ is an estimate of p_i. An alternative estimator of t_y is the so-called Hajek estimator:

$$\widehat{t}_{y,H} = \frac{N}{\widehat{N}} \sum_{i \in S_r} \frac{w_i}{\widehat{p}_i} y_i, \tag{2.6.8}$$

where $\widehat{N} = \sum_{i \in S_r} \frac{w_i}{\widehat{p}_i}$ is an estimate of the population size N based on the respondents.

The estimated response probabilities in (5.0.3) or (5.0.4) may be obtained through parametric or nonparametric methods. In the context of parametric estimation, we assume that

$$p_i = f(\mathbf{x}_i, \boldsymbol{\alpha}), \tag{2.6.9}$$

for some function $f(\mathbf{z}_i, .)$, where $\boldsymbol{\alpha}$ is a vector of unknown parameters. The estimated response probabilities are given by

$$\widehat{p}_i = f(\mathbf{x}_i, \widehat{\boldsymbol{\alpha}}),$$

where $\widehat{\boldsymbol{\alpha}}$ is a suitable estimator (e.g., maximum likelihood estimator) of $\boldsymbol{\alpha}$. The class of parametric models (5.0.5) includes the popular linear logistic regression model as a special case. It is given by

$$p_i = \frac{exp(\mathbf{x}_i^\top \boldsymbol{\alpha})}{1 + exp(1 + \mathbf{x}_i^\top \boldsymbol{\alpha})}.$$

There are several issues associated with the use of a parametric model: (i) they are not robust to the misspecification of the form of $f(\mathbf{x}_i, .)$; (ii) they can fail to account properly on local violations of the parametric assumption such as nonlinearities or interaction effects, both of which may not have been detected during model selection; (iii) they may yield very small estimated response probabilities, resulting in very large nonresponse adjustment factors $\widehat{p}_i^{-1}$, ultimately leading to potentially unstable estimates; e.g., Little and Vartivarian (2005) and Beaumont (2005).

The estimated response probabilities are used to correct for nonresponse bias. Consequently, the NWA estimators reduce nonresponse bias by incorporating the estimated response probabilities in the estimators. Some authors as Little and Vartivarian (2005) and Kim and Kim (2007) also argued that the estimators using the estimated response probability could be more efficient than the estimators

using the true response probability.

In chapter 3 and 4, we propose methods to overcome specific problems induced by item nonresponse : preserving correlation coefficient between two variables of interest with MIVQUE based imputation (chapter 3) and preserving the distribution function in case of imputation for zero inflated data (chapter 4). In chapter 5, we compare different *machine learning methods* (parametric, non parametric and including models agregation) to estimate response probabilities in case of unit nonresponse, aiming at estimating finite population totals in the best way.

References

Ardilly, P. (2006). *Les techniques de sondage*. Technip.

Beaumont, J.-F. (2005). Calibrated imputation in surveys under a quasi-model-assisted approach. *Journal of the Royal Statistical Society: Series B* (Statistical Methodology), 67, 445–458.

Cochran, W. G. (1977). *Sampling techniques*, Third edition. New York: Wiley.

Biemer P.P. and Christ S.L. (2008). *International Handbook of Survey Methodology*, 317-341. New York: Lawrence Erlbaum Associates.

Deville, J-C., and Särndal, C-E. (1992). Calibration estimators in survey sampling. *Journal of the American statistical Association*, 87(418), 376–382.

Favre-Martinoz C. (2015). *Estimation robuste en population finie et infinie*, PhD thesis, University of Rennes 1, European University of Brittany.

Groves, R. M., Dillman, D., Eltinge, J. L., Little,R. J. A. (2002). *Survey Nonresponse*. New York: Wiley.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R.(2004). *Survey methodology*. New York: Wiley.

Hajek, J. (1971). Comment on An essay on the logical foundations of survey sampling by Basu, D. in *Foundations of Statistical Inference* (Godambe, V.P. and Sprott, D.A. eds.), p. 236. Holt, Rinehart and Winston.

Haziza, D. (2009). Imputation and inference in the presence of missing data. *Handbook of statistics*, 29, 215-246.

Haziza, D., Kuromi G. (2007). Handling Item Non Response in Surveys. *CS-BIGS*, 1(2), 102-118.

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

Kim J.-K. (2014), *Theory and Applications of Sample Surveys*, Textbook, Iowa State University.

Kim, J. K., and Kim, J. J. (2007). Nonresponse weighting adjustment using estimated response probability. *The Canadian Journal of Statistics*, 35, 501–514.

Lavallée, P. and Beaumont, J.-F. (2015). Why We Should Put Some Weight on Weights. *Survey Insights: Methods from the Field*, Weighting: Practical Issues and 'How to' Approach, Invited article,
Retrieved from $http://surveyinsights.org/?p=6255$.

Leeuw, E. D., Hox, J. J., Dillman, D. A., & European Association of Methodology. (2008). *International handbook of survey methodology.* New York: Lawrence Erlbaum Associates.

Lesage E. (2013), *Utilisation d'information auxiliaire en théorie des sondages à l'étape de l'échantillonnage et à l'étape de l'estimation*, PhD thesis, University of Rennes 1, European University of Brittany.

Little, R.J.A., Rubin, D.B. (2002). *Statistical analysis with missing data.* (Second Edition.) New York: Wiley.

Little R. J. A. and Vartivarian S. (2005). Does weighting for nonresponse increase the variance of survey means? *Survey Methodology*, 31, 161–168.

Lohr S. L.(2008). *International Handbook of Survey Methodology*, 97–112. New York: Lawrence Erlbaum Associates.

Montaquila, J. M., Kalton, G. (2010). *Sampling from Finite Populations*, Westat 1600 Research Blvd., Rockville, MD 20850, USA.

Oh, H.L. and Scheuren, F.S. (1983). Weighting adjustments for unit nonresponse, In *Incomplete Data in Sample Surveys*, 2,(Madow, W.G., Olkin, I. and Rubin, D.B., eds.) New York: Academic Press.

Pfeffermann, D., and Rao, C.R., eds. (2009). *Handbook of Statistics. Volume 29A, Sample Surveys: Design, Methods and Application and Volume 29B, Sample Surveys: Inference and Analysis.* New York: Elsevier.

Rässler, S., Rubin D. B., Schenker N. (2008) *International Handbook of Survey Methodology*, 370-386. New York: Lawrence Erlbaum Associates.

Schafer, J.L., and Graham, J.W. (2002). Missing Data: Our View of the State of the Art, *Psychological Methods*, 7(2), 147-177.

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

Särndal, C.-E., and Lundström, S. (2005). *Estimation in Surveys with Nonresponse.* New York: Wiley.

Sen, A.R. (1953). On the estimate of the variance in sampling with varying probabilities, *Journal of the Indian Society of Agricultural Statistics*, 5, 119-127.

Thompson, S. T.(2012). *Sampling*, Third Edition. John Wiley & Sons, Inc.

Tillé, Y. (2001). *Théorie des sondages : Echantillonnage et estimation en populations finies : cours et exercices.* Dunod.

Yates, F. and Grundy, P. M. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society: Series B*, 15, 235-261.

Chapter 3

Preserving relationships between variables with MIVQUE based imputation

In this paper, we study the problem of preserving the relationships between items requiring imputation. Marginal imputation, which consists of treating items separately, tends to distort the relationships because this type of imputation procedure does not account for the existing relationships between items. When the interest lies in estimating a simple parameter such as a population mean or total, it is generally possible to produce a simple imputation procedure such as deterministic or random regression imputation. The latter leads to asymptotically unbiased estimator of simple parameters provided that the assumed imputation model is correctly specified. On the other hand, these types of imputation may lead to severely biased estimators of parameters measuring relationships (e.g., a coefficient of correlation), if applied separately for each variable requiring imputation.

To overcome this problem, two main approaches have been studied in the literature: the first consists of using a marginal imputation procedure followed by a bias-adjustment procedure at the estimation stage. This approach was investigated by Skinner and Rao (2002) and Chauvet and Haziza (2012), among others. In the second approach, the missing values are imputed using a joint procedure,

which accounts for the relationships between items. Shao and Wang proposed a joint random regression imputation procedure and showed that it leads to asymptotically unbiased estimators of coefficients of correlation; see also Chauvet and Haziza (2012) for a fully efficient version of the Shao-Wang procedure.

The Shao-Wang procedure belongs to the class of random imputation procedures. Unlike deterministic imputation procedures, random procedures suffer from an additional variability, called the imputation variance, leading to somehow inefficient estimators. In this paper, we propose a modification of the Shao-Wang procedure, that can be implemented in two steps. In the first step, initial imputed values are obtained using the Shao-Wang procedure. In the second step, the initial imputed values are iteratively modified so that appropriate calibration constraints are satisfied. We propose to calibrate on Minimum In Variance Quadratic Unbiased Estimators (MIVQUE), which are based on a geometrical interpretation of the covariance structure of variables. The choice of calibrating on MIVQUEs is first motivated by the unbiasedness condition which is expected to provide robustness of correlation estimation with respect to nonresponse and consequently to help in preserving the relationship between variables. Moreover, as the theory of MIVQUE is not based on a likelihood assumption but on algebraic considerations on moment estimators, its application to nonresponse issues (see Causeur, 2006) in sample surveys is straightforward.

On the one hand, satisfying the calibration constraints ensures that the imputation variance is virtually eliminated. On the other hand, calibrating on the MIVQUE leads to efficient estimators of parameters such as marginal first and second moments as well as coefficients of correlation when the bivariate distribution of the study variables is symmetric or exhibits a low degree of asymmetry. The idea of calibrated imputation has been investigated in the context of outliers and robust estimation by Ren and Chambers (2002). The idea of finding imputed values that satisfy constraints can also be found in Beaumont (2005), Favre *et al.*

(2005), Rancourt and Liu (2001), Chauvet *et al.* (2011) and Chauvet and Haziza (2012).

This paper is organised as follows. In Section 2, we describe the theoretical set-up and present the joint imputation procedure of Shao and Wang (2002). In Section 3, we describe the MIVQUE approach which is based on a geometrical interpretation of the covariance structure. A weighted version of the MIVQUE, which is useful in the context of survey sampling, is also introduced. A two steps MIVQUE based imputation is proposed in Section 4 and its properties are discussed. In section 5, the results of a simulation study, comparing the Shao-Wang procedure and the proposed procedure in terms of bias and relative efficiency, are presented. We give some final remarks in Section 6.

3.1 Theoretical set-up

Let U be a finite population of size N. We are interested in estimating the finite population coefficient of correlation between the variables y_1 and y_2 :

$$R_{12} = \frac{t^{11} - t^{10}t^{01}/N}{\{t^{20} - (t^{10})^2/N\}^{1/2}\{t^{02} - (t^{01})^2/N\}^{1/2}},$$

where $t^{kl} = \sum_{i\in U}(y_{1i})^k(y_{2i})^l$ with $(k,l) \in \{(1,0),(2,0),(1,1),(0,1),(0,2)\}$. For example, $t^{10} = \sum_{i\in U} y_{1i}$ and $t^{11} = \sum_{i\in U} y_{1i}y_{2i}$.

We select a sample S of size n according to a sampling design $p(.)$. Let $w_i = 1/\pi_i$ be the sampling weight attached to unit i, where $\pi_i = P(i \in S)$ denotes its first-order inclusion probability in the sample.

A complete data estimator of R_{12} is the plug-in estimator

$$\hat{R}_{12,\pi} = \frac{\hat{t}_\pi^{11} - \hat{t}_\pi^{10}\hat{t}_\pi^{01}/\hat{N}_\pi}{\{\hat{t}_\pi^{20} - (\hat{t}_\pi^{10})^2/\hat{N}_\pi\}^{1/2}\{\hat{t}_\pi^{02} - (\hat{t}_\pi^{01})^2/\hat{N}_\pi\}^{1/2}},$$

where $\hat{t}_{\pi}^{kl} = \sum_{i \in S} w_i (y_{1i})^k (y_{2i})^l$ and $\hat{N}_{\pi} = \sum_{i \in S} w_i$ denote the expansion type estimators of t^{kl} and N, respectively. Under some regularity conditions, the estimator $\hat{R}_{12,\pi}$ is asymptotically design-unbiased for R_{12} (e.g., Deville, 1999).

We now turn to the case where both y_1 and y_2 are subject to missingness. Let r_{1i} be a response indicator variable corresponding to y_1. Let y_{1i}^* denote the imputed value used to replace the missing y_{1i}, and let $\tilde{y}_{1i} = y_{1i}$ if $r_{1i} = 1$ and $\tilde{y}_{1i} = y_{1i}^*$ if $r_{1i} = 0$. The quantities r_{2i}, y_{2i}^* and $\tilde{y}_{2i}$ are similarly defined for y_{2i}. An imputed estimator of R_{12} based on observed and imputed values is defined as

$$\hat{R}_{12,I} = \frac{\hat{t}_I^{11} - \hat{t}_I^{10}\hat{t}_I^{01}/\hat{N}_{\pi}}{\{\hat{t}_I^{20} - (\hat{t}_I^{10})^2/\hat{N}_{\pi}\}^{1/2}\{\hat{t}_I^{02} - (\hat{t}_I^{01})^2/\hat{N}_{\pi}\}^{1/2}}, \tag{3.1.1}$$

where $\hat{t}_I^{kl} = \sum_{i \in S} w_i (\tilde{y}_{1i})^k (\tilde{y}_{2i})^l$. Note that (3.1.1) can be readily computed by secondary analysts using complete data software as it does not require the response indicators to be available in the imputed data file.

We consider the class of linear regression imputation procedures. We assume that the following bivariate model holds:

$$\begin{aligned} y_{1i} &= \boldsymbol{\beta}_1^{\top} \boldsymbol{x}_i + \epsilon_{1i}, \\ y_{2i} &= \boldsymbol{\beta}_2^{\top} \boldsymbol{x}_i + \epsilon_{2i}, \end{aligned} \tag{3.1.2}$$

where $\boldsymbol{x}_i$ is a vector of auxiliary variables attached to unit i available for all the sample units (respondents and nonrespondents), $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are vectors of unknown parameters. The errors ϵ_{1i} (respectively, ϵ_{2i}) are independent random variables with mean 0 and unknown variance σ_{11} (respectively, σ_{22}). We assume that the covariance matrix of $(\epsilon_{1i}, \epsilon_{2i})^{\top}$ is

$$\begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix},$$

where $\sigma_{12} \equiv Cov_m(\epsilon_{1i}, \epsilon_{2i})$ and the subscript m denotes model (3.1.2).

In this paper, the properties of point estimators are evaluated with respect to the Imputation Model (IM) approach; e.g., Haziza (2009). In this approach, inference is made with respect to the joint distribution induced by the imputation model, the sampling design, and the nonresponse mechanism. We adopt the following notation: let $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_N)$ with $\mathbf{y}_i = (y_{1i}, y_{2i})^\top$; let $\boldsymbol{\delta}= (\delta_1, \ldots, \delta_N)^\top$, where $\delta_i = 1$ if unit $i \in S$ and $\delta_i = 0$, otherwise; finally, let $\mathbf{r} = (\mathbf{r}_1, \ldots, \mathbf{r}_N)$ with $\mathbf{r}_i = (r_{1i}, r_{2i})^\top$.

We denote by $E_q(.) \equiv E(. \mid \mathbf{y}, \delta)$ the expectation with respect to the nonresponse model: except for the response indicators $\mathbf{r}_i$, all the other variables involved in point and variance estimators are treated as fixed. We denote by $E_m(.) \equiv E(. \mid \delta, \mathbf{r})$ the expectation with respect to the imputation model : except for the variables of interest $\mathbf{y}$, all the other variables involved in point and variance estimators are treated as fixed. Finally, we denote by $E_I(.) \equiv E(. \mid \mathbf{y}, \delta, \mathbf{r})$ the expectation with respect to the imputation mechanism in the case of random imputation procedure: except for the imputed values y_{1i}^* and y_{2i}^*, all the other variables involved in point and variance estimators are treated as fixed. Note that the auxiliary variables $\mathbf{x}$ in (3.1.2) are always treated as fixed.

Under a random imputation procedure, the total error of $\hat{R}_{12,I}$ can be expressed as

$$\hat{R}_{12,I} - R_{12} = (\hat{R}_{12,\pi} - R_{12}) + (\check{R}_{12,I} - \hat{R}_{12,\pi}) + (\hat{R}_{12,I} - \check{R}_{12,I}), \tag{3.1.3}$$

where the first term on the right hand side of (3.1.3) denotes the sampling error, the second and third terms denote the nonresponse error and imputation error, respectively, and $\check{R}_{12,I} = E_I\left(\hat{R}_{12,I}\right)$.

Under random imputation, the conditional nonresponse bias of $\hat{R}_{12,I}$ is defined as

$$B_{mqI}(\hat{R}_{12,I}) = E_m E_q E_I(\hat{R}_{12,I} - \hat{R}_{12,\pi}) = E_q E_m E_I(\hat{R}_{12,I} - \hat{R}_{12,\pi}),$$

where the subscript q denotes the unknown nonresponse mechanism. The second equality in the previous formula is justified when the sampling design is nonformative and the data are Missing At Random (Rubin, 1976), which we assume to be the case in this paper. That is, model (3.1.2) holds for the respondents.

If each term $t^{kl} = \sum_{i \in U} (y_{1i})^k (y_{2i})^l$ is consistently estimated, then $\hat{R}_{12,I}$ is a consistent estimator of R_{12}, provided some mild regularity conditions are satisfied; see, for example Cardot et al. (2013). For the marginal first moments t^{01} and t^{10}, an appropriate deterministic or random marginal regression imputation procedure may be used, whereas the marginal second moments t^{02} and t^{20} require a marginal random imputation procedure (see Appendix A). The main difficulty lies in estimating the cross product term t^{11} in an (asymptotically) unbiased fashion. Unlike for the marginal first and second moments, marginal imputation procedures may lead to a severely biased estimator of t^{11} because they do not account for the existing relationship between y_1 and y_2. To overcome this problem, Shao and Wang (2002) proposed a joint regression imputation procedure, which is described next.

Missing y_{1i} and y_{2i} are imputed by y_{1i}^* and y_{2i}^* with

$$\begin{aligned} y_{1i}^* &= \hat{\boldsymbol{\beta}}_1^{r\top} \boldsymbol{x}_i + \epsilon_{1i}^*, \\ y_{2i}^* &= \hat{\boldsymbol{\beta}}_2^{r\top} \boldsymbol{x}_i + \epsilon_{2i}^* \end{aligned} \tag{3.1.4}$$

where

$$\begin{aligned} \hat{\boldsymbol{\beta}}_1^r &= \left(\sum_{i \in S} w_i r_{1i} \boldsymbol{x}_i \boldsymbol{x}_i^\top \right)^{-1} \sum_{i \in S} w_i r_{1i} \boldsymbol{x}_i y_{1i}, \\ \hat{\boldsymbol{\beta}}_2^r &= \left(\sum_{i \in S} w_i r_{2i} \boldsymbol{x}_i \boldsymbol{x}_i^\top \right)^{-1} \sum_{i \in S} w_i r_{2i} \boldsymbol{x}_i y_{2i}. \end{aligned} \tag{3.1.5}$$

In addition, we define the estimators of σ_{11} , σ_{22} and σ_{12} by

$$\begin{aligned}
\hat{\sigma}^r_{11} &= \frac{1}{\sum_{i\in S} w_i r_{1i} r_{2i}} \sum_{i\in S} w_i r_{1i} r_{2i} e^2_{1i}, \\
\hat{\sigma}^r_{22} &= \frac{1}{\sum_{i\in S} w_i r_{1i} r_{2i}} \sum_{i\in S} w_i r_{1i} r_{2i} e^2_{2i}, \\
\hat{\sigma}^r_{12} &= \frac{1}{\sum_{i\in S} w_i r_{1i} r_{2i}} \sum_{i\in S} w_i r_{1i} r_{2i} e_{1i} e_{2i},
\end{aligned} \tag{3.1.6}$$

respectively, where $e_{1i} = y_{1i} - \hat{\boldsymbol{\beta}}_1^{r\top} \boldsymbol{x}_i$ and $e_{2i} = y_{2i} - \hat{\boldsymbol{\beta}}_2^{r\top} \boldsymbol{x}_i$.

The random residuals ϵ^*_{1i} and ϵ^*_{2i} are generated as follows:

(i) If y_{1i} is missing but y_{2i} is observed, we use

$$\epsilon^*_{1i} = \frac{\hat{\sigma}^r_{12}}{\hat{\sigma}^r_{22}} (y_{2i} - \hat{\boldsymbol{\beta}}_2^{r\top} \boldsymbol{x}_i) + \tilde{\epsilon}^*_{1i}, \tag{3.1.7}$$

where the $\tilde{\epsilon}^*_{1i}$ are independent random variables with mean 0 and variance $\hat{\sigma}^r_{11} - (\hat{\sigma}^r_{12})^2/\hat{\sigma}^r_{22}$.

(ii) If y_{1i} is observed but y_{2i} is missing, we use

$$\epsilon^*_{2i} = \frac{\hat{\sigma}^r_{12}}{\hat{\sigma}^r_{11}} (y_{1i} - \hat{\boldsymbol{\beta}}_1^{r\top} \boldsymbol{x}_i) + \tilde{\epsilon}^*_{2i}, \tag{3.1.8}$$

where the $\tilde{\epsilon}^*_{2i}$ are independent random variables with mean 0 and variance $\hat{\sigma}^r_{22} - (\hat{\sigma}^r_{12})^2/\hat{\sigma}^r_{11}$.

(iii) If y_{1i} and y_{2i} are both missing, the ϵ^*_{1i} and ϵ^*_{2i} are independently distributed with mean 0 and covariance matrix :

$$\begin{pmatrix} \hat{\sigma}^r_{11} & \hat{\sigma}^r_{12} \\ \hat{\sigma}^r_{12} & \hat{\sigma}^r_{22} \end{pmatrix} = \sum_{i\in S} w_i r_{1i} r_{2i} \begin{pmatrix} e^2_{1i} & e_{1i} e_{2i} \\ e_{1i} e_{2i} & e^2_{2i} \end{pmatrix} / \sum_{i\in S} w_i r_{1i} r_{2i}.$$

Shao and Wang (2002) showed that $\hat{R}_{12,I}$ based on the above joint imputation procedures is asymptotically unbiased for R_{12}, provided that the bivariate imputation model (3.1.2) is correctly specified. This property holds whether or not

the bivariate distribution of the variables y_1 and y_2 is symmetric. In other words, the Shao-Wang procedure does not make any assumption about the distribution of the error terms. In particular, it doesn't require normally distributed errors terms. Note that marginal random regression imputation is obtained from the above imputation procedure by setting $\hat{\sigma}_{12}^r = 0$ in (i)-(iii). One drawback of the Shao-Wang procedure is that it introduces an additional amount of variability due to the random selection of residuals. As a result, the imputed estimator $\hat{R}_{12,I}$ is potentially inefficient. Chauvet and Haziza (2012) proposed a balanced version of the Shao-Wang procedure, which consists of selecting the residuals ϵ_{1i}^* and ϵ_{2i}^* at random so that the imputation error, $\hat{R}_{12,I} - \check{R}_{12,I}$, is (approximately) equal to zero. That is, the residuals are selected at random so that the following balancing constraints are satisfied:

$$\hat{t}_I^{kl} - \check{t}_I^{kl} = 0 \tag{3.1.9}$$

for $(k,l) \in \{(1,0),(2,0),(1,1),(0,1),(0,2)\}$, where $\check{t}_I^{kl} = E_I(\hat{t}_I^{kl})$. Balanced imputation can be implemented by adapting the Cube algorithm originally developed by Deville and Tillé (2004) in the context of balanced sampling.

3.2 The MIVQUE approach

In the Gauss-Markov approach of estimation of expectation parameters, Best Linear Unbiased Estimators (BLUEs) are defined as the unbiased linear combinations of observed values with minimum variance. Similarly, for variance parameters, the class of Gauss-Markov estimators can be defined as the unbiased quadratic forms of the observed values with minimum variance, also called Minimum Variance Quadratic Unbiased Estimator (MIVQUE); see Rao, 1970, 1971a, 1971b for details.

3.2.1 MIVQUE through the bivariate case with fully observed covariates

We consider here an arbitrary bivariate case of nonresponse pattern on two variables of interest (here, y_1 and y_2) with auxiliary variables. For any combination c of variables and $j \in c$, we denote by $\mathbf{y}_j^{(c)}$ the n_c-vector of observed values for the j-th variable on the sample $\mathcal{S}_c$ for which only the variables with indices in c are observed. In the Gauss-Markov approach described by Causeur (2006), the class of quadratic estimators of the variance parameters is defined as the set of quadratic forms of $\breve{\mathbf{y}} = (\mathbf{y_1^{(1)\top}}, \mathbf{y_2^{(2)\top}}, \mathbf{y_1^{(12)\top}}, \mathbf{y_2^{(12)\top}})^\top$. In what follows, $\mathbf{x}^{(c)}$ and $\mathbf{w}^{(c)}$ denote respectively the $n_c \times p$ matrix of the observed values of the covariates in $\mathcal{S}_c$ and the $n_c \times n_c$ diagonal matrix, whose l-th diagonal element is w_l for l in $\mathcal{S}_c$.

The first two moments of $\breve{\mathbf{y}}$ are therefore given in their partitioned form as follows:

$$E_m(\breve{\mathbf{y}}) = \mathbf{X}\boldsymbol{\beta},\ V_m(\breve{\mathbf{y}}) = \mathbf{V},$$

where

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}^{(1)} & \mathbf{0}_{n_1} \\ \mathbf{0}_{n_2} & \mathbf{x}^{(2)} \\ \mathbf{x}^{(12)} & \mathbf{0}_{n_{12}} \\ \mathbf{0}_{n_{12}} & \mathbf{x}^{(12)} \end{pmatrix},\ \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \text{ and } \mathbf{V} = \left(\begin{array}{c|c|cc} \sigma_{11}\mathbf{I}_{n_1} & \mathbf{0}_{n_1,n_2} & \mathbf{0}_{n_1,n_{12}} & \mathbf{0}_{n_1,n_{12}} \\ \hline \mathbf{0}_{n_2,n_1} & \sigma_{22}\mathbf{I}_{n_2} & \mathbf{0}_{n_2,n_{12}} & \mathbf{0}_{n_2,n_{12}} \\ \hline \mathbf{0}_{n_{12},n_1} & \mathbf{0}_{n_{12},n_2} & \sigma_{11}\mathbf{I}_{n_{12}} & \sigma_{12}\mathbf{I}_{n_{12}} \\ \mathbf{0}_{n_{12},n_1} & \mathbf{0}_{n_{12},n_2} & \sigma_{12}\mathbf{I}_{n_{12}} & \sigma_{22}\mathbf{I}_{n_{12}} \end{array}\right).$$

In the above expressions, $\mathbf{I}_{n_c}$, $\mathbf{0}_{n_c}$ and $\mathbf{0}_{n_c,n_{c'}}$ denote respectively the $n_c \times n_c$ identity matrix, the $n_c \times p$ matrix consisting of zero entries and the $n_c \times n_{c'}$ matrix consisting of zero entries.

A linear Gauss-Markov estimator $\hat{\theta}$ of any linear contrast $\theta = \boldsymbol{\lambda}^\top\boldsymbol{\beta}$ is defined as follows: $\hat{\theta} = \boldsymbol{\ell}^\top\breve{\mathbf{y}}$, where $\boldsymbol{\ell}$ is chosen so that $\hat{\theta}$ is unbiased with minimum variance in the class of linear unbiased estimators. This optimization issue leads to the well-known general least-squares solution:

$$\boldsymbol{\ell} = \mathbf{V^{-1}X}(\mathbf{X}^\top\mathbf{V^{-1}X})^{-1}\boldsymbol{\lambda},$$

which depends on the variance parameters through $\mathbf{V}$.

A weighted version of $\hat{\boldsymbol{\theta}}$ can be obtained by replacing $\breve{\mathbf{y}}$ by $\mathbf{W^{1/2}y}$ and $\mathbf{X}$ by $\mathbf{W^{1/2}X}$ in the above expressions, where the weighting matrix $\boldsymbol{W}$ is defined as

$$\mathbf{W} = \left(\begin{array}{c|c|cc} \mathbf{w^{(1)}} & \mathbf{0}_{n_1,n_2} & \mathbf{0}_{n_1,n_{12}} & \mathbf{0}_{n_1,n_{12}} \\ \hline \mathbf{0}_{n_2,n_1} & \mathbf{w^{(2)}} & \mathbf{0}_{n_2,n_{12}} & \mathbf{0}_{n_2,n_{12}} \\ \hline \mathbf{0}_{n_{12},n_1} & \mathbf{0}_{n_{12},n_2} & \mathbf{w^{(12)}} & \mathbf{0}_{n_{12},n_{12}} \\ \mathbf{0}_{n_{12},n_1} & \mathbf{0}_{n_{12},n_2} & \mathbf{0}_{n_{12},n_{12}} & \mathbf{w^{(12)}} \end{array}\right).$$

The corresponding Gauss-Markov estimator $\hat{\boldsymbol{\theta}}$ is given by $\hat{\boldsymbol{\theta}} = \boldsymbol{\ell}_{\boldsymbol{w}}^{\top}\breve{\mathbf{y}}$, where $\boldsymbol{\ell}_{\boldsymbol{w}} = \mathbf{W^{-1/2}}\boldsymbol{\ell}$. The weighted version will be able to handle survey regression imputation for which the weights are defined as the inverse of the inclusion probabilities. Note that the Shao-Wang procedure makes use of the survey weights when estimating the variance and covariance parameters; see Section 2.

The vector of variance parameters $\boldsymbol{\sigma} = (\sigma_{11}, \sigma_{22}, \sigma_{12})^{\top}$ is now estimated using similar Gauss-Markov techniques in a quadratic estimation framework. First, in order to ensure the invariance of the estimation of any linear contrast $\theta = \boldsymbol{\lambda}^{\top}\boldsymbol{\sigma}$ of the variance parameters with respect to translation on the mean parameters, most authors (see Rao and Kleffe, 1988 for a detailed review) suggest to define quadratic estimators as quadratic forms $\hat{\theta} = \breve{\mathbf{y}}_{\boldsymbol{x}}^{\top}\boldsymbol{A}\breve{\mathbf{y}}_{\boldsymbol{x}}$ of the linear projection $\breve{\mathbf{y}}_{\boldsymbol{x}} = (\boldsymbol{I}_{\boldsymbol{n_y}} - \mathbf{X}(\mathbf{X}^{\top}\mathbf{X})^{-1}\mathbf{X}^{\top})\breve{\mathbf{y}}$ of $\breve{\boldsymbol{y}}$ onto the linear subspace orthogonal to the linear subspace spanned by $\mathbf{X}$. Causeur (2006) showed that the MIVQUE estimators belong to a complete subclass of the quadratic estimators which is described by the linear combinations of all possible cross-products of the variables on the subsamples defined by each missingness pattern.

After restriction to unbiased estimators, the general expression used for quadratic unbiased estimators of variance parameters in the bivariate situation introduced above is:

$$\hat{\boldsymbol{\sigma}}_{\boldsymbol{jj'}} = \tilde{\boldsymbol{\sigma}}_{\boldsymbol{jj'}} + \boldsymbol{\alpha}_{\mathbf{1}}^{\boldsymbol{(jj')}}[\tilde{\boldsymbol{\sigma}}_{\mathbf{11}}^{\mathbf{(12)}} - \tilde{\boldsymbol{\sigma}}_{\mathbf{11}}] + \boldsymbol{\alpha}_{\mathbf{2}}^{\boldsymbol{(jj')}}[\tilde{\boldsymbol{\sigma}}_{\mathbf{22}}^{\mathbf{(12)}} - \tilde{\boldsymbol{\sigma}}_{\mathbf{22}}], \tag{3.2.1}$$

where $\tilde{\sigma}_{jj'}$ is the empirical estimator of $\sigma_{jj'}$ computed on the largest sample for which y_j and $y_{j'}$ are jointly observed, $\tilde{\sigma}^{(c)}_{jj'}$ is the empirical estimator of $\sigma_{jj'}$ computed on $\mathcal{S}_c$ and $\alpha_l^{(jj')}$ are unknown constants (Causeur 2006). Thus, the unbiased estimators are additively corrected versions of the empirical estimator, which additive correction is optimally adjusted so that the resulting estimator has the minimum variance.

The MIVQUEs are obtained for each variance parameters, by replacing the $\alpha_l^{(jj')}$ in expression (3.2.1) of $\hat{\sigma}_{jj'}$ by the coefficients providing the estimator with minimum variance. In the present situation of only two variables prone to missing values, the optimal coefficients in the case of equal weights ($w_i = 1/n$) are given by

$$\begin{aligned}
\begin{pmatrix} \alpha_1^{(11)} \\ \alpha_2^{(11)} \end{pmatrix} &= -\frac{\gamma_{2.2}^2 \frac{f_{1.2}}{1+f_{1.2}}}{1-\gamma_{1.2}^2\gamma_{2.2}^2\frac{1}{1+f_{1.2}}\frac{1}{1+f_{2.2}}} \begin{pmatrix} -\gamma_{1.2}^2\frac{1}{1+f_{2.2}} \\ 1 \end{pmatrix}, \\
\begin{pmatrix} \alpha_1^{(22)} \\ \alpha_2^{(22)} \end{pmatrix} &= -\frac{\gamma_{1.2}^2 \frac{f_{2.2}}{1+f_{2.2}}}{1-\gamma_{1.2}^2\gamma_{2.2}^2\frac{1}{1+f_{1.2}}\frac{1}{1+f_{2.2}}} \begin{pmatrix} 1 \\ -\gamma_{2.2}^2\frac{1}{1+f_{1.2}} \end{pmatrix}, \\
\begin{pmatrix} \alpha_1^{(12)} \\ \alpha_2^{(12)} \end{pmatrix} &= -2\frac{1}{1-\gamma_{1.2}^2\gamma_{2.2}^2\frac{1}{1+f_{1.2}}\frac{1}{1+f_{2.2}}} \begin{pmatrix} \gamma_{1.2}-\gamma_{2.2}\gamma_{1.2}^2\frac{1}{1+f_{2.2}} \\ \gamma_{2.2}-\gamma_{1.2}\gamma_{2.2}^2\frac{1}{1+f_{1.2}} \end{pmatrix},
\end{aligned}$$

where $\gamma_{1.2} = \sigma_{12}/\sigma_{11}$, $\gamma_{2.2} = \sigma_{12}/\sigma_{22}$, $f_{1.2} = n_{12}/n_1$ and $f_{2.2} = n_{12}/n_2$.

As shown in Causeur (2006), in the present missing data issue, no uniformly optimal estimator can be obtained for the variance parameters. This explains why the above locally MIVQUE depends itself on the variance parameters. In the following, we introduce the iterated MIVQUE procedure, which takes advantage of the explicit expressions of locally MIVQUE to define an estimating algorithm.

3.2.2 Iterated version of MIVQUE

Let Σ_0 denote a known $q \times q$ positive definite symmetric matrix and MIVQUE(Σ, Σ_0) be the $q \times q$ symmetric matrix, whose element (i, j) is MIVQUE(σ_{ij}, Σ_0), the local MIVQUE under the hypothesis $\Sigma = \Sigma_0$. For instance, when $\Sigma_0 = I_q$,

$\widehat{\Sigma}_1 = \widetilde{\Sigma} = MIVQUE(\Sigma, \Sigma_0)$ is the empirical variance-covariance estimator of Σ. Therefore, this choice for Σ_0 seems to be a natural starting point for estimating Σ.

In order to reduce arbitrariness of this starting point, the preceding procedure can be iterated, leading to a sequence $(\widehat{\Sigma}_d)_{d\geq 0}$ of estimators of Σ defined by the recurrence relation:

$$\widehat{\Sigma}_d = MIVQUE(\Sigma, \widehat{\Sigma}_{d-1}), d \geq 1.$$

Though MIVQUE and MLE are technically different estimators, Harville (1977) bridged the gap between the two by showing that the MLE could be viewed as an infinitely iterated version of MIVQUE. As a difference between the EM and iterated MIVQUE approaches, it may be noted that the latter is a coordinate-free approach which does not belong to the class of data augmentation algorithms for which the imputation step can turn out to be sensitive. In a linear regression framework and small-sample conditions, Causeur (2006) showed that once or twice iterated MIVQUE can show marked improvements with respect to MLE.

3.3 MIVQUE based imputation

In practice, some form of calibration is used in virtually all the medium to large scale surveys. Calibration consists of modifying the sampling (initial) weights so that survey estimates of totals coincide with true, known population totals (also called benchmarks) from external sources. The interested reader is referred to Deville and Särndal (1992), Särndal (2007) and Kim and Park (2010), among other for excellent discussions on calibration. In this section, we use the idea of calibration to modify initial imputed values rather than initial weights so that appropriate calibration constraints are satisfied.

We propose a calibrated version of the Shao-Wang procedure, which consists of two distinct steps :

(1) Use the Shao-Wang imputation procedure (see Section 2) and obtain the initial imputed values y_{1i}^* and y_{2i}^*.

(2) Determine final imputed values by modifying the initial values obtained in Step (1) so that the following calibration constraints are satisfied:

$$\begin{aligned} \hat{t}_I^{10} &= N\ \hat{\mu}_1 \\ \hat{t}_I^{01} &= N\ \hat{\mu}_2 \\ \hat{t}_I^{20} &= N\ (\hat{\mu}_1)^2 + (N-1)\ \hat{\sigma}_{11} \\ \hat{t}_I^{02} &= N\ (\hat{\mu}_2)^2 + (N-1)\ \hat{\sigma}_{22} \\ \hat{t}_I^{11} &= N\ \hat{\mu}_1\ \hat{\mu}_2 + (N-1)\ \hat{\sigma}_{12}, \end{aligned} \tag{3.3.1}$$

where $\hat{\mu}_1$, $\hat{\mu}_2$, $\hat{\sigma}_{11}$, $\hat{\sigma}_{22}$ and $\hat{\sigma}_{12}$ denote, respectively, iterated MIVQUE of μ_1, μ_2, σ_{11}, σ_{22} and σ_{12}. Let $F(.)$ be a monotonic and twice differentiable function satisfying $F(0) = 1$ and $F'(0) > 0$. The function $F(.)$ is the so-called calibration function; e.g., Deville and Särndal (1992). The final imputed values $\hat{y}_{1i}$ and $\hat{y}_{2i}$ are defined as

$$\begin{aligned} \hat{y}_{1i} &= y_{1i}^* F(\lambda_1 + \lambda_2 y_{1i}^* + \lambda_3 y_{2i}^*) \\ \hat{y}_{2i} &= y_{2i}^* F(\lambda_4 + \lambda_5 y_{2i}^* + \lambda_3 y_{1i}^*), \end{aligned} \tag{3.3.2}$$

where the coefficients $\lambda_1, \ldots, \lambda_5$ are determined so that the calibration constraints (3.3.1) are satisfied. From (3.3.2), the final imputed values are expressed as the product of the initial value y_{ki}^*, $k = 1, 2$, and an adjustment factor. In the calibration literature, several choices of $F(.)$ are available; see e.g., Deville and Särndal (1992). For example, on may use the linear method, which corresponds to the calibration function $F(u) = 1 + u$. However, the latter may produce negative imputed values. For this reason, we prefer using the exponential function, $F(u) = \exp(u)$, which is frequently

utilized in the context of weighting in surveys. Solving for the coefficients $\lambda_1, \ldots, \lambda_5$ may be done using the Newton-Raphson algorithm with initial values $\lambda_1 = \ldots = \lambda_5 = 0$.

The estimators $\hat{\mu}_1$, $\hat{\mu}_2$, $\hat{\sigma}_{11}$, $\hat{\sigma}_{22}$ and $\hat{\sigma}_{12}$ are all unbiased for their corresponding parameter provided that model (3.1.2) holds. This is true even if the bivariate distribution of the variables y_1 and y_2 is not symmetric. From the calibration constraints (3.3.1), it follows that

$$B_{qmI}(\hat{t}_I^{kl}) = 0 \tag{3.3.3}$$

for $(k,l) \in \{(1,0),(2,0),(1,1),(0,1),(0,2)\}$. Once again, the bias in (3.3.3) is equal to zero regardless of the bivariate distribution of the variables y_1 and y_2. Since the coefficient of correlation R_{12} can be expressed as a smooth function of t^{kl} for $(k,l) \in \{(1,0),(2,0),(1,1),(0,1),(0,2)\}$, the imputed estimator $\hat{R}_{12,I}$ is asymptotically unbiased for R_{12} and this property holds whether or not the bivariate distribution is symmetric.

When the bivariate distribution of y_1 and y_2 is symmetric, we expect the proposed imputation procedure to be significantly more efficient than the Shao-Wang procedure. We now explain why this is the case. As mentioned in Section 2 , the total error of $\hat{R}_{12,I}$ can be expressed as the sum of three terms: the sampling error, the nonresponse error and the imputation error; see expression (3.1.3). While the sampling error depends on the finite population U under study, the sampling design used to select the sample S and the sample size n, it does not depend on nonresponse and imputation. Therefore, nothing can be done at the imputation stage about the sampling error and the latter is identical for both the Shao and Wang procedure and the proposed procedure. On the other hand, the nonresponse error depends on the response rate and the predictive power of the imputation model (3.1.2). Finally, the imputation error, $\hat{R}_{12,I} - \check{R}_{12,I}$, in (3.1.3) vanishes under the proposed procedure, unlike the Shao-Wang procedure. The

calibration constraints (3.3.1) ensure that, conditionally on the sample and the set of respondents, the imputed estimator $\hat{R}_{12,I}$ always takes the same value if the imputation process is repeated. Therefore, the proposed procedure does not suffer from the imputation variance, which makes it fully efficient, a term coined by Kim and Fuller (2004). In contrast, $\hat{R}_{12,I}$ exhibits some variability under the Shao-Wang procedure, due to the random selection of residuals. Finally, note that the MIVQUE estimators are designed to perform well in terms of mean square error when the distribution of the variables is symmetric or near symmetric. In this case, calibrating on the MIVQUE estimator leads to efficient estimators of t^{kl} for $(k,l) \in \{(1,0),(2,0),(1,1),(0,1),(0,2)\}$. On the other hand, if the distribution is not symmetric, the MIVQUE estimator may be unstable, which in turns, leads to inefficient estimators of t^{kl}. This is confirmed empirically in the next section.

3.4 Simulation study

We conducted a simulation study in order to assess the proposed method in terms of relative bias and relative efficiency. We performed 5000 iterations of the following process. First, a finite population was generated; we considered the case of a census. Then, nonresponse in the population was generated and missing values were imputed. Below, we describe one such iteration in further details.

We generated a finite population of size $N = 1000$ consisting of two study variables, y_1 and y_2, and an auxiliary variable x. The x-values were first generated according to a Gamma distribution. Given the x-values, N values of $(y_1, y_2)^\top$ were generated according to the following bivariate model:

$$y_{1i} = \beta_1 x_i + \varepsilon_{i1},$$
$$y_{2i} = \beta_2 x_i + \varepsilon_{i2},$$

where $\beta_1 = \beta_2 = 1$ and the error terms ε_{1i} and ε_{2i} were independently generated

Obs.	Non response pattern		
	y_1	y_2	x
1	?	?	x_1
2	?	?	x_2
3	y_{13}	?	x_3
4	y_{14}	?	x_4
5	?	y_{25}	x_5
6	?	y_{26}	x_6
7	y_{17}	y_{27}	x_7
8	y_{18}	y_{28}	x_8
...	...	...	...
N	y_{1N}	y_{2N}	x_N

a - Initial missing values are indicated by the symbol '?'

$\Rightarrow$ 1

Obs.	First imputation		
	y_1	y_2	x
1	y^*_{11}	y^*_{21}	x_1
2	y^*_{12}	y^*_{22}	x_2
3	y_{13}	y^*_{23}	x_3
4	y_{14}	y^*_{24}	x_4
5	y^*_{15}	y_{25}	x_5
6	y^*_{16}	y_{26}	x_6
7	y_{17}	y_{27}	x_7
8	y_{18}	y_{28}	x_8
...	...	...	...
N	y_{1N}	y_{2N}	x_N

b - First imputed values are indicated by y^*_{ji}

$\Rightarrow$ 2

Obs.	Calibration on MIVQUE		
	y_1	y_2	x
1	$\hat{y}_{11}$	$\hat{y}_{21}$	x_1
2	$\hat{y}_{12}$	$\hat{y}_{22}$	x_2
3	y_{13}	$\hat{y}_{23}$	x_3
4	y_{14}	$\hat{y}_{24}$	x_4
5	$\hat{y}_{15}$	y_{25}	x_5
6	$\hat{y}_{16}$	y_{26}	x_6
7	y_{17}	y_{27}	x_7
8	y_{18}	y_{28}	x_8
...	...	...	...
N	y_{1N}	y_{2N}	x_N

c - Final imputed values are indicated by $\hat{y}_{ji}$

Table 3.3.1: Example of two steps MIVQUE calibrated imputation

according to

$$\varepsilon_{1i} = \kappa \times \chi_i + \kappa_1 \times \nu_i,$$

$$\varepsilon_{2i} = \kappa \times \chi_i + \kappa_2 \times \varsigma_i,$$

with χ_i, ν_i and ς_i denoting error terms independently generated according to a normal distribution with mean 0 and variance 1, and κ, κ_1 and κ_2 are parameters. A large value of κ corresponds to a large value of the coefficient of correlation between y_1 and y_2. The parameters κ_1 and κ_2 are used to control the skewness and kurtosis of y_1 and y_2, respectively.

The p-values showed in Table 3.4.1 come from a multivariate normality test. This test uses skewness measured by two different location estimates as described in Kankainen et al. (2007). This test, implemented with the R package ICS, is based on the regular mean vector and the location estimate based on third moments.

We generated eight types of finite populations. The degree of asymmetry in each population varied from none to high. Also, in each population, the variables y_1 and y_2 were generated so that the finite population coefficient of correlation R_{12} was either approximately equal to 0.5 or approximately equal to 0.8. This led to eight different populations.

We were interested in estimating five finite population parameters: the population mean of y_1 and y_2, given by $\bar{Y}_1 = t_{10}/N$ and $\bar{Y}_2 = t_{01}/N$, respectively; the variability of y_1 and y_2 in the population given by $S_1^2 = (N-1)^{-1}\left\{t^{20} - (t^{10})^2/N\right\}$ and $S_2^2 = (N-1)^{-1}\left\{t^{02} - (t^{01})^2/N\right\}$, respectively and the finite population coefficient of correlation between y_1 and y_2, R_{12}. Table 3.4.1 shows the Monte Carlo averages of several characteristics for each type of populations.

In order to focus on the nonresponse/imputation error, we considered the case of a census, $n = N = 1000$. Let p_{1i} and p_{2i} be the response probabilities for unit

	Symmetric		Low asymmetry		Medium asymmetry		Stronger asymmetry	
R_{12}	0.5	0.8	0.5	0.8	0.5	0.8	0.5	0.8
Population	1	2	3	4	5	6	7	8
Parameters								
$\bar{Y}_1$	8.0	8.0	8.0	8.0	2.0	2.0	2.0	2.0
$\bar{Y}_2$	8.0	8.0	8.0	8.0	2.0	2.0	2.0	2.0
S_1^2	2.6	2.6	92.1	92.1	10.2	10.3	19.3	19.3
S_2^2	2.6	2.6	92.0	92.2	10.2	10.3	19.4	19.3
R_{12}	0.5	0.8	0.5	0.8	0.5	0.8	0.5	0.8
κ	0.0	0.9	0.0	5.3	0.0	1.7	0.0	2.4
$\kappa_1 = \kappa_2$	1.1	0.7	6.8	4.3	2.3	1.4	3.1	2.0
skewness(y_1)	0.1	0.1	0.6	0.6	0.8	0.8	1.1	1.1
skewness(y_2)	0.1	0.1	0.6	0.6	0.8	0.8	1.1	1.1
Multivariate normality tests (p-values) for (y_1,y_2) based on :								
skewness	0.24	0.33	0.00	0.00	0.00	0.00	0.00	0.00

Table 3.4.1: Average characteristics of the populations and multi-normality test p-values for the generated populations

i to the study variables y_1 and y_2, respectively. We generated nonresponse to y_1 and y_2 according to

$$p_{1i} = \left\{1 + exp\left(\frac{-0.4055}{\overline{X}}x_i\right)\right\}^{-1} \tag{3.4.1}$$

and

$$p_{2i} = \left\{1 + exp\left(\frac{-0.4055}{\overline{X}}x_i\right)\right\}^{-1}, \tag{3.4.2}$$

where $\overline{X} = N^{-1}\sum_{i\in U} x_i$ denotes the population mean of the x-values. The coefficients in (3.4.1) and (3.4.2) were chosen so that the average response rates for the study variables y_1 and y_2 were approximately equal to 60%.

The response indicators r_{1i} and r_{2i} were then generated independently from a Bernoulli distribution with parameter p_{1i} and p_{2i}, respectively, which led to a set of respondents. Then, missing values to y_1 and y_2 were imputed according to (i) the Shao-Wang (SW) procedure and (ii) the calibrated Shao-Wang procedure (CSW).

As a measure of bias of an estimator $\hat{\gamma}$ of a finite population parameter γ, we computed the Monte Carlo percent relative bias

$$RB_{MC}(\hat{\gamma}) = \frac{1}{K}\sum_{k=1}^{K}\frac{(\hat{\gamma}_{(k)} - \gamma)}{\gamma} \times 100,$$

where $\hat{\gamma}_{(k)}$ denotes the estimator estimator $\hat{\gamma}$ in the k-th sample. As a measure of relative efficiency, we computed

$$RE = \frac{MSE_{MC}(\hat{\gamma}_{CSW})}{MSE_{MC}(\hat{\gamma}_{SW})} \times 100,$$

where $\hat{\gamma}_{SW}$ and $\hat{\gamma}_{CSW}$ denote the estimator $\hat{\gamma}$ obtained under the SW and CSW procedures, respectively, and

$$MSE_{MC}(\hat{\gamma}) = \frac{1}{K}\sum_{k=1}^{K}(\hat{\gamma}_{(k)} - \gamma)^2.$$

Tables 3.4.2 and 3.4.3 show the Monte Carlo percent relative bias of several imputed estimators for the populations. In terms of relative bias, both SW and CSW performed well, as expected. Both procedures led to negligible bias in most scenarios regardless of the nature of the distribution (symmetric or asymmetric).

We now turn the relative efficiency shown in Table 3.4.3. For scenarios 1-6 (which corresponds to the populations exhibiting no asymmetry, a low asymmetry or a medium asymmetry) , the proposed CSW procedure was significantly more efficient than the SW procedure with values of relative efficiency ranging from 60% to 75%. When the population exhibited a large degree of asymmetry, the proposed CSW procedure was more efficient than the SW procedure for scenario 7 (which corresponds to a coefficient of correlation equal to 0.5) but was significantly less efficient for scenario 8 (which corresponds to a coefficient of correlation equal to 0.8) with values of relative efficiency ranging from 90% to 250%.. These results suggests that applying the proposed procedure in the case of highly asymmetric distribution may lead to unstable estimators.

	Symmetric				Low asymmetry				Medium asymmetry				Stronger asymmetry			
Population	Number 1		Number 2		Number 3		Number 4		Number 5		Number 6		Number 7		Number 8	
R_{12}	0.5		0.8		0.5		0.8		0.5		0.8		0.5		0.8	
Method	SW	CSW	SW	CSW	SW	CSW	SW	CSW	SW	CSW	SW	CSW	SW	CSW	SW	CSW
Parameters																
$\bar{Y}_1$	0.0%	0.0%	0.0%	0.0%	0.1%	0.1%	0.0%	0.0%	0.0%	0.0%	-0.1%	-0.1%	0.1%	0.1%	-0.0%	0.0%
$\bar{Y}_2$	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	-0.1%	0.0%	0.0%	0.0%	0.1%	0.0%	0.0%	-0.0%
S_1^2	0.0%	0.2%	-0.1%	0.2%	-0.1%	0.2%	-0.1%	0.2%	-0.1%	0.2%	-0.1%	0.3%	-0.2%	0.2%	-0.0%	0.4%
S_2^2	-0.1%	0.2%	0.0%	0.3%	0.0%	0.2%	-0.1%	0.2%	-0.1%	0.2%	-0.1%	0.2%	-0.2%	0.2%	-0.2%	0.1%
R_{12}	0.1%	0.1%	0.0%	0.1%	0.2%	0.2%	0.0%	0.1%	0.2%	0.1%	0.0%	0.1%	0.3%	0.2%	0.1%	0.2%

Table 3.4.2: Monte Carlo percent relative bias of several parameters under SW and CSW procedures (in %)

	Symmetric		Low asymmetry		Medium asymmetry		Stronger asymmetry	
Population	Number 1	Number 2	Number 3	Number 4	Number 5	Number 6	Number 7	Number 8
R_{12}	0.5	0.8	0.5	0.8	0.5	0.8	0.5	0.8
Parameters								
$\bar{Y}_1$	64%	61%	68%	67%	71%	71%	69.3%	92.3%
$\bar{Y}_2$	62%	60%	67%	67%	69%	71%	69.0%	91.8%
S_1^2	72%	70%	63%	62%	63%	61%	57.7%	228.6%
S_2^2	72%	69%	62%	63%	61%	60%	57.4%	246.0%
R_{12}	75%	75%	71%	75%	74%	75%	72.6%	250.0%

Table 3.4.3: Relative efficiency (in %)

3.5 Discussion

In this paper, we proposed a calibrated version of the Shao-Wang procedure. We showed empirically that the proposed procedure leads to an asymptotically unbiased estimator of a coefficient of correlation and is much more efficient than the Shao-Wang procedure when the underlying distribution of the variables being imputed is symmetric or near symmetric. In this paper, we considered the case of two variables requiring imputation. In practice, we may want to preserve the relationship between more than two items. Shao and Wang (2002) extended their method to handle this situation. Causeur (2006) derived MIVQUE in a multivariate setting. Therefore, our procedure can be extended in a relatively straightforward fashion to the case of more than two items requiring imputation.

Variance estimation in the presence of imputed values is an important problem as naive variance estimators (which are those computed by treating the imputed values as observed values) tend to underestimate the true variance of point estimators, which is turns leads to confidence intervals that are too narrow. Because of the complexity of the proposed procedure, methods relying on first-order Taylor expansions are virtually infeasible. If the overall sampling fraction is small, one may use the bootstrap procedure of Shao and Sitter (1996), which consists of selecting repeated samples from the population and imputing the nonrespondents in each bootstrap samples using the same procedure that was used in the original sample. Bootstrap variance estimation in the case of nonnegligible sampling fractions is currently under investigation.

Appendix A : Properties of imputed estimators of first and second moments under marginal imputation

We assume that the following imputation model holds for the responding units:

$$y_{1i} = \boldsymbol{x}_i^\top \boldsymbol{\beta}_1 + \epsilon_{1i}. \tag{3.5.1}$$

where $\boldsymbol{x}_i$ is a vector of auxiliary variables attached to unit i available for all $i \in S$ and $\boldsymbol{\beta}_1$ is a vector of unknown parameters. We make the usual assumptions:

$$E_m(\epsilon_{1i}) = 0, \quad E_m(\epsilon_{1i}\epsilon_{1j}) = 0 \quad for \;\; i \neq j \text{ and } \quad V_m(\epsilon_{1i}) = \sigma_{11}$$

where σ_{11} is an an unknown parameter and the subscript m denotes model (3.5.1).

A.1 Deterministic marginal imputation

In the case of deterministic marginal imputation, missing y_{1i} is imputed by $y_{1i}^* = \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_1^r$ where $\widehat{\boldsymbol{\beta}}_1^r$ is obtained from (3.1.5). It follows that

$$E_m(y_{1i}^*) = \boldsymbol{x}_i^\top \boldsymbol{\beta}_1 \tag{3.5.2}$$

$$V_m(\hat{\boldsymbol{\beta}}_{\mathbf{1}}{}^r) = \sigma_{11}\hat{\mathbf{T}}_r^{-1}\left(\sum_{i\in S} w_i^2 r_i \boldsymbol{x}_i \boldsymbol{x}_i^\top\right)\hat{\mathbf{T}}_r^{-1}, \tag{3.5.3}$$

where $\hat{\mathbf{T}}_r^{-1} = \sum_{i\in S} w_i r_i \boldsymbol{x}_i \boldsymbol{x}_i^\top$.

A.1.1 Imputed estimator of the first moment

The imputed estimator for the marginal first moment t^{10} is

$$\hat{t}_I^{10} = \sum_{i\in S} w_i r_i y_{1i} + \sum_{i\in S} w_i (1 - r_i) y_{1i}^*.$$

Using (3.5.2), the nonresponse error of $\hat{t}_I^{10}$ is given by

$$E_m(\hat{t}_I^{10} - \hat{t}_\pi^{10}) = -\sum_{i\in S} w_i (1 - r_i) E_m(y_{1i} - \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_{\mathbf{1}}^{\boldsymbol{r}}) = 0.$$

Consequently, marginal deterministic regression imputation leads to an unbiased estimator of the marginal first moment, provided that model (3.5.1) holds.

A.1.2 Imputed estimator of the second moment

The imputed estimator of the marginal second moment t^{20} is given by

$$\begin{aligned}\hat{t}_I^{20} &= \sum_{i\in S} w_i r_i y_{1i}^2 + \sum_{i\in S} w_i(1-r_i)(y_{1i}^*)^2 \\ &= \sum_{i\in S} w_i r_i y_{1i}^2 + \sum_{i\in S} w_i(1-r_i)(\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_1^r)^2.\end{aligned}$$

It follows that

$$\begin{aligned}E_m(\hat{t}_I^{20}) &= \sum_{i\in S} w_i r_i \left\{\sigma_{11} + (\boldsymbol{x}_i^\top \boldsymbol{\beta}_1)^2\right\} + \sum_{i\in S} w_i(1-r_i)\left\{V_m(\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_1^r) + (\boldsymbol{x}_i^\top \boldsymbol{\beta}_1)^2\right\} \\ &= \sum_{i\in S} w_i(\boldsymbol{x}_i^\top \boldsymbol{\beta}_1)^2 + \sigma_{11}\sum_{i\in S} w_i r_i \\ &\quad + \sigma_{11}\sum_{i\in S} w_i(1-r_i)\boldsymbol{x}_i^\top \hat{\mathbf{T}}_r^{-1}\left(\sum_{i\in S} w_i^2 r_i \boldsymbol{x}_i \boldsymbol{x}_i^\top\right)\hat{\mathbf{T}}_r^{-1}\boldsymbol{x}_i. \qquad (3.5.4)\end{aligned}$$

Assuming that (i)$\max(w_i) = O(N/n)$ and (ii) $\sum_{i\in s} w_i / \sum_{i\in s} w_i r_i = O_p(1)$, the third term in the right hand side of (3.5.4) is of lower order of magnitude than the two first terms. We have

$$E_m(\hat{t}_I^{20}) = \sum_{i\in S} w_i(\boldsymbol{x}_i^\top \boldsymbol{\beta_1})^2 + \sigma_{11}\sum_{i\in S} w_i r_i + O_p(N/n_r),$$

where n_r denotes the number of respondents to item y_1.

Noting that $\hat{t}_\pi^{20} = \sum_{i\in S} w_i y_{1i}^2$ and ignoring the higher order terms, we obtain

$$E_m(\hat{t}_I^{20} - \hat{t}_\pi^{20}) \approx -\sigma_{11}\sum_{i\in S} w_i(1-r_i),$$

which is not negligible. This result shows that deterministic marginal regression leads generally to asymptotically biased estimator of the marginal second moment t^{20}.

A.2 Marginal random regression imputation

In the case of marginal random imputation, missing y_{1i} is imputed by

$$y_{1i}^{*} = \boldsymbol{x}_i^{\top}\widehat{\boldsymbol{\beta}}_1^{r} + \epsilon_{1i}^{*},$$

where the residuals ϵ_{1i}^{*} are generated so that

$$\begin{aligned} E_I(\epsilon_{1i}^{*}) &= 0, && (3.5.5)\\ V_I(\epsilon_{1i}^{*}) &= \hat{\sigma}_{11} = \frac{1}{\sum_{i\in S} w_i r_i}\sum_{i\in S} w_i r_i (y_{1i} - \boldsymbol{x}_i^{\top}\widehat{\boldsymbol{\beta}}_1^{r}), && (3.5.6)\\ Cov_I(\epsilon_{1i}^{*}, \epsilon_{1k}^{*}) &= 0 \quad for \quad i \neq j. && (3.5.7) \end{aligned}$$

A.2.1 Imputed estimator for the first moment

The imputed estimator for the marginal first moment t^{10} is

$$\hat{t}_I^{10} = \sum_{i\in S} w_i r_i y_{1i} + \sum_{i\in S} w_i (1-r_i)\left(\boldsymbol{x}_i^{\top}\widehat{\boldsymbol{\beta}}_1^{r} + \epsilon_{1i}^{*}\right).$$

Using (3.5.5), we obtain

$$E_I(\hat{t}_I^{10}) = \sum_{i\in S} w_i r_i y_{1i} + \sum_{i\in S} w_i (1-r_i)\boldsymbol{x}_i^{\top}\widehat{\boldsymbol{\beta}}_1^{r}.$$

Using (3.5.2) and noting that $E_I(\hat{t}_\pi^{10}) = \hat{t}_\pi^{10}$, we obtain

$$\begin{aligned} E_m E_I(\hat{t}_I^{10} - \hat{t}_\pi^{10}) &= E_m\left\{\sum_{i\in S} w_i r_i y_{1i} + \sum_{i\in S} w_i (1-r_i)\boldsymbol{x}_i^{\top}\widehat{\boldsymbol{\beta}}_1^{r} - \hat{t}_\pi^{10}\right\}\\ &= E_m\left\{\sum_{i\in S} w_i (1-r_i)(\boldsymbol{x}_i^{\top}\widehat{\boldsymbol{\beta}}_1^{r} - y_{1i})\right\}\\ &= \sum_{i\in S} w_i (1-r_i) E_m(\boldsymbol{x}_i^{\top}\widehat{\boldsymbol{\beta}}_1^{r} - y_{1i})\\ &= 0 \end{aligned}$$

As a result, marginal random regression imputation preserves the marginal first moment t^{10}, provided that model (3.5.1) holds.

A.2.2 Imputed estimator for the second moment

The imputed estimator of the second moment t^{20} is

$$\hat{t}_I^{20} = \sum_{i \in S} w_i r_i y_{1i}^2 + \sum_{i \in S} w_i (1 - r_i)(\boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_1^r + \epsilon_{1i}^*)^2.$$

Now,

$$\begin{aligned} E_I(\hat{t}_I^{20}) &= \sum_{i \in S} w_i r_i y_{1i}^2 + \sum_{i \in S} w_i (1 - r_i) \left\{ (\boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_1^r)^2 + E_I(\epsilon_{1i}^{*2}) + 2\boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_1^r E_I(\epsilon_{1i}^*) \right\} \\ &= \sum_{i \in S} w_i r_i y_{1i}^2 + \sum_{i \in S} w_i (1 - r_i) \{ (\boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_1^r)^2 + \hat{\sigma}_{11} \}, \end{aligned}$$

On the other hand, we have

$$\begin{aligned} E_m(\hat{\sigma}_{11}) &= \sigma_{11} + \frac{1}{\sum_{i \in S} w_i r_i} \left\{ \sum_{i \in S} w_i r_i \boldsymbol{x}_i^\top \hat{\mathbf{T}}_r^{-1} \sigma_{11} \left(\sum_{i \in S} w_i^2 r_i \boldsymbol{x}_i^\top \right) \hat{\mathbf{T}}_r^{-1} \boldsymbol{x}_i \right. \\ &\quad \left. -2 \sum_{i \in S} w_i^2 r_i \boldsymbol{x}_i^\top \hat{\mathbf{T}}_r^{-1} \boldsymbol{x}_i \sigma_{11} \right\} \\ &= \sigma_{11} + O_p(1/n_r). \end{aligned} \tag{3.5.8}$$

Ignoring higher-order terms, we have:

$$\begin{aligned} E_m E_I(\hat{t}_I^{20}) &\approx \sum_{i \in S} w_i r_i \left\{ (\boldsymbol{x}_i^\top \boldsymbol{\beta}_1)^2 + \sigma_{11} \right\} + \sum_{i \in S} w_i (1 - r_i) \left\{ (\boldsymbol{x}_i^\top \boldsymbol{\beta}_1)^2 + \sigma_{11} \right\} \\ &= \sum_{i \in S} w_i \left\{ (\boldsymbol{x}_i^\top \boldsymbol{\beta}_1)^2 + \sigma_{11} \right\} \\ &= E_m E_I(\hat{t}_\pi^{20}), \end{aligned}$$

which completes the proof. As a result, marginal random regression imputation asymptotically preserves the marginal second moment t^{20}, provided that model (3.5.1) holds.

References

Beaumont, J. F. (2005). Calibrated imputation in surveys under a quasi-model-assisted approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67, 445–458.

Causeur, D. (2006). MIVQUE and maximum likelihood estimation for multivariate linear models with incomplete observations. *Sankhya: The Indian Journal of Statistics*, 409–435.

Chauvet, G., Deville, J. C., and Haziza, D. (2011). On balanced random imputation in surveys, *Biometrika*, 98, 459–471.

Chauvet G. and Haziza D. (2012). Fully efficient estimation of coefficients of correlation in the presence of imputed survey data, *The Canadian Journal of Statistics*, 40, 124–149.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 39, 1–38.

Deville, J. C. (1999). Variance estimation for complex statistics and estimators: linearization and residual techniques, *Survey Methodology*, 25, 193–203.

Deville, J. C. and Tillé, Y. (2004). Efficient balanced sampling: the cube method, *Biometrika*, 91, 893–912.

Favre, A. C., Matei, A., and Tillé, Y. (2005). Calibrated random imputation for qualitative data, *Journal of statistical planning and inference*, 128, 411–425.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems, *Journal of the American Statistical Association*, 72, 320–338.

Haziza, D. (2009). Imputation and inference in the presence of missing data, *Handbook of Statistics*, 29, 215–246.

Kankainen, A., Taskinen, S. and Oja, H. (2007). Tests of multinormality based on location vectors and scatter matrices, *Statistical Methods and Applications*, 16, 357–379.

Kim, J.K. and Fuller, W.A. (2004). Fractional hot-deck imputation. *Biometrika*, 91, 559–578.

Kim, J.K. and Park, M. (2010). Calibration Estimation in Survey Sampling. *International Statistical Review*, 78, 21–39.

Liu, T.-P. and Rancourt, E. (2001). Constrained categorical imputation for non-response in surveys, *Working Paper HSMD-2001-012E*, Methodology Branch, Statistics Canada, Ottawa.

Rao, C. R. (1970). Estimation of heteroscedastic variances in linear models, *Journal of the American Statistical Association*, 65, 161–172.

Rao, C.R. (1971a). Estimation of variance covariance components - MINQUE theory, *Journal of Multivariate Analysis*, 1, 257–275.

Rao, C. R. (1971b). Minimum variance quadratic unbiased estimation of variance components, *Journal of Multivariate Analysis*, 1, 445–456.

Ren, R. and Chambers, R. L. (2002). Outlier robust imputation of survey data via reverse calibration, *Methodology Working Paper M03/19* , Southampton Statistical Sciences Research Institute, University of Southampton, Southampton.

Rao, C.R. and Kleffe, J. (1988). *Estimation of Variance Components and Applications.* North-Holland Series in Statistics and Probability, Elsevier, Amsterdam.

Rubin, D. B. (1974). Characterizing the estimation of parameters in incomplete-data problems, *Journal of the American Statistical Association*, 69, 467–474.

Rubin, D. B. (1976). Inference and Missing Data, *Biometrika*, 63, 581–590.

Särndal, C.-E. (2007). The calibration approach in survey theory and practice. *Survey Methodology*, 33, 99–119.

Shao, J. and Sitter, R.R. (1996). Bootstrap for imputed survey data. *Journal of the American Statistical Association*, 93, 819–831.

Shao, J., and Wang, H. (2002). Sample correlation coefficients based on survey data under regression imputation, *Journal of the American Statistical Association*, 97, 544–552.

Skinner, C. J., and Rao, J. N. K. (2002). Jackknife variance for multivariate statistics under hot deck imputation from common donors, *Journal of Statistical Planning and Inference*, 102, 149–167.

Chapter 4

Preserving the distribution function in case of imputation for zero inflated data

Imputation methods need to be adapted to the study variable which has to be imputed. For instance, in business surveys, the variables of interest often contain a large number of zeros. In the Capital Expenditure Survey conducted at statistics Canada, approximately 70% of businesses reported a value of zero to Capital Machinery and 50% reported a value of zero to Capital Construction (Haziza et al., 2014). In case of some variable of interest containing a large amount of zeroes, Haziza et al. (2014) proposed imputation methods based on a mixture regression model. They proved that these methods led to doubly robust estimators of the population mean, in the sense that the imputed estimator of the mean is consistent whether the variable of interest or the non-response mechanism is adequately modeled. However, these methods are not necessarily appropriate when estimating more complex parameters such as the population distribution function.

In this work, we consider estimating the population distribution function in case of imputation for zero inflated data. We use the IM approach, without explicit assumptions on the non-response mechanism for the variable of interest. We propose a random imputation method which leads to a consistent estimator of the

population distribution function. As recalled in Haziza et al. (2014), random imputation methods usually suffer from an additional variability due to the imputation variance. Therefore, we also propose a balanced version of our method, which enables to reduce the imputation variance. Roughly speaking, it consists of randomly generating the imputed values while satisfying appropriate balancing constraints, by using an adaptation of the Cube algorithm (Deville & Tillé (2004); Chauvet, Deville & Haziza 2011).

The paper is organized as follows. In Section 4.1, we describe the theoretical set-up and the notation used throughout the paper. In Section 4.2, we briefly recall the two imputation procedures proposed by Haziza et al.(2014), and introduce our two proposed imputation methods. In Section 4.3, we prove that the proposed random imputation procedure yields a consistent estimator of the total and of the population distribution function. The results of a simulation study comparing the four procedures in terms of bias and relative efficiency are presented in Section 4.4.

4.1 Theoretical set-up

We are interested in some finite population U of size N, with some variable of interest y taking the value y_i for unit $i \in U$. We note $\mathbf{y}_U = (y_1, \ldots, y_N)^\top$ for the vector of values for the variable y. We are interested in estimating the total $t_y = \sum_{i \in U} y_i$, and the finite population distribution function

$$F_N(t) = \frac{1}{N} \sum_{i \in U} 1(y_i \leq t) \tag{4.1.1}$$

where $1(\cdot)$ is the indicator function.

A sample s of size n is selected according to a sampling design $p(.)$, with π_i the first-order inclusion probability in the sample for unit i. We suppose that $\pi_i > 0$ for any unit $i \in U$, and we note $d_i = \pi_i^{-1}$ the design weight. We note

$\boldsymbol{\delta}_U = (\delta_1, \ldots, \delta_N)^\top$ for the vector of sample membership indicators. In case of full response, a complete data estimator of t_y is the expansion estimator

$$\hat{t}_{y\pi} = \sum_{i \in s} d_i y_i. \tag{4.1.2}$$

This estimator is design-unbiased for t_y, in the sense that $E_p(\hat{t}_{y\pi}) = t_y$ with E_p the expectation under the sampling design $p(.)$, conditionally on $\mathbf{y}_U$. We also note V_p the variance under the sampling design $p(.)$. Concerning the population distribution function F_N, plugging into (4.1.1) the expansion estimators of the involved totals yields the plug-in estimator

$$\hat{F}_N(t) = \frac{1}{\hat{N}_\pi} \sum_{i \in s} d_i 1(y_i \leq t) \quad \text{with} \quad \hat{N}_\pi = \sum_{i \in s} d_i. \tag{4.1.3}$$

Under some mild assumptions on the variable of interest and the sampling design (see Deville, 1999; Cardot, Chaouch, Goga et Labruère, 2010), $\hat{F}_N(t)$ is approximately unbiased and mean-square consistent for $F_N(t)$.

We now turn to the case when the variable of interest y is subject to missingness. Let r_i be the response indicator, such that $r_i = 1$ if unit i responded to item y, and $r_i = 0$ otherwise. We note $\mathbf{r} = (r_1, \ldots, r_N)^\top$ for the vector of the response indicators. We assume that each unit responds independently of one another. Let E_q and V_q denote the expectation and variance under the non-response mechanism, conditionally on the vector $\mathbf{y}_U$ of population values and on the vector $\boldsymbol{\delta}_U$ of sample membership indicators. An imputation mechanism is used to replace the missing values. That is, an artificial value y_i^* is used to replace the missing y_i. An imputed estimator for t_y based on observed and imputed values is

$$\hat{t}_{yI} = \sum_{i \in s} d_i r_i y_i + \sum_{i \in s} d_i (1 - r_i) y_i^*. \tag{4.1.4}$$

Similarly, an imputed estimator of the distribution function based on observed and imputed values is

$$\hat{F}_I(t) = \frac{1}{\hat{N}_\pi} \left\{ \sum_{i \in s} d_i r_i 1(y_i \leq t) + \sum_{i \in s} d_i (1 - r_i) 1(y_i^* \leq t) \right\}. \tag{4.1.5}$$

In comparison with the estimators obtained in (4.1.2) and (4.1.3) with complete data, there are two additional random mechanisms involved in the estimators given in (4.1.4) and (4.1.5). First, the non-response mechanism leads to observe the values of y for a part of s only. Then, the imputation mechanism is used to replace missing y_i's with artificial values.

The imputation mechanism is motivated by an underlying imputation model, which corresponds to a set of assumptions on the variable y subject to missingness. In the context of a zero-inflated variable of interest, we consider the mixture regression model introduced in Haziza et al. (2014). Namely, we assume that

$$y_i = \eta_i \left\{ \mathbf{z}_i^\top \beta + \sigma \sqrt{v_i} \epsilon_i \right\}, \tag{4.1.6}$$

where the η_i's are independent Bernoulli random variables equal to 1 with probability ϕ_i, and equal to 0 otherwise; the ϵ_i's are independent and identically distributed random variables of mean 0, variance 1 and with a common distribution function F_ϵ; the parameters β and σ are unknown, and v_i is a known constant. The vector $\mathbf{z}_i$ is a vector of auxiliary variables, which is assumed to be known on the whole sample including non-respondents. To sum up, according to the imputation model (4.1.6) the variable y_i follows a regression model with a probability ϕ_i, and is equal to 0 otherwise. Let E_m et V_m denote respectively the expectation and variance under the imputation model.

In practice, the ϕ_i's are unknown and need to be estimated. We assume that they may be parametrically modeled as

$$\phi_i = f(\mathbf{u}_i, \gamma) \tag{4.1.7}$$

where f is a known function, $\mathbf{u}_i$ is a vector of variables recorded for all sampled units, and γ is an unknown parameter. An estimator of ϕ_i is given by

$$\hat{\phi}_i = f(\mathbf{u}_i, \hat{\gamma}_r) \tag{4.1.8}$$

with $\hat{\gamma}_r$ an estimator of the unknown coefficient γ computed on the responding units. We assume that η_i and ϵ_i are independent, conditionally on the vectors $\mathbf{z}_i$ and $\mathbf{u}_i$. We will also assume that there exists some vector $\boldsymbol{\lambda}$ such that

$$v_i^{1/2} = \boldsymbol{\lambda}^\top \mathbf{z}_i. \tag{4.1.9}$$

In this paper, we use the Imputation Model (IM) approach where the inference is made with respect to the imputation model, the sampling design, the response mechanism and the imputation mechanism. This does not require an explicit modeling of the non-response mechanism unlike the Non-response Model approach (Haziza, 2009), but we assume that the data are missing at random, which means that model (4.1.6) holds for both the respondents and the non-respondents. We note E_I and V_I the expectation and variance under the imputation mechanism, conditionally on the vectors $\mathbf{y}_U$, $\boldsymbol{\delta}_U$ and $\mathbf{r}_U$.

4.2 Imputation methods

In this Section, we first briefly recall in Sections 4.2.1 and 4.2.2 the random imputation methods proposed by Haziza et al. (2014) for zero-inflated data. We then introduce the new methods that we propose in Sections 4.2.3 and 4.2.4.

4.2.1 Haziza-Nambeu-Chauvet random imputation

A first proposal of Haziza et al. (2014) is to use the imputation mechanism

$$y_i^* = \eta_i^* \left\{ \mathbf{z}_i^\top \hat{\mathbf{B}}_r^* \right\}, \tag{4.2.1}$$

where the unknown regression parameter β is estimated by

$$\hat{B}_r^* = \left(\sum_{i \in s} \omega_i r_i \hat{\phi}_i v_i^{-1} \mathbf{z}_i \mathbf{z}_i^\top \right)^{-1} \sum_{i \in s} \omega_i r_i v_i^{-1} \mathbf{z}_i y_i, \tag{4.2.2}$$

where ω_i denotes a so called imputation weight, and $\hat{\phi}_i$ is given in (4.1.8). The η_i^*'s are independently generated, and η_i^* is equal to 1 with the probability $\hat{\phi}_i$, and

is equal to 0 otherwise.

There are several possible choices for the imputation weights ω_i. Using a modeling of the response mechanism for the variable y_i, Haziza et al. (2014) propose to choose the imputation weights so that $\hat{t}_{yI}$ is a doubly robust estimator for t_y. This means that the imputed estimator is approximately unbiased for t_y whether the imputation model or the non-response model is adequately specified. Haziza et al. (2014) also prove that the resulting imputed estimator is consistent for t_y under either approach.

The random imputation mechanism in (4.2.1) has two drawbacks. Firstly, it leads to an additional imputation variance due to the η_i^*'s. To overcome this problem, Haziza et al. (2014) proposed a balanced version of their imputation mechanism that is presented in Section 4.2.2. Secondly, the imputation mechanism in (4.2.1) does not lead to an approximately unbiased estimator of the distribution function, as will be illustrated in the simulation study conducted in Section 4.4.

4.2.2 Haziza-Nambeu-Chauvet balanced random imputation

The balanced random imputation procedure of Haziza et al. (2014) consists of replacing a missing value with

$$y_i^* \quad = \quad \tilde{\eta}_i^* \left\{ \mathbf{z}_i^\top \hat{\mathbf{B}}_r^* \right\}, \tag{4.2.3}$$

where the $\tilde{\eta}_i^*$'s are not independently generated, but so that the imputation variance of $\hat{t}_{yI}$ is approximately equal to zero. Indeed, the imputation variance of $\hat{t}_{yI}$ is eliminated if the following constraint is satisfied:

$$\sum_{i \in s} d_i (1 - r_i)(y_i^* - \hat{\phi}_i \mathbf{z}_i^\top \hat{\mathbf{B}}_r^*) \quad = \quad 0, \tag{4.2.4}$$

which is equivalent to generate the $\tilde{\eta}_i^*$'s so that

$$\sum_{i \in s} d_i(1-r_i)(\tilde{\eta}_i^* - \hat{\phi}_i)(\mathbf{z}_i^\top \hat{\mathbf{B}}_{\mathbf{r}}^*) \quad = \quad 0. \tag{4.2.5}$$

Haziza et al. (2014) propose a procedure adapted from the Cube method (Deville and Tillé, 2004; Chauvet and Tillé, 2006) which enables to generate the $\tilde{\eta}_i^*$'s so that (4.2.5) is satisfied, at least approximately. As a result, the imputation variance is eliminated or at least significantly reduced.

The corresponding imputation procedure is called balanced random ϕ-regression (BRR_ϕ) imputation by Haziza et al. (2014). They prove that under the BRR_ϕ imputation, an appropriate choice for the imputation weights ω_i leads to a doubly robust estimator for t_y. Also, their empirical results indicate that it performs well in reducing the imputation variance. A drawback of the BRR_ϕ imputation mechanism is that it does not preserve the distribution function of the imputed variable, because it does not take into account the error terms ϵ_i in the imputation model (4.1.6). This is empirically illustrated in section 4.4. In order to overcome this problem, two new imputation procedures are proposed in Sections 4.2.3 and 4.2.4 below.

4.2.3 Proposed random imputation procedure

The random imputation procedure that we propose consists in mimicking as closely as possible the imputation model (4.1.6), by replacing some missing y_i with the imputed value

$$y_i^* \quad = \quad \eta_i^* \left\{ \mathbf{z}_i^\top \hat{\mathbf{B}}_r^* + \hat{\sigma}\sqrt{v_i}\epsilon_i^* \right\}, \tag{4.2.6}$$

where $\hat{\mathbf{B}}_r^*$ is defined in equation (4.2.2), and η_i^* is a Bernoulli random variable as defined in (4.2.1). In the imputed value given in (4.2.6), $\hat{\sigma}$ is an estimator of σ and the ϵ_i^*'s are selected independently and with replacement in the set of observed

estimated residuals

$$G_r = \{e_j \ ; \ r_j = 1 \text{ and } \eta_j = 1\} \quad \text{where} \quad e_j = \frac{y_j - \mathbf{z}_j^\top \hat{\mathbf{B}}_r^*}{\hat{\sigma}\sqrt{v_j}}, \tag{4.2.7}$$

with $Pr(\epsilon_i^* = e_j) = \tilde{\omega}_j$ for any $j \in s$ such that $r_j = 1$ and $\eta_j = 1$, where

$$\tilde{\omega}_j \ = \ \frac{\omega_j}{\sum_{k\in s} \omega_j r_k \eta_k}. \tag{4.2.8}$$

Under this imputation procedure, we have

$$\begin{aligned} E_I(y_i^*) &= \hat{\phi}_i(\mathbf{z}_i^\top \hat{\mathbf{B}}_r^*), \\ V_I(y_i^*) &= \hat{\phi}_i(1-\hat{\phi}_i)(\mathbf{z}_i^\top \hat{\mathbf{B}}_r^*)^2 + (\hat{\phi}_i v_i) \sum_{j\in s} \tilde{\omega}_j r_j \eta_j \left(\frac{y_j - \mathbf{z}_j^\top \hat{\mathbf{B}}_r^*}{\sqrt{v_j}}\right)^2. \end{aligned} \tag{4.2.9}$$

It easily follows from the first line in equation (4.2.9) that under this imputation procedure, the imputed estimator $\hat{t}_{yI}$ is approximately unbiased for t_y. Further theoretical properties of this imputation procedure are examined in Section 4.3. We prove in Theorem 4.1 that this random imputation method leads to an imputed estimator $\hat{t}_{yI}$ which is mean-square consistent for the true total. Also, we prove in Theorem 4.2 that this method leads to an imputed estimator $\hat{F}_I(t)$ which is L_1-consistent for the population distribution function. However, this imputation procedure leads to an additional variability for $\hat{t}_{yI}$ due to the imputation variance. Therefore, a balanced version of this imputation procedure is proposed in Section 4.2.4 below.

4.2.4 The proposed random balanced imputation procedure

The balanced random imputation procedure consists in replacing a missing value with

$$y_i^* \ = \ \tilde{\eta}_i^* \left\{ \mathbf{z}_i^\top \hat{\mathbf{B}}_r^* + \hat{\sigma}\sqrt{v_i}\tilde{\epsilon}_i^* \right\}, \tag{4.2.10}$$

where $\hat{\mathbf{B}}_r^*$ is as defined in (4.2.2), but where the $\tilde{\eta}_i^*$'s and the $\tilde{\epsilon}_i^*$'s are not independently generated, but so as to eliminate the imputation variance of $\hat{t}_{yI}$. The

imputation variance is zero if equation (4.2.4) holds. A sufficient condition consists in generating the residuals $\tilde{\eta}_i^*$ and $\tilde{\epsilon}_i^*$ so that

$$\sum_{i \in s} d_i (1 - r_i)(\tilde{\eta}_i^* - \hat{\phi}_i)(\mathbf{z}_i^\top \hat{\mathbf{B}}_r^*) = 0, \quad (4.2.11)$$

$$\sum_{i \in s} d_i (1 - r_i) \tilde{\eta}_i^* \sqrt{v_i} \tilde{\epsilon}_i^* = 0. \quad (4.2.12)$$

This is done in a two-step procedure: first, the $\tilde{\eta}_i^*$'s are generated by means of Algorithm 1 in Haziza et al. (2014), so that (4.2.11) is approximately respected; then, the $\tilde{\epsilon}_i^*$'s are generated by using algorithm 1 described in Chauvet et al. (2011), so that (4.2.12) is approximately respected. Like with the procedure described in Section 4.2.3, it leads to an approximately unbiased estimation of the total and of the distribution function, as empirically illustrated in Section 4.4. Also, this imputation procedure is fully efficient for the estimation of the total, and the imputation variance is reduced for the estimation of the population distribution function.

4.3 Properties of the proposed imputation methods

In this section, we prove that the proposed random imputation procedure leads to a consistent estimator for the total and the distribution function. In order to study the asymptotic properties of the sampling designs and the estimators in this article, we use the asymptotic framework proposed by Isaki et Fuller (1982). We suppose that population U belongs to a nested sequence $\{U_\tau\}$ of finite populations with increasing sizes N_τ, and that the vector of values for the variable of interest $\mathbf{y}_{U\tau} = (y_{1\tau}, \ldots, y_{N\tau})^\top$ belongs to a nested sequence $\{\mathbf{y}_{U\tau}\}$ with increasing sizes N_τ. For simplicity, the index τ is omitted in what follows and all limits are computed when $\tau \to \infty$.

Theorem 4.1. *Let us suppose that the imputation model (4.1.6) holds and that the following assumptions are satisfied:*

H1: There exists some constants C_1 and C_2 such that $d_i \leq C_1 N n^{-1}$ and $\tilde{\omega}_i \leq C_2 n^{-1}$ for any unit $i \in U$. Also, there exists some constant C_3 such that $\max_{i \neq j \in U} |\pi_{ij} - \pi_i \pi_j| \leq C_3 n^{-1}$, with π_{ij} the probability that units i and j are selected together in the sample.

H2: There exists a constant K_1 such that for all $i \in U$, $0 < K_1 < p_i$.

H3: There exists some constants K_2, K_3, K_4 such that $\|z_i\| \leq K_2$, $\|v_i\| \leq K_3$ and $\|v_i^{-1}\| \leq K_4$ for all $i \in U$.

H4: We have $E\left(\|\hat{\boldsymbol{\gamma}}_r - \boldsymbol{\gamma}\|^2\right) = O(n^{-1})$, where $\boldsymbol{\gamma}$ is given in (4.1.7).

H5: There exists a constant K_5 such that for any vector $\tilde{\gamma}$

$$|f(u_i, \tilde{\boldsymbol{\gamma}}) - f(u_i, \boldsymbol{\gamma})| \leq K_5 \|\tilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\| \text{ for all } i \in U.$$

H6: We have $E\left(\|\hat{\mathbf{B}}_r^* - \boldsymbol{\beta}\|^2\right) = O(n^{-1})$.

Then under the random imputation mechanism proposed in Section 4.2.3, we have

$$E\left\{N^{-1}(\hat{t}_{yI} - t_y)\right\}^2 \;=\; O(n^{-1}), \tag{4.3.1}$$

so that the imputed estimator $\hat{t}_{yI}$ is mean-square consistent for the true total.

Theorem 4.2. *Let us suppose that the imputation model given by (4.1.6) holds and that the assumptions (H1)-(H6) are satisfied. Let us also suppose that the distribution function F_ϵ is absolutely continuous. Then under the random imputation mechanism proposed in Section 4.2.3, we have for any $t \in \mathbb{R}$*

$$E|\hat{F}_I(t) - F_N(t)| \;=\; o(1), \tag{4.3.2}$$

so that the imputed estimator of the distribution function $\hat{F}_I$ is L_1-consistent for the true population distribution function at any point t.

4.4 Simulation study

In order to evaluate the performance of the imputation methods that we propose, we implement a simulation study inspired by Haziza et al. (2014). We generate twelve finite populations of size $N = 10,000$ with a variable of interest y and an auxiliary variable z. The values of z are generated according to a Gamma distribution with shift parameter 2 and scale parameter 5. The values of y are generated according to the following mixture model:

$$y_i = \eta_i(a_0 + a_1 z_i + \epsilon_i), \tag{4.4.1}$$

where the ϵ_i's are generated according to a centered Normal distribution with variance σ^2. We use $a_0 = 30$ and $a_1 = 1.5$. Also, we choose three different values of σ^2 so that the coefficient of determination R^2 equals 0.4, 0.5 or 0.6 for the units i such that $\eta_i = 1$.

The η_i's are generated according to a Bernoulli distribution with parameter ϕ_i, and

$$\log\left(\frac{\phi_i}{1-\phi_i}\right) = b_0 + b_1 z_i, \tag{4.4.2}$$

and with four possible values for the parameters b_0 and b_1, chosen so that the proportion of non-null values is approximately equal to 0.60, 0.70, 0.80 or 0.90. The four different proportion of non-null values, crossed with the three different levels for the R^2, lead to the twelve finite populations.

We are interested in estimating the total t_y, and the distribution function $F_N(t)$ with $t = t_\alpha$, the α-th quintile. In this simulation study, we consider the values $\alpha = 0.50, 0.75$ and 0.95. In each population, we select $R = 1,000$ without-replacement simple random samples of size $n = 500$. In each sample, we generate a response indicator r_i for unit i according to a Bernoulli distribution with parameter p_i such

that

$$\log\left(\frac{p_i}{1-p_i}\right) \quad = \quad c_0 + c_1 z_i. \tag{4.4.3}$$

We use four possible values for the parameters c_0 and c_1, chosen so that the proportion of respondents is approximately equal to 0.50, 0.60, 0.70 or 0.80.

In this simulation study, we compare four imputation methods to handle non-response:

(i) RR_ϕ: random imputation proposed by Haziza et al. (2014), and presented in Section 4.2.1;

(ii) BRR_ϕ: balanced random imputation proposed by Haziza et al. (2014), and presented in Section 4.2.2;

(iii) MRR_ϕ : proposed random imputation method, presented in Section 4.2.3;

(iv) $BMRR_\phi$: proposed balanced random imputation method, presented in Section 4.2.4.

For each of the four methods, we use imputation weights $\omega_i = 1$, and the ϕ_i's and p_i's are estimated by means of logistic regression modeling. In each sample, missing values are replaced by imputed values according to imputation methods (i) to (iv), and the imputed estimators $\hat{t}_{yI}$ and $\hat{F}_I(t_\alpha)$ are computed.

As a measure of bias of an estimator $\hat{\theta}_I$ of a finite population parameter θ, we compute the Monte Carlo percent relative bias

$$RB_{MC}(\hat{\theta}_I) \quad = \quad \frac{100}{R}\sum_{k=1}^{R}\frac{(\hat{\theta}_{I(k)} - \theta)}{\theta}, \tag{4.4.4}$$

where $\hat{\theta}_{I(k)}$ denotes the imputed estimator computed in the k-th sample. As a measure of relative efficiency for each imputation method, using $BMRR_\phi$ as a

		RR_ϕ		BRR_ϕ		MRR_ϕ		$BMRR_\phi$	
R^2	$\overline{\phi}$	RB %	RE	RB %	RE	RB %	RE	RB %	RE
0.4	0.6	0.04	1.11	0.04	1.00	-0.01	1.17	0.04	1.00
0.4	0.7	-0.05	1.16	-0.05	1.00	-0.07	1.21	-0.04	1.00
0.4	0.8	-0.17	1.17	-0.13	1.00	-0.14	1.25	-0.13	1.00
0.4	0.9	-0.09	1.13	-0.14	1.00	-0.09	1.25	-0.15	1.00
0.5	0.6	0.17	1.20	0.04	1.00	0.14	1.23	0.04	1.00
0.5	0.7	0.03	1.25	-0.04	1.00	0.03	1.27	-0.04	1.00
0.5	0.8	-0.11	1.18	-0.12	1.00	-0.13	1.24	-0.13	1.00
0.5	0.9	-0.17	1.08	-0.13	1.00	-0.18	1.14	-0.13	1.00
0.6	0.6	0.04	1.21	0.03	1.00	0.02	1.24	0.03	1.00
0.6	0.7	-0.19	1.23	-0.03	1.00	-0.19	1.28	-0.03	1.00
0.6	0.8	-0.16	1.21	-0.12	1.00	-0.18	1.25	-0.12	1.00
0.6	0.9	-0.10	1.14	-0.13	1.00	-0.11	1.18	-0.12	1.00

Table 4.4.1: Relative bias (RB %) and Relative efficiency (RE) of four imputed estimators of the total with an average response probability of 50%

benchmark, we computed

$$RE_{MC}(\hat{\theta}_I) = \frac{MSE_{MC}(\hat{\theta}_I)}{MSE_{MC}(\hat{\theta}_{BMRR_\phi})},$$

with

$$MSE_{MC}(\hat{\theta}_I) = \frac{1}{R}\sum_{k=1}^{R}(\hat{\theta}_{I(k)} - \theta)^2 \tag{4.4.5}$$

the Mean Square Error of $\hat{\theta}_I$ approximated by means of the R simulations.

We first consider the estimation of the total t_y, for which the simulation results are given in Tables 4.4.1 to 4.4.4. The four imputation methods lead to approximately unbiased estimators of the total, as expected. Turning to the relative efficiency (RE), we note that in all studied cases the balanced version of an imputation method outperforms its unbalanced version. Also, the two balanced imputation procedures BRR_ϕ and $BMRR_\phi$ exhibit the same efficiency.

		RR_ϕ		BRR_ϕ		MRR_ϕ		$BMRR_\phi$	
R^2	$\overline{\phi}$	RB %	RE	RB %	RE	RB %	RE	RB %	RE
0.4	0.6	0.04	1.18	-0.05	1.00	0.07	1.19	-0.06	1.00
0.4	0.7	-0.15	1.22	-0.10	0.99	-0.16	1.29	-0.11	1.00
0.4	0.8	-0.08	1.17	-0.12	1.00	-0.09	1.24	-0.12	1.00
0.4	0.9	-0.17	1.10	-0.16	1.00	-0.20	1.21	-0.16	1.00
0.5	0.6	0.03	1.18	-0.05	1.00	0.04	1.21	-0.05	1.00
0.5	0.7	-0.12	1.17	-0.09	0.99	-0.13	1.20	-0.10	1.00
0.5	0.8	-0.14	1.19	-0.11	1.00	-0.12	1.25	-0.11	1.00
0.5	0.9	-0.16	1.11	-0.15	1.00	-0.13	1.17	-0.15	1.00
0.6	0.6	0.09	1.20	-0.05	1.00	0.08	1.23	-0.06	1.00
0.6	0.7	-0.04	1.24	-0.09	0.99	-0.01	1.24	-0.10	1.00
0.6	0.8	-0.14	1.18	-0.11	1.00	-0.12	1.22	-0.10	1.00
0.6	0.9	-0.14	1.12	-0.14	1.00	-0.14	1.18	-0.14	1.00

Table 4.4.2: Relative bias (RB %) and Relative efficiency (RE) of four imputed estimators of the total with an average response probability of 60%

		RR_ϕ		BRR_ϕ		MRR_ϕ		$BMRR_\phi$	
R^2	$\overline{\phi}$	RB %	RE	RB %	RE	RB %	RE	RB %	RE
0.4	0.6	0.04	1.13	0.00	1.00	0.06	1.16	0.00	1.00
0.4	0.7	-0.10	1.18	-0.05	1.00	-0.09	1.20	-0.05	1.00
0.4	0.8	-0.13	1.15	-0.10	1.00	-0.13	1.22	-0.10	1.00
0.4	0.9	-0.09	1.09	-0.12	1.00	-0.10	1.14	-0.12	1.00
0.5	0.6	-0.09	1.17	0.00	1.00	-0.09	1.19	0.01	1.00
0.5	0.7	0.02	1.17	-0.04	1.00	0.04	1.18	-0.05	1.00
0.5	0.8	-0.10	1.16	-0.09	1.00	-0.13	1.21	-0.09	1.00
0.5	0.9	-0.14	1.11	-0.11	0.99	-0.17	1.16	-0.12	1.00
0.6	0.6	-0.03	1.16	0.01	1.00	-0.07	1.17	0.01	1.00
0.6	0.7	-0.06	1.14	-0.03	1.00	-0.04	1.14	-0.03	1.00
0.6	0.8	-0.10	1.22	-0.08	1.00	-0.08	1.24	-0.08	1.00
0.6	0.9	-0.14	1.10	-0.10	1.00	-0.14	1.13	-0.10	1.00

Table 4.4.3: Relative bias (RB %) and Relative efficiency (RE) of four imputed estimators of the total with an average response probability of 70%

		RR_ϕ		BRR_ϕ		MRR_ϕ		$BMRR_\phi$	
R^2	$\overline{\phi}$	RB %	RE	RB %	RE	RB %	RE	RB %	RE
0.4	0.6	0.08	1.15	0.07	1.00	0.05	1.19	0.07	1.00
0.4	0.7	-0.03	1.14	0.00	1.00	0.00	1.19	0.00	1.00
0.4	0.8	-0.11	1.11	-0.08	1.00	-0.11	1.16	-0.08	1.00
0.4	0.9	-0.08	1.09	-0.09	1.00	-0.11	1.16	-0.09	1.00
0.5	0.6	0.06	1.08	0.06	1.00	0.07	1.09	0.07	1.00
0.5	0.7	0.06	1.14	0.01	1.00	0.10	1.16	0.01	1.00
0.5	0.8	-0.06	1.11	-0.08	1.00	-0.10	1.15	-0.08	1.00
0.5	0.9	-0.10	1.08	-0.09	1.00	-0.12	1.09	-0.08	1.00
0.6	0.6	-0.01	1.13	0.07	1.00	-0.04	1.15	0.06	1.00
0.6	0.7	-0.04	1.16	0.01	1.00	-0.01	1.18	0.01	1.00
0.6	0.8	-0.11	1.10	-0.07	1.00	-0.08	1.12	-0.08	1.00
0.6	0.9	-0.07	1.08	-0.08	1.00	-0.05	1.12	-0.08	1.00

Table 4.4.4: Relative bias (RB %) and Relative efficiency (RE) of four imputed estimators of the total with an average response probability of 80%

We now consider the estimation of the population distribution function, for which the simulation results are presented in Tables 4.4.5 to 4.4.12. In all the cases considered, the two proposed imputation methods MRR_ϕ and $BMRR_\phi$ lead to approximately unbiased estimators of the distribution function, with absolute relative biases no greater than 2 % . On the contrary, the RR_ϕ and the BRR_ϕ imputation methods lead to biased estimators. The absolute relative bias can be as large as 16 % . We note that the bias can be particularly large when the response probability is lower, which corresponds to imputing more missing values using an imputation method which does not mimic the imputation model adequately. Turning to the relative efficiency, we note that MRR_ϕ and $BMRR_\phi$ always outperform RR_ϕ and BRR_ϕ, which is partly due to the bias under these latter imputation methods. The gap may be very large in places, with a value of RE as large as 8.65 for BRR_ϕ in comparison with $BMRR_\phi$ for $\alpha = 75\%$, a response probability of 50% and $R^2 = 0.4$,see Table 4.4.5. Comparing the two proposed imputation methods, we note that $BMRR_\phi$ is equivalent or better than MRR_ϕ in terms of efficiency, with values of RE ranging from 0.98 to 1.27.

The gain in accuracy is particularly appreciable with lower response probabilities, that is, when more imputed values are generated according to the imputation mechanism.

4.5 Conclusion

In this paper, we considered imputation for zero-inflated data. We proposed two imputation methods which enable to respect the nature of the data, and in particular which preserve the finite population distribution function. In particular, we proposed a balanced imputation method which enables to preserve the distribution of the imputed variable while being fully efficient for the estimation of a total.

Our imputation methods rely upon the mixture regression imputation model proposed by Haziza et al. (2014). As mentioned by these authors, the proposed methods could be extended to more general mixture regression models, for example to handle count data.

In practice, we may not be interested in the distribution function in itself, but rather in complex parameters such as quantiles. Establishing the theoretical properties of estimators of such parameters under the proposed imputation procedures is a challenging task. This is a topic for further research.

		F50								F75							
		RR_ϕ		BRR_ϕ		MRR_ϕ		$BMRR_\phi$		RR_ϕ		BRR_ϕ		MRR_ϕ		$BMRR_\phi$	
R^2	$\overline{\phi}$	RB %	RE	RB %	RE	RB %	RE	RB %	RE	RB %	RE	RB %	RE	RB %	RE	RB %	RE
0.4	0.6	-10.26	3.93	-10.24	3.80	-1.07	1.12	-0.95	1.00	6.56	4.65	6.54	4.63	0.75	1.10	0.67	1.00
0.4	0.7	-16.22	8.88	-16.20	8.65	-1.01	1.18	-0.97	1.00	7.24	4.52	7.23	4.53	0.65	1.07	0.68	1.00
0.4	0.8	-7.38	4.42	-7.51	4.42	-0.36	1.18	-0.28	1.00	8.10	5.10	8.06	5.11	0.78	1.05	0.79	1.00
0.4	0.9	0.97	2.13	0.99	2.12	0.11	1.12	0.12	1.00	7.66	5.30	7.67	5.32	0.43	1.09	0.48	1.00
0.5	0.6	-10.30	4.25	-10.14	3.92	-1.33	1.24	-1.22	1.00	5.64	3.61	5.70	3.58	0.72	1.06	0.73	1.00
0.5	0.7	-13.53	6.60	-13.40	6.30	-1.24	1.18	-1.09	1.00	6.12	3.51	6.15	3.45	0.78	1.09	0.82	1.00
0.5	0.8	-4.57	2.96	-4.60	2.90	-0.21	1.11	-0.42	1.00	6.95	3.79	6.90	3.77	1.04	1.07	1.11	1.00
0.5	0.9	1.71	1.96	1.64	1.96	0.26	1.10	0.15	1.00	6.29	3.77	6.24	3.73	0.72	1.05	0.65	1.00
0.6	0.6	-9.93	3.84	-9.93	3.60	-1.56	1.22	-1.50	1.00	5.11	2.89	5.09	2.86	0.97	1.05	0.90	1.00
0.6	0.7	-10.79	4.83	-10.99	4.79	-1.28	1.24	-1.42	1.00	5.34	2.66	5.27	2.63	0.99	1.09	0.96	1.00
0.6	0.8	-2.74	2.38	-2.85	2.26	-0.37	1.14	-0.48	1.00	5.77	2.86	5.74	2.83	1.20	1.07	1.09	1.00
0.6	0.9	1.59	1.73	1.63	1.69	-0.16	1.11	-0.03	1.00	5.07	2.62	5.12	2.63	0.77	1.03	0.84	1.00

Table 4.4.5: Relative bias (RB %) and Relative efficiency (RE) of four imputed estimators of the distribution function evaluated at the 50% and the 75% quartiles with an average response probability of 50%

		F50								F75							
		RR_ϕ		BRR_ϕ		MRR_ϕ		$BMRR_\phi$		RR_ϕ		BRR_ϕ		MRR_ϕ		$BMRR_\phi$	
R^2	$\overline{\phi}$	RB %	RE	RB %	RE	RB %	RE	RB %	RE	RB %	RE	RB %	RE	RB %	RE	RB %	RE
0.4	0.6	-8.24	3.35	-8.11	3.07	-0.99	1.14	-0.91	1.00	5.28	4.27	5.24	4.27	0.44	1.08	0.50	1.00
0.4	0.7	-13.03	7.15	-13.13	6.98	-0.95	1.27	-0.89	1.00	5.80	4.17	5.79	4.11	0.52	1.13	0.53	1.00
0.4	0.8	-6.23	3.44	-6.20	3.31	-0.31	1.15	-0.35	1.00	6.47	4.65	6.46	4.63	0.66	1.12	0.67	1.00
0.4	0.9	0.62	1.85	0.58	1.78	0.21	1.14	0.15	1.00	6.09	4.54	6.07	4.52	0.47	1.10	0.41	1.00
0.5	0.6	-8.17	3.41	-8.07	3.16	-1.09	1.16	-1.16	1.00	4.56	3.32	4.58	3.29	0.53	1.05	0.57	1.00
0.5	0.7	-10.93	5.62	-10.93	5.45	-0.96	1.19	-0.91	1.00	4.96	3.17	4.96	3.17	0.67	1.07	0.57	1.00
0.5	0.8	-3.85	2.39	-3.98	2.35	-0.38	1.17	-0.35	1.00	5.52	3.47	5.52	3.48	0.82	1.09	0.81	1.00
0.5	0.9	1.17	1.67	1.12	1.65	0.01	1.14	0.06	1.00	4.99	3.28	4.99	3.27	0.56	1.05	0.58	1.00
0.6	0.6	-8.07	3.32	-7.90	3.07	-1.50	1.24	-1.34	1.00	4.04	2.74	4.03	2.70	0.65	1.07	0.76	1.00
0.6	0.7	-9.11	4.36	-9.10	4.15	-1.32	1.21	-1.21	1.00	4.24	2.45	4.25	2.45	0.71	1.01	0.78	1.00
0.6	0.8	-2.44	1.86	-2.50	1.81	-0.44	1.13	-0.51	1.00	4.53	2.64	4.54	2.61	0.79	1.07	0.89	1.00
0.6	0.9	1.15	1.53	1.16	1.51	-0.01	1.12	-0.10	1.00	4.05	2.41	4.04	2.38	0.74	1.06	0.69	1.00

Table 4.4.6: Relative bias (RB %) and Relative efficiency (RE) of four imputed estimators of the distribution function evaluated at the 50% and the 75% quartiles with an average response probability of 60%

		F50								F75							
		RR_ϕ		BRR_ϕ		MRR_ϕ		$BMRR_\phi$		RR_ϕ		BRR_ϕ		MRR_ϕ		$BMRR_\phi$	
R^2	$\overline{\phi}$	RB %	RE	RB %	RE	RB %	RE	RB %	RE	RB %	RE	RB %	RE	RB %	RE	RB %	RE
0.4	0.6	-6.12	2.42	-6.07	2.25	-0.82	1.12	-0.79	1.00	3.91	3.36	3.89	3.34	0.35	1.08	0.38	1.00
0.4	0.7	-9.85	4.93	-9.92	4.81	-0.69	1.19	-0.76	1.00	4.35	3.36	4.36	3.34	0.36	1.12	0.43	1.00
0.4	0.8	-4.73	2.73	-4.76	2.63	-0.13	1.18	-0.14	1.00	4.89	3.60	4.86	3.61	0.51	1.10	0.52	1.00
0.4	0.9	0.22	1.55	0.27	1.51	0.06	1.04	0.03	1.00	4.52	3.34	4.53	3.34	0.30	1.05	0.29	1.00
0.5	0.6	-5.86	2.40	-6.01	2.29	-0.76	1.16	-0.90	1.00	3.38	2.81	3.38	2.80	0.40	1.09	0.40	1.00
0.5	0.7	-8.34	3.91	-8.27	3.71	-0.90	1.15	-0.79	1.00	3.68	2.72	3.69	2.70	0.42	1.07	0.45	1.00
0.5	0.8	-3.08	1.99	-3.09	1.92	-0.21	1.16	-0.27	1.00	4.17	2.85	4.17	2.85	0.77	1.10	0.68	1.00
0.5	0.9	0.78	1.45	0.77	1.41	0.28	1.10	0.20	1.00	3.70	2.56	3.70	2.56	0.45	1.04	0.40	1.00
0.6	0.6	-5.89	2.37	-5.94	2.23	-1.00	1.15	-1.20	1.00	3.01	2.41	3.02	2.36	0.50	1.06	0.52	1.00
0.6	0.7	-6.87	3.00	-6.89	2.89	-1.00	1.03	-1.05	1.00	3.21	2.17	3.15	2.17	0.56	1.03	0.61	1.00
0.6	0.8	-1.94	1.65	-2.02	1.57	-0.42	1.10	-0.31	1.00	3.42	2.24	3.42	2.22	0.65	1.06	0.62	1.00
0.6	0.9	0.80	1.32	0.75	1.29	0.06	1.06	0.04	1.00	3.00	1.94	3.01	1.94	0.55	0.99	0.48	1.00

Table 4.4.7: Relative bias (RB %) and Relative efficiency (RE) of four imputed estimators of the distribution function evaluated at the 50% and the 75% quartiles with an average response probability of 70%

		F50								F75							
		RR_ϕ		BRR_ϕ		MRR_ϕ		$BMRR_\phi$		RR_ϕ		BRR_ϕ		MRR_ϕ		$BMRR_\phi$	
R^2	$\overline{\phi}$	RB %	RE	RB %	RE	RB %	RE	RB %	RE	RB %	RE	RB %	RE	RB %	RE	RB %	RE
0.4	0.6	-4.07	1.81	-4.05	1.60	-0.56	1.16	-0.58	1.00	2.50	2.30	2.52	2.29	0.22	1.07	0.18	1.00
0.4	0.7	-6.59	3.07	-6.61	2.93	-0.54	1.13	-0.50	1.00	2.84	2.30	2.80	2.28	0.20	1.08	0.17	1.00
0.4	0.8	-3.14	1.97	-3.15	1.91	-0.05	1.11	-0.11	1.00	3.22	2.41	3.22	2.39	0.37	1.07	0.32	1.00
0.4	0.9	0.13	1.32	0.09	1.29	0.10	1.08	0.03	1.00	2.98	2.23	2.98	2.22	0.16	1.05	0.15	1.00
0.5	0.6	-4.00	1.75	-4.00	1.64	-0.71	1.10	-0.70	1.00	2.12	2.04	2.15	2.04	0.23	1.07	0.20	1.00
0.5	0.7	-5.60	2.52	-5.56	2.37	-0.65	1.08	-0.52	1.00	2.39	1.98	2.41	1.97	0.21	1.04	0.28	1.00
0.5	0.8	-2.15	1.55	-2.08	1.50	-0.07	1.08	-0.07	1.00	2.77	2.08	2.79	2.07	0.50	1.08	0.47	1.00
0.5	0.9	0.47	1.22	0.43	1.21	0.18	1.05	0.10	1.00	2.43	1.91	2.43	1.89	0.31	1.05	0.28	1.00
0.6	0.6	-3.90	1.73	-3.97	1.60	-0.74	1.13	-0.84	1.00	1.92	1.87	1.88	1.84	0.32	1.07	0.25	1.00
0.6	0.7	-4.58	2.06	-4.63	1.97	-0.65	1.09	-0.73	1.00	2.10	1.75	2.11	1.74	0.33	1.07	0.39	1.00
0.6	0.8	-1.29	1.35	-1.32	1.33	-0.23	1.09	-0.17	1.00	2.30	1.75	2.29	1.73	0.41	1.02	0.47	1.00
0.6	0.9	0.48	1.19	0.46	1.18	-0.04	1.06	0.05	1.00	1.99	1.64	1.98	1.62	0.30	1.05	0.33	1.00

Table 4.4.8: Relative bias (RB %) and Relative efficiency (RE) of four imputed estimators of the distribution function evaluated at the 50% and the 75% quartiles with an average response probability of 80%

		F90								F95							
		RR_ϕ		BRR_ϕ		MRR_ϕ		$BMRR_\phi$		RR_ϕ		BRR_ϕ		MRR_ϕ		$BMRR_\phi$	
R^2	$\overline{\phi}$	RB %	RE	RB %	RE	RB %	RE	RB %	RE	RB %	RE	RB %	RE	RB %	RE	RB %	RE
0.4	0.6	3.87	2.38	3.85	2.35	0.80	1.03	0.76	1.00	2.05	1.50	2.04	1.48	0.61	1.02	0.61	1.00
0.4	0.7	3.87	2.43	3.85	2.42	0.79	1.03	0.72	1.00	2.04	1.55	2.02	1.55	0.53	0.99	0.48	1.00
0.4	0.8	4.06	2.76	4.06	2.77	0.82	1.02	0.80	1.00	2.11	1.76	2.11	1.77	0.57	0.99	0.53	1.00
0.4	0.9	3.92	3.26	3.92	3.27	0.60	1.07	0.65	1.00	2.04	2.03	2.05	2.03	0.41	1.00	0.43	1.00
0.5	0.6	3.44	1.83	3.46	1.82	1.06	1.00	1.02	1.00	1.78	1.24	1.78	1.23	0.72	1.00	0.68	1.00
0.5	0.7	3.39	1.86	3.41	1.85	0.92	1.05	0.97	1.00	1.78	1.31	1.80	1.29	0.67	1.02	0.68	1.00
0.5	0.8	3.55	2.06	3.55	2.04	1.00	1.00	1.02	1.00	1.88	1.51	1.88	1.49	0.68	1.06	0.71	1.00
0.5	0.9	3.32	2.33	3.30	2.32	0.78	1.01	0.72	1.00	1.76	1.65	1.75	1.64	0.52	1.00	0.51	1.00
0.6	0.6	3.13	1.49	3.15	1.49	1.32	0.98	1.35	1.00	1.63	1.11	1.62	1.11	0.86	0.99	0.85	1.00
0.6	0.7	2.97	1.50	2.96	1.50	1.14	1.00	1.06	1.00	1.54	1.13	1.52	1.14	0.67	1.01	0.64	1.00
0.6	0.8	3.11	1.68	3.10	1.67	1.22	1.01	1.15	1.00	1.64	1.30	1.65	1.30	0.77	1.03	0.74	1.00
0.6	0.9	2.72	1.88	2.73	1.87	0.81	1.05	0.76	1.00	1.50	1.41	1.50	1.41	0.59	1.05	0.55	1.00

Table 4.4.9: Relative bias (RB %) and Relative efficiency (RE) of four imputed estimators of the distribution function evaluated at the 90% and the 95% quartiles with an average response probability of 50%

		F90								F95							
		RR_ϕ		BRR_ϕ		MRR_ϕ		$BMRR_\phi$		RR_ϕ		BRR_ϕ		MRR_ϕ		$BMRR_\phi$	
R^2	$\overline{\phi}$	RB %	RE	RB %	RE	RB %	RE	RB %	RE	RB %	RE	RB %	RE	RB %	RE	RB %	RE
0.4	0.6	3.24	2.32	3.24	2.33	0.70	1.02	0.72	1.00	1.75	1.46	1.75	1.47	0.59	1.00	0.61	1.00
0.4	0.7	3.22	2.41	3.22	2.38	0.67	1.07	0.67	1.00	1.73	1.62	1.73	1.59	0.50	1.06	0.49	1.00
0.4	0.8	3.34	2.73	3.35	2.74	0.69	1.01	0.67	1.00	1.78	1.76	1.79	1.78	0.49	0.99	0.48	1.00
0.4	0.9	3.23	3.01	3.23	3.00	0.59	1.10	0.56	1.00	1.71	1.94	1.71	1.94	0.37	1.03	0.35	1.00
0.5	0.6	2.88	1.90	2.90	1.88	0.90	1.07	0.95	1.00	1.57	1.35	1.57	1.33	0.67	1.04	0.68	1.00
0.5	0.7	2.85	1.87	2.84	1.87	0.84	1.03	0.83	1.00	1.54	1.34	1.55	1.35	0.60	1.01	0.64	1.00
0.5	0.8	2.96	2.11	2.94	2.11	0.87	1.03	0.89	1.00	1.61	1.50	1.61	1.51	0.63	1.01	0.60	1.00
0.5	0.9	2.74	2.23	2.73	2.23	0.69	0.98	0.65	1.00	1.50	1.55	1.48	1.55	0.46	0.98	0.45	1.00
0.6	0.6	2.62	1.54	2.63	1.52	1.13	1.05	1.12	1.00	1.42	1.20	1.44	1.17	0.77	1.04	0.80	1.00
0.6	0.7	2.47	1.57	2.51	1.57	0.91	1.04	0.95	1.00	1.31	1.22	1.32	1.21	0.58	1.07	0.61	1.00
0.6	0.8	2.58	1.66	2.58	1.66	1.03	1.03	1.01	1.00	1.42	1.28	1.42	1.30	0.67	1.02	0.69	1.00
0.6	0.9	2.27	1.79	2.25	1.77	0.70	1.04	0.74	1.00	1.29	1.40	1.28	1.39	0.50	1.04	0.51	1.00

Table 4.4.10: Relative bias (RB %) and Relative efficiency (RE) of four imputed estimators of the distribution function evaluated at the 90% and the 95% quartiles with an average response probability of 60%

		F90								F95							
		RR_ϕ		BRR_ϕ		MRR_ϕ		$BMRR_\phi$		RR_ϕ		BRR_ϕ		MRR_ϕ		$BMRR_\phi$	
R^2	$\overline{\phi}$	RB %	RE	RB %	RE	RB %	RE	RB %	RE	RB %	RE	RB %	RE	RB %	RE	RB %	RE
0.4	0.6	2.47	2.13	2.47	2.12	0.52	1.08	0.55	1.00	1.37	1.42	1.36	1.42	0.48	1.03	0.48	1.00
0.4	0.7	2.46	2.06	2.45	2.06	0.48	1.04	0.51	1.00	1.33	1.44	1.33	1.45	0.37	0.99	0.39	1.00
0.4	0.8	2.56	2.33	2.56	2.33	0.57	1.06	0.55	1.00	1.36	1.58	1.36	1.58	0.37	1.06	0.38	1.00
0.4	0.9	2.45	2.34	2.44	2.32	0.46	1.04	0.44	1.00	1.32	1.58	1.32	1.58	0.31	1.01	0.30	1.00
0.5	0.6	2.22	1.69	2.23	1.69	0.68	1.03	0.70	1.00	1.21	1.21	1.23	1.20	0.52	0.99	0.54	1.00
0.5	0.7	2.14	1.68	2.16	1.67	0.62	1.03	0.64	1.00	1.18	1.27	1.20	1.26	0.47	1.00	0.49	1.00
0.5	0.8	2.23	1.83	2.25	1.83	0.69	1.05	0.69	1.00	1.24	1.38	1.25	1.38	0.52	1.03	0.50	1.00
0.5	0.9	2.07	1.90	2.07	1.90	0.53	1.05	0.48	1.00	1.14	1.40	1.15	1.41	0.37	1.00	0.37	1.00
0.6	0.6	2.05	1.47	2.06	1.45	0.91	1.04	0.91	1.00	1.13	1.15	1.12	1.14	0.63	1.03	0.64	1.00
0.6	0.7	1.92	1.43	1.90	1.42	0.74	0.98	0.72	1.00	1.04	1.14	1.03	1.15	0.50	1.01	0.46	1.00
0.6	0.8	1.95	1.53	1.96	1.52	0.77	1.06	0.78	1.00	1.10	1.23	1.11	1.22	0.54	1.06	0.55	1.00
0.6	0.9	1.71	1.51	1.70	1.51	0.51	0.98	0.52	1.00	1.01	1.28	1.00	1.28	0.42	0.98	0.41	1.00

Table 4.4.11: Relative bias (RB %) and Relative efficiency (RE) of four imputed estimators of the distribution function evaluated at the 90% and the 95% quartiles with an average response probability of 70%

		F90								F95							
		RR_ϕ		BRR_ϕ		MRR_ϕ		$BMRR_\phi$		RR_ϕ		BRR_ϕ		MRR_ϕ		$BMRR_\phi$	
R^2	$\overline{\phi}$	RB %	RE	RB %	RE	RB %	RE	RB %	RE	RB %	RE	RB %	RE	RB %	RE	RB %	RE
0.4	0.6	1.63	1.62	1.62	1.61	0.30	1.07	0.30	1.00	0.93	1.21	0.92	1.20	0.32	1.02	0.31	1.00
0.4	0.7	1.62	1.60	1.62	1.61	0.31	1.06	0.33	1.00	0.89	1.22	0.90	1.22	0.25	1.06	0.26	1.00
0.4	0.8	1.75	1.79	1.74	1.78	0.38	1.07	0.40	1.00	0.93	1.31	0.92	1.30	0.26	1.03	0.29	1.00
0.4	0.9	1.64	1.69	1.64	1.69	0.33	1.03	0.31	1.00	0.89	1.31	0.88	1.31	0.21	1.01	0.21	1.00
0.5	0.6	1.47	1.46	1.47	1.46	0.42	1.01	0.43	1.00	0.82	1.16	0.82	1.16	0.34	1.03	0.35	1.00
0.5	0.7	1.44	1.46	1.45	1.45	0.38	1.05	0.41	1.00	0.81	1.20	0.82	1.19	0.32	1.03	0.34	1.00
0.5	0.8	1.53	1.56	1.52	1.54	0.48	1.03	0.47	1.00	0.87	1.21	0.86	1.20	0.37	1.01	0.36	1.00
0.5	0.9	1.37	1.49	1.38	1.48	0.35	1.01	0.33	1.00	0.78	1.22	0.79	1.22	0.26	1.01	0.25	1.00
0.6	0.6	1.41	1.31	1.39	1.30	0.61	1.00	0.59	1.00	0.79	1.09	0.76	1.08	0.44	1.01	0.43	1.00
0.6	0.7	1.28	1.29	1.27	1.30	0.46	1.03	0.51	1.00	0.71	1.08	0.72	1.09	0.31	1.00	0.34	1.00
0.6	0.8	1.33	1.39	1.33	1.40	0.53	1.02	0.55	1.00	0.76	1.11	0.76	1.11	0.40	0.97	0.40	1.00
0.6	0.9	1.12	1.34	1.12	1.33	0.33	1.02	0.32	1.00	0.69	1.15	0.69	1.15	0.29	0.99	0.28	1.00

Table 4.4.12: Relative bias (RB %) and Relative efficiency (RE) of four imputed estimators of the distribution function evaluated at the 90% and the 95% quartiles with an average response probability of 80%

References

Cardot, H., Chaouch, M., Goga, C., Labruère, C. (2010). Properties of Design-Based Functional Principal Components Analysis. *Journal of Statistical Planning and Inference*, 140, 75–91.

Chauvet, G., Deville, J-C. and Haziza, D. (2011). On random balanced imputation in surveys. *Biometrika*, 98, 459–471.

Chauvet, G. and Haziza, D. (2012). Fully efficient estimation of coefficients of correlation in the presence of imputed survey data. *The Canadian Journal of Statistics*, 40, 124–149.

Chauvet, G. and Tillé, Y. (2006). A fast algorithm for balanced sampling. *Computational Statistics*, 21, 53–62.

Chen, J., Chen, S.-Y. and Rao, J.N.K. (2003). Empirical likelihood confidence intervals for the mean of a population containing many zero values. *The Canadian Journal of Statistics*, 31, 53–68.

Deville, J-C. and Tillé, Y. (2004). Efficient balanced sampling: The cube method. *Biometrika*, 91, 893–912.

Deville, J-C. and Särndal, C.-E. (1994). Variance estimation for the regression imputed Horvitz-Thompson estimator. *Journal of Official Statistics*, 10, 381–394.

Fay, R.E. (1991). A design-based perspective on missing data variance. *Proceedings of the 1991 Annual Research Conference, US Bureau of the Census*, 429–440.

Haziza, D. (2009). Imputation and inference in the presence of missing data. *Handbook of statistics*, 29, 215–246.

Haziza, D., and Nambeu, C-O., and Chauvet, G. (2014), Doubly robust imputation procedures for populations containing a large amount of zeroes in surveys. *Canadian Journal of Statistics*, 42 (4), 650-669.

Haziza, D. and Picard, F. (2012). On doubly robust point and variance estimation in the presence of imputed data. *Canadian Journal of Statistics*, 40, 259–281.

Haziza, D. and Rao, J. N. K. (2006). A nonresponse model approach to inference under imputation for missing survey data. *Survey Methodology*, 32, 53–64.

Isaki, J.K. and Fuller, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89–96.

Kim, J.K. and Fuller, W.A. (2004). Fractional hot-deck imputation. *Biometrika*, 91, 559–578.

Kim, J.K. and Haziza, D. (2014). Doubly robust inference with missing data in survey sampling. *Statistica Sinica*, 24, 375–394.

Rao, J.N.K. and Shao, J. (1992). On variance estimation under imputation for missing data. *Biometrika*, 79, 811–822.

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

Rubin, D.B. (1976). Inference and Missing Data. *Biometrika*, 63, 581–590.

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

Shao, J. and Steel, P. (1999). Variance estimation for survey data with composite imputation and nonnegligible sampling fractions. *Journal of the American Statistical Association*, 94, 254–265.

4.6 Appendix

Proof of equation (4.2.9)

We first prove that

$$\sum_{i\in s}\omega_i r_i \eta_i e_i \;=\; 0. \tag{4.6.1}$$

We have

$$\sum_{i\in s}\omega_i r_i \eta_i v_i^{-1} z_i (y_i - \mathbf{z}_i^\top \hat{B}_r) = 0 \;\Rightarrow\; \sum_{i\in s}\omega_i r_i \eta_i v_i^{-1}(\lambda^\top z_i)(y_i - \mathbf{z}_i^\top \hat{B}_r) = 0, \tag{4.6.2}$$

where λ is defined in (4.1.9). From (4.1.9), this leads to (4.6.1).

We now consider equation (4.2.9). From (4.6.1), we obtain successively

$$\begin{aligned} E_I(y_i^*|\eta_i^*) &= \eta_i^* \mathbf{z}_i^\top \hat{\mathbf{B}}_r^*, \\ E_I(y_i^*) &= \hat{\phi}_i \mathbf{z}_i^\top \hat{\mathbf{B}}_r^*, \end{aligned} \tag{4.6.3}$$

which gives the first line in (4.2.9). Also, since

$$V_I(y_i^*|\eta_i^*) \;=\; \hat{\sigma}^2 v_i \eta_i^* \left\{\sum_{j\in s}\tilde{\omega}_j r_j \eta_j e_j^2\right\}, \tag{4.6.4}$$

we obtain

$$\begin{aligned} V_I(y_i^*) &= V_I E_I(y_i^*|\eta_i^*) + E_I V_I(y_i^*|\eta_i^*) \\ &= V_I\left(\eta_i^* \mathbf{z}_i^\top \hat{\mathbf{B}}_r^*\right) + E_I\left(\hat{\sigma}^2 v_i \eta_i^* \left\{\sum_{j\in s}\tilde{\omega}_j r_j \eta_j e_j^2\right\}\right) \\ &= \hat{\phi}_i(1-\hat{\phi}_i)\left(\mathbf{z}_i^\top \hat{\mathbf{B}}_r^*\right)^2 + \hat{\sigma}^2 \hat{\phi}_i v_i \sum_{j\in s}\tilde{\omega}_j r_j \eta_j e_j^2 \\ &= \hat{\phi}_i(1-\hat{\phi}_i)(\mathbf{z}_i^\top \hat{\mathbf{B}}_r^*)^2 + \hat{\phi}_i v_i \sum_{j\in s}\tilde{\omega}_j r_j \eta_j \left(\frac{y_j - \mathbf{z}_j^\top \hat{\mathbf{B}}_r^*}{\sqrt{v_j}}\right)^2. \end{aligned}$$

Proof of Theorem 4.1

We can write

$$N^{-1}(\hat{t}_{yI} - t_y) = N^{-1}(\hat{t}_{y\pi} - t_y) + N^{-1}(\hat{t}_{yI} - \hat{t}_{y\pi}). \quad (4.6.5)$$

It follows from Assumptions (H1), (H3) and from the model assumptions that

$$E\left[\left\{N^{-1}(\hat{t}_{y\pi} - t_y)\right\}^2\right] = O(n^{-1}). \quad (4.6.6)$$

Therefore, we focus on the second term in the right-hand side of (4.6.5) only, for which we can write $N^{-1}(\hat{t}_{yI} - t_y) = T_1 + T_2 + T_3 + T_4$, with

$$\begin{aligned}
T_1 &= N^{-1}\sum_{i\in s} d_i(1-r_i)(y_i^* - \hat{\phi}_i \mathbf{z}_i^\top \hat{\mathbf{B}}_r^*),\\
T_2 &= N^{-1}\sum_{i\in s} d_i(1-r_i)\hat{\phi}_i \mathbf{z}_i^\top (\hat{\mathbf{B}}_r^* - \boldsymbol{\beta}),\\
T_3 &= N^{-1}\sum_{i\in s} d_i(1-r_i)(\hat{\phi}_i - \phi_i)\mathbf{z}_i^\top \boldsymbol{\beta},\\
T_4 &= N^{-1}\sum_{i\in s} d_i(1-r_i)(\phi_i \mathbf{z}_i^\top \boldsymbol{\beta} - y_i).
\end{aligned}$$

We proceed by showing that $E\{(T_k)^2\} = O(n^{-1})$ for any $k = 1, \ldots, 4$.

Study of the term T_1

We have

$$\begin{aligned}
E_I(T_1) &= N^{-1}\sum_{i\in s} d_i(1-r_i)\{E_I(y_i^*) - \hat{\phi}_i \mathbf{z}_i^\top \hat{\mathbf{B}}_r^*\}\\
&= N^{-1}\sum_{i\in s} d_i(1-r_i)\{\hat{\phi}_i \mathbf{z}_i^\top \hat{\mathbf{B}}_r^* - \hat{\phi}_i \mathbf{z}_i^\top \hat{\mathbf{B}}_r\} = 0, \quad (4.6.7)
\end{aligned}$$

which leads to $E(T_1) = 0$ and

$$E\{(T_1)^2\} = V(T_1) = EV_I(T_1). \quad (4.6.8)$$

Also, since the y_i^*'s are independent conditionally on $\mathbf{y}_U$, $\boldsymbol{\delta}_U$ and $\mathbf{r}_U$, we obtain

$$\begin{aligned} V_I(T_1) &= N^{-2}\sum_{i\in s} d_i^2(1-r_i)V_I(y_i^*) \\ &= N^{-2}\sum_{i\in s} d_i^2(1-r_i)\hat{\phi}_i(1-\hat{\phi}_i)(\mathbf{z}_i^\top\hat{\mathbf{B}}_r^*)^2 + \\ &\quad N^{-2}\sum_{i\in s} d_i^2(1-r_i)\hat{\phi}_i v_i \sum_{j\in s}\tilde{\omega}_j r_j\eta_j\left(\frac{y_j-\mathbf{z}_j^\top\hat{\mathbf{B}}_r^*}{\sqrt{v_j}}\right)^2, \end{aligned} \tag{4.6.9}$$

where the second line in (4.6.9) follows from equation (4.2.9). We first consider the first term in the right-hand side of (4.6.9), that we denote as T_{11}. By using the Cauchy-Schwarz inequality and Assumption (H3), we obtain

$$(\mathbf{z}_i^\top\hat{\mathbf{B}}_r^*)^2 \leq \|z_i\|^2\|\hat{\mathbf{B}}_r^*\|^2 \leq 2(K_2)^2(\|\boldsymbol{\beta}\|^2+\|\hat{\mathbf{B}}_r^*-\boldsymbol{\beta}\|^2), \tag{4.6.10}$$

which leads to

$$\begin{aligned} T_{11} &\leq N^{-2}(K_2)^2\|\boldsymbol{\beta}\|^2\sum_{i\in s} d_i^2 + N^{-2}(K_2)^2\|\hat{\mathbf{B}}_r^*-\boldsymbol{\beta}\|^2\sum_{i\in s} d_i^2 \\ &\leq \frac{(K_2C_1)^2}{n}(\|\boldsymbol{\beta}\|^2+\|\hat{\mathbf{B}}_r^*-\beta\|^2), \end{aligned} \tag{4.6.11}$$

where the second line in (4.6.11) follow from Assumption (H1). From Assumption (H6), we obtain that

$$E(T_{11}) = O(n^{-1}). \tag{4.6.12}$$

We now consider the second term in the right-hand side of (4.6.8), that we denote as T_{12}. By using Assumption (H3), we obtain successively

$$\left(\frac{y_j-\mathbf{z}_j^\top\hat{\mathbf{B}}_r^*}{\sqrt{v_j}}\right)^2 \leq \left(\frac{y_j-\mathbf{z}_j^\top\boldsymbol{\beta}}{\sqrt{v_j}}\right)^2 + K_4(K_2)^2\|\hat{\mathbf{B}}_r^*-\boldsymbol{\beta}\|^2 \tag{4.6.13}$$

and

$$\begin{aligned} T_{12} &\leq K_3N^{-2}\sum_{i\in s} d_i^2\sum_{j\in s}\tilde{\omega}_j\eta_j r_j\left(\frac{y_j-\mathbf{z}_j^\top\boldsymbol{\beta}}{\sqrt{v_j}}\right)^2 + \\ &\quad K_3K_4(K_2)^2\|\hat{\mathbf{B}}_r^*-\boldsymbol{\beta}\|^2N^{-2}\sum_{i\in s} d_i^2. \end{aligned} \tag{4.6.14}$$

Since the sampling design is non-informative and the non-response mechanism is unconfounded, we can write $E(T_{12}) = E_p E_q E_m(T_{12})$. From (4.6.14), we obtain

$$\begin{aligned} E_m(T_{12}) &\leq K_3 N^{-2} \sum_{i\in s} d_i^2 \left\{\sigma^2 + K_4 (K_2)^2 E_m \|\hat{\mathbf{B}}_r^* - \boldsymbol{\beta}\|^2\right\} \\ &\leq \frac{K_3 (C_1)^2}{n} \left\{\sigma^2 + K_4 (K_2)^2 E_m \|\hat{\mathbf{B}}_r^* - \boldsymbol{\beta}\|^2\right\} \qquad (4.6.15) \\ , & \qquad (4.6.16) \end{aligned}$$

where the second line in (4.6.15) follows from Assumption (H1). By using Assumption (H6), this leads to

$$E(T_{12}) = O(n^{-1}). \qquad (4.6.17)$$

From (4.6.8), (4.6.12) and (4.6.17), we obtain $E\{(T_1)^2\} = O(n^{-1})$.

Study of the term T_2

Using the Cauchy-Schwartz inequality we have

$$(T_2)^2 \leq \|N^{-1} \sum_{i\in s} d_i (1 - r_i) \hat{\phi}_i \mathbf{z}_i\|^2 \times \|\hat{\mathbf{B}}_r^* - \boldsymbol{\beta}\|^2. \qquad (4.6.18)$$

Also, from Assumptions (H1) and (H3), we have

$$\|N^{-1} \sum_{i\in s} d_i (1 - r_i) \hat{\phi}_i \mathbf{z}_i\|^2 \leq C_1 K_2. \qquad (4.6.19)$$

Using (4.6.18) and (4.6.19), we obtain

$$E\{(T_2)^2\} \leq C_1 K_2 \times E\|\hat{\mathbf{B}}_r^* - \boldsymbol{\beta}\|^2, \qquad (4.6.20)$$

and from Assumption (H6), $E\{(T_2)^2\} = O(n^{-1})$.

Study of the term T_3

We have $T_3 = \left\{N^{-1} \sum_{i\in s} d_i (1 - r_i)(\hat{\phi}_i - \phi_i)\mathbf{z}_i\right\}^\top \boldsymbol{\beta}$, and by using the Cauchy-Schwartz inequality we obtain

$$(T_3)^2 \leq \|N^{-1} \sum_{i\in s} d_i (1 - r_i)(\hat{\phi}_i - \phi_i)\mathbf{z}_i\|^2 \times \|\boldsymbol{\beta}\|^2. \qquad (4.6.21)$$

Also, by using Assumptions (H1), (H3) and (H5), we have

$$\begin{aligned}\|N^{-1}\sum_{i\in s} d_i(1-r_i)(\hat{\phi}_i-\phi_i)\mathbf{z}_i\| &\leq N^{-1}\sum_{i\in s} d_i(1-r_i)|\hat{\phi}_i-\phi_i|\times\|\mathbf{z}_i\| \\ &\leq N^{-1}K_2\sum_{i\in s} d_i(1-r_i)|\hat{\phi}_i-\phi_i| \\ &\leq N^{-1}K_2K_5\{\sum_{i\in s} d_i(1-r_i)\}\times\|\hat{\gamma}_r-\gamma\| \\ &\leq K_2K_5C_1\|\hat{\gamma}_r-\gamma\|. \qquad (4.6.22)\end{aligned}$$

By plugging (4.6.22) into (4.6.21), we have

$$(T_3)^2 \leq \{K_2K_5C_1\|\boldsymbol{\beta}\|\}^2\times\|\hat{\gamma}_r-\gamma\|, \qquad (4.6.23)$$

and from Assumption (H4) we obtain $E\{(T_3)^2\}=O(n^{-1})$.

Study of the term T_4

We have

$$E_m(T_4) = N^{-1}\sum_{i\in s} d_i(1-r_i)E_m(\phi_i\mathbf{z}_i^\top\boldsymbol{\beta}-y_i)=0, \qquad (4.6.24)$$

which leads to $E(T_4)=0$ and

$$E\{(T_4)^2\}=V(T_4)=EV_m(T_4). \qquad (4.6.25)$$

Also, we have

$$\begin{aligned}V_m(T_4) &= N^{-2}\sum_{i\in s} d_i^2(1-r_i)V_m(y_i) \\ &= N^{-2}\sum_{i\in s} d_i^2(1-r_i)\{V_mE_m(y_i|\eta_i)+E_mV_m(y_i|\eta_i)\} \\ &= N^{-2}\sum_{i\in s} d_i^2(1-r_i)\{\phi_i(1-\phi_i)(\mathbf{z}_i^\top\boldsymbol{\beta})^2+\phi_i\sigma^2v_i\}, \qquad (4.6.26)\end{aligned}$$

and from Assumptions (H1) and (H3), we obtain

$$\begin{aligned}V_m(T_4) &\leq N^{-2}\sum_{i\in s} d_i^2(1-r_i)\{\|\mathbf{z}_i\|^2\|\boldsymbol{\beta}\|^2+\sigma^2|v_i|\}, \\ &\leq N^{-2}\{\sum_{i\in s} d_i^2\}\{(K_2)^2\|\boldsymbol{\beta}\|^2+\sigma^2K_3\}, \\ &\leq \frac{(C_1)^2}{n}\{(K_2)^2\|\boldsymbol{\beta}\|^2+\sigma^2K_3\}. \qquad (4.6.27)\end{aligned}$$

From (4.6.25) and (4.6.27), we obtain $E\{(T_4)^2\} = O(n^{-1})$.

To conclude the proof, we note that

$$\begin{aligned} E\left[\{N^{-1}(\hat{t}_{yI} - t_y)\}^2\right] &= E\left\{(T_1 + T_2 + T_3 + T_4)^2\right\} \\ &\leq 4\left[E\{(T_1)^2\} + E\{(T_2)^2\} + E\{(T_3)^2\} + E\{(T_4)^2\}\right], \end{aligned} \tag{4.6.28}$$

where $E\{(T_k)^2\} = O(n^{-1})$ for any $k = 1, \ldots, 4$. The proof is complete.

Proof of Theorem 4.2

We can write

$$\hat{F}_I(t) - F_N(t) = \left\{\hat{F}_N(t) - F_N(t)\right\} + \left\{\hat{F}_I(t) - \hat{F}_N(t)\right\}. \tag{4.6.29}$$

It follows from Assumption (H1) that

$$E\left[\left\{\hat{F}_N(t) - F_N(t)\right\}^2\right] = O(n^{-1}). \tag{4.6.30}$$

Therefore, we focus on the second term in the right-hand side of (4.6.29) only, namely $\hat{F}_I(t) - \hat{F}_N(t)$, which is the imputation error. In order to study this term, we now describe a way of getting the imputed value y_i^* in (4.2.6) from the true value y_i in (4.1.6) for a non-respondent in three steps.

Firstly, ϵ_i is replaced by a random residual $\hat{\epsilon}_i$, selected with-replacement from the set

$$G_r = \{\epsilon_j \ ; \ r_j = 1 \text{ and } \eta_j = 1\}. \tag{4.6.31}$$

Let $j(i)$ denote the donor selected for unit i. This leads to the value:

$$\begin{aligned} \hat{y}_i &= \eta_i\left\{\mathbf{z}_i^\top \boldsymbol{\beta} + \sigma\sqrt{v_i}\hat{\epsilon}_i\right\} \\ &= \eta_i\left\{\mathbf{z}_i^\top \boldsymbol{\beta} + \sigma\sqrt{v_i}\epsilon_{j(i)}\right\}. \end{aligned} \tag{4.6.32}$$

Secondly, the unknown parameters $\boldsymbol{\beta}$ and σ are estimated, and the exact residual $\hat{\epsilon}_i = \epsilon_{j(i)}$ is replaced in (4.6.32) by the estimated residual $e_{j(i)}$ (see equation 4.2.7). This leads to the value:

$$\begin{aligned} y_i^{**} &= \eta_i \left\{ \mathbf{z}_i^\top \hat{\mathbf{B}}_r^* + \hat{\sigma}\sqrt{v_i} e_{g(i)} \right\} \\ &= \eta_i \left\{ \mathbf{z}_i^\top \hat{\mathbf{B}}_r^* + \hat{\sigma}\sqrt{v_i} \epsilon_i^* \right\}. \end{aligned} \tag{4.6.33}$$

Finally, the unknown indicator η_i is replaced in (4.6.33) by η_i^*. This leads to the final imputed value in (4.2.6). Making use of this decomposition, the imputation error can be written as

$$\hat{F}_I(t) - \hat{F}_N(t) = T_5 + T_6 + T_7, \tag{4.6.34}$$

with

$$T_5 = N^{-1} \sum_{i \in s} d_i (1 - r_i) \left\{ 1(y_i^* \leq t) - 1(y_i^{**} \leq t) \right\}, \tag{4.6.35}$$

$$T_6 = N^{-1} \sum_{i \in s} d_i (1 - r_i) \left\{ 1(y_i^{**} \leq t) - 1(\hat{y}_i \leq t) \right\}, \tag{4.6.36}$$

$$T_7 = N^{-1} \sum_{i \in s} d_i (1 - r_i) \left\{ 1(\hat{y}_i \leq t) - 1(y_i \leq t) \right\}. \tag{4.6.37}$$

The term T_5 in (4.6.35) represents the error due to the replacement of the binary indicator η_i by the generated indicator η_i^*. The term T_6 in (4.6.36) represents the error due to the replacement of the binary $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}$ by estimators, and by the replacement of the true imputed residual by the estimated imputed residual. Lastly, the term T_7 in (4.6.37) represents the error due to the replacement of the original residual ϵ_i by a residual randomly selected in the set of the respondent residuals. We consider these three terms separately.

Error in predicting the zero/non-zero value

We first focus on T_5, for which we prove that

$$E\{(T_5)^2\} = O(n^{-1}). \tag{4.6.38}$$

After some algebra, we can write

$$\begin{aligned} 1(y_i^* \leq t) - 1(y_i^{**} \leq t) &= (\eta_i^* - \eta_i)\{1(\varepsilon_i^* \leq \hat{t}_i) - 1(t \geq 0)\} \qquad (4.6.39) \\ \text{with } \hat{t}_i &= \frac{t - \mathbf{z}_i^\top \hat{\mathbf{B}}_r^*}{\hat{\sigma}\sqrt{v_i}}. \end{aligned}$$

This leads to $(T_5)^2 = T_{51} + T_{52}$, with

$$\begin{aligned} T_{51} &= N^{-2} \sum_{i \in s} d_i^2 (1 - r_i)(\eta_i^* - \eta_i)^2 \{1(\varepsilon_i^* \leq \hat{t}_i) - 1(t \geq 0)\}^2, \qquad (4.6.40) \\ T_{52} &= N^{-2} \sum_{i \neq j \in s} d_i (1 - r_i) d_j (1 - r_j)(\eta_i^* - \eta_i)(\eta_j^* - \eta_j) \times \\ &\quad \{1(\varepsilon_i^* \leq \hat{t}_i) - 1(t \geq 0)\}\{1(\varepsilon_j^* \leq \hat{t}_j) - 1(t \geq 0)\}. \qquad (4.6.41) \end{aligned}$$

From Assumption (H1), we have

$$T_{51} \leq N^{-2} \textstyle\sum_{i \in s} d_i^2 \leq \frac{(C_1)^2}{n}. \qquad (4.6.42)$$

Also, since η_i^*, η_j^*, ε_i^* and ε_j^* are independent with respect to the imputation mechanism, we have:

$$\begin{aligned} E_I(T_{52}) &= N^{-2} \sum_{i \neq j \in s} d_i (1 - r_i) d_j (1 - r_j)(\hat{\phi}_i - \eta_i)(\hat{\phi}_j - \eta_j) \times \\ &\quad \{\hat{F}_{\varepsilon_r}(\hat{t}_i) - 1(t \geq 0)\}\{\hat{F}_{\varepsilon_r}(\hat{t}_j) - 1(t \geq 0)\} \qquad (4.6.43) \end{aligned}$$

where $\hat{F}_{\varepsilon_r}(t) = \sum_{j \in s} \tilde{\omega}_j r_j \eta_j 1(e_j \leq t)$. This leads to

$$\begin{aligned} E_m\{E_I(T_{52}) | \varepsilon_j, j \in s; \eta_g, g \in S_r\} &= N^{-2} \sum_{i \neq j \in s} d_i (1 - r_i) d_j (1 - r_j)(\hat{\phi}_i - \phi_i)(\hat{\phi}_j - \phi_j) \times \\ &\quad \{\hat{F}_{\varepsilon_r}(\hat{t}_i) - 1(t \geq 0)\}\{\hat{F}_{\varepsilon_r}(\hat{t}_j) - 1(t \geq 0)\}, \qquad (4.6.44) \end{aligned}$$

and

$$\begin{aligned} E(T_{52}) &\leq E\left\{ N^{-2} \sum_{i \neq j \in s} d_i (1 - r_i) d_j (1 - r_j) |\hat{\phi}_i - \phi_i| \times |\hat{\phi}_j - \phi_j| \right\} \\ &\leq E\left\{ N^{-2} K_5^2 \parallel \hat{\gamma}_r - \gamma \parallel^2 (\sum_{i \in s} d_i)^2 \right\} \\ &\leq (K_5 C_1)^2 E(\parallel \hat{\gamma}_r - \gamma \parallel^2), \qquad (4.6.45) \end{aligned}$$

where the second line in (4.6.45) follows from Assumption (H5), and the last line in (4.6.45) follows from Assumption (H1). From Assumption H_4, we obtain

$$E(T_{52}) \;=\; O(n^{-1}). \tag{4.6.46}$$

From (4.6.42) and (4.6.46), we obtain (4.6.38).

Error in estimating the regression parameters

We now focus on T_6, for which we prove that

$$E(|T_6|) \;=\; o(1). \tag{4.6.47}$$

After some algebra, we can write

$$1(y_i^{**} \leq t) - 1(\hat{y}_i \leq t) \;=\; \eta_i\{1(\varepsilon_i^* \leq \hat{t}_i) - 1(\hat{\varepsilon}_i \leq t_i)\}, \tag{4.6.48}$$

where

$$t_i = \frac{t - \mathbf{z}_i^\top \boldsymbol{\beta}}{\sigma\sqrt{v_i}} \quad \text{and} \quad \hat{t}_i = \frac{t - \mathbf{z}_i^\top \hat{\mathbf{B}}_r^*}{\hat{\sigma}\sqrt{v_i}}.$$

This leads to

$$T_6 \;=\; N^{-1}\sum_{i\in s} d_i(1-r_i)\eta_i\{1(\varepsilon_i^* \leq \hat{t}_i) - 1(\hat{\varepsilon}_i \leq t_i)\} \tag{4.6.49}$$

and

$$E_I(|T_6|) \;\leq\; N^{-1}\sum_{i\in s} d_i(1-r_i)\eta_i \sum_{j\in s} \tilde{\omega}_j r_j \eta_j \mid 1(e_j \leq \hat{t}_i) - 1(\varepsilon_j \leq t_i) \mid . \tag{4.6.50}$$

We can rewrite

$$e_j \;=\; \frac{\sigma}{\hat{\sigma}}\varepsilon_j - \frac{\mathbf{z}_j^\top(\hat{\mathbf{B}}_r^* - \boldsymbol{\beta})}{\hat{\sigma}\sqrt{v_j}},$$

which leads to

$$e_j \leq \hat{t}_i \;\Leftrightarrow\; \varepsilon_j \leq t_i + \left(\frac{\mathbf{z}_j}{\sqrt{v_j}} - \frac{\mathbf{z}_i}{\sqrt{v_i}}\right)^\top \frac{(\hat{\mathbf{B}}_r^* - \boldsymbol{\beta})}{\sigma} \equiv t_{ij},$$

and from (4.6.50) we obtain

$$E_I(|T_6|) \leq N^{-1} \sum_{i \in s} d_i (1 - r_i) \eta_i \sum_{j \in s} \tilde{\omega}_j r_j \eta_j \mid 1(\varepsilon_j \leq t_{ij}) - 1(\varepsilon_j \leq t_i) \mid \equiv T_6'. \tag{4.6.51}$$

Let us take some constant $\nu > 0$. Since the distribution function F_ε is continuous, there exists some τ_ν such that

$$|t - u| \leq \tau_\nu \quad \Rightarrow \quad |F_\varepsilon(t) - F_\varepsilon(u)| \leq \nu \tag{4.6.52}$$

We note

$$1_A = 1\left(\parallel \hat{\mathbf{B}}_r^* - \boldsymbol{\beta} \parallel \geq \frac{\sigma \tau_\nu}{4 \sup_j \left\| \frac{\mathbf{z}_j}{\sqrt{v_j}} \right\|} \right) \text{ and } 1_B = 1\left(\parallel \hat{\mathbf{B}}_r^* - \boldsymbol{\beta} \parallel < \frac{\sigma \tau_\nu}{4 \sup_j \left\| \frac{\mathbf{z}_j}{\sqrt{v_j}} \right\|} \right). \tag{4.6.53}$$

From assumption (H1), we obtain that $T_6' \leq C_1$, so that

$$T_6' 1_A \quad \leq \quad C_1 1_A,$$

and

$$E(T_6' 1_A) \quad \leq \quad C_1 P\left(\parallel \hat{\mathbf{B}}_r^* - \boldsymbol{\beta} \parallel \geq \frac{\sigma \tau_\nu}{4 \sup_j \parallel \frac{\mathbf{z}_j}{\sqrt{v_j}} \parallel} \right) \tag{4.6.54}$$

$$\leq \quad C_1 \left(\tau_\nu \frac{\sigma}{4 sup_j \left\| \frac{\mathbf{z}_j}{\sqrt{v_j}} \right\|} \right)^{-2} E \parallel \hat{\mathbf{B}}_r^* - \boldsymbol{\beta} \parallel^2 \tag{4.6.55}$$

where the second line in (4.6.54) follows from the Bienaymé -Chebyshev inequality. Therefore, from Assumptions (H3) and (H6), we have

$$E\{T_6' 1_A\} \quad = \quad O(n^{-1}). \tag{4.6.56}$$

Now, note that under $1_B = 1$, we have

$$\begin{aligned} |t_{ij} - t_i| &= \left| \frac{1}{\sigma} (\frac{\mathbf{z}_j}{\sqrt{v_j}} - \frac{\mathbf{z}_i}{\sqrt{v_i}})^\top (\hat{\mathbf{B}}_r^* - \boldsymbol{\beta}) \right| \\ &\leq \frac{1}{\sigma} \left\| \frac{\mathbf{z}_j}{\sqrt{v_j}} - \frac{\mathbf{z}_i}{\sqrt{v_i}} \right\| \times \left\| \hat{\mathbf{B}}_r^* - \boldsymbol{\beta} \right\| \\ &\leq \frac{2}{\sigma} \sup_j \left\| \frac{\mathbf{z}_j}{\sqrt{v_j}} \right\| \times \left\| \hat{\mathbf{B}}_r^* - \boldsymbol{\beta} \right\| \\ &\leq \frac{\tau_\nu}{2}. \end{aligned}$$

Therefore

$$| \, 1(\varepsilon_j \leq t_{ij}) - 1(\varepsilon_j \leq t_i) \, | \, 1_B \leq 1\left(t_i - \frac{\tau_\nu}{2} \leq \varepsilon_j \leq t_i + \frac{\tau_\nu}{2} \right). \quad (4.6.57)$$

From equations (4.6.51) and (4.6.57), we have

$$\begin{aligned} T_6' 1_B &= N^{-1} \sum_{i \in s} d_i (1 - r_i) \eta_i \sum_{j \in s} \tilde{\omega}_j r_j \eta_j \, | \, 1(\varepsilon_j \leq t_{ij}) - 1(\varepsilon_j \leq t_i) \, | \, 1_B \\ &\leq N^{-1} \sum_{i \in s} d_i (1 - r_i) \eta_i \sum_{j \in s} \tilde{\omega}_j r_j \eta_j 1(t_i - \frac{\tau_\nu}{2} \leq \varepsilon_j \leq t_i + \frac{\tau_\nu}{2}) \end{aligned}$$

and

$$\begin{aligned} E_m(T_6' 1_B) &\leq N^{-1} \sum_{i \in s} d_i (1 - r_i) \phi_i \left\{ F_\varepsilon(t_i + \frac{\tau_\nu}{2}) - F_\varepsilon(t_i - \frac{\tau_\nu}{2}) \right\} \\ &\leq \nu N^{-1} \sum_{i \in s} d_i (1 - r_i) \phi_i \\ &\leq C_1 \nu, \end{aligned} \quad (4.6.58)$$

where the second line in (4.6.58) follows from (4.6.52), and the last line in (4.6.58) follows from Assumption (H1). Since ν is arbitrary small, we obtain

$$E\{T_6' 1_B\} = o(1). \quad (4.6.59)$$

From equations (4.6.56) and (4.6.59), we obtain (4.6.47).

Error in replacing the random residuals

Finally, we focus on T_7 for which we prove that

$$E\{(T_7)^2\} = O(n^{-1}). \tag{4.6.60}$$

After some algebra, we can write

$$1(\hat{y}_i \leq t) - 1(y_i \leq t) = \eta_i\{1(\hat{\varepsilon}_i \leq t_i) - 1(\varepsilon_i \leq t_i)\} \text{ with } t_i = \frac{t - \mathbf{z}_i^\top \boldsymbol{\beta}}{\sigma\sqrt{v_i}}. \tag{4.6.61}$$

This leads successively to

$$\begin{aligned}
T_7 &= N^{-1}\sum_{i\in s} d_i(1-r_i)\eta_i\left\{1(\hat{\varepsilon}_i \leq t_i) - 1(\varepsilon_i \leq t_i)\right\}, \\
E_I(T_7) &= N^{-1}\sum_{i\in s} d_i(1-r_i)\eta_i \sum_{j\in s}\tilde{\omega}_j r_j \eta_j \left\{1(\varepsilon_j \leq t_i) - 1(\varepsilon_i \leq t_i)\right\}, \\
E_m\{E_I(T_7)|\eta_i, i\in s\} &= N^{-1}\sum_{i\in s} d_i(1-r_i)\eta_i \sum_{j\in s}\tilde{\omega}_j r_j \eta_j \left\{F_\varepsilon(t_i) - 1(F_\varepsilon(t_i)\right\} = 0, \\
E_m E_I(T_7) &= 0.
\end{aligned} \tag{4.6.62}$$

This leads to $E(T_7) = 0$ and

$$\begin{aligned}
E\{(T_7)^2\} &= V(T_7) \\
&= E_p E_q E_m V_I(T_7) + E_p E_q V_m E_I(T_7).
\end{aligned} \tag{4.6.63}$$

We have

$$\begin{aligned}
V_I(T_7) &= N^{-2}\sum_{i\in s} d_i^2(1-r_i)\eta_i \sum_{j\in s}\tilde{\omega}_j r_j \eta_j \left\{1(\varepsilon_j \leq t_i) - \sum_{k\in s}\omega_k r_k \eta_k 1(\varepsilon_k \leq t_i)\right\}^2 \\
&\leq N^{-2}\sum_{i\in s} d_i^2 \\
&\leq \frac{(C_1)^2}{n},
\end{aligned} \tag{4.6.64}$$

where the last line in (4.6.64) follows from Assumption (H1). Also, we obtain from (4.6.62)

$$\begin{aligned}
V_m\{E_I(T_7)\} &= V_m E_m\{E_I(T_7)|\eta_i, i\in s\} + E_m V_m\{E_I(T_7)|\eta_i, i\in s\} \\
&= E_m V_m\{E_I(T_7)|\eta_i, i\in s\}.
\end{aligned} \tag{4.6.65}$$

From the rewriting

$$\begin{aligned} E_I(T_7) &= N^{-1}\sum_{j\in s}\tilde{\omega}_j r_j \eta_j \sum_{i\in s} d_i(1-r_i)\eta_i 1(\varepsilon_j \le t_i) \\ &- N^{-1}\sum_{i\in s} d_i(1-r_i)\eta_i 1(\varepsilon_i \le t_i), \end{aligned} \tag{4.6.66}$$

we obtain

$$\begin{aligned} V_m\{E_I(T_7)|\eta_i, i\in s\} &= N^{-2}\sum_{j\in s}\omega_j^2 r_j\eta_j V_m\{\sum_{i\in s} d_i(1-r_i)\eta_i 1(\varepsilon_j \le\ t_i)|\eta_i, i\in s\} \\ &+ N^{-2}\sum_{i\in s} d_i^2(1-r_i)\eta_i F_\varepsilon(t_i)\{1-F_\varepsilon(t_i)\} \\ &= N^{-2}(\sum_{i\in s} d_i)^2 \sum_{j\in s}\tilde{\omega}_j^2 r_j\eta_j V_m\left\{\frac{\sum_{i\in s} d_i(1-r_i)\eta_i 1(\varepsilon_j \le\ t_i)}{\sum_{i\in s} d_i}|\eta_i, i\in s\right\} \\ &+ N^{-2}\sum_{i\in s} d_i^2(1-r_i)\eta_i F_\varepsilon(t_i)\{1-F_\varepsilon(t_i)\} \\ &\le N^{-2}(\sum_{i\in s} d_i)^2 \sum_{j\in s}\tilde{\omega}_j^2 + N^{-2}\sum_{i\in s} d_i^2 \\ &\le \frac{(C_1)^2\{(C_2)^2+1\}}{n}, \end{aligned} \tag{4.6.67}$$

where the last line in (4.6.67) follows from Assumption (H1). From (4.6.65) and (4.6.67), we obtain

$$V_m\{E_I(T_7)\} = O(n^{-1}). \tag{4.6.68}$$

From (4.6.63) and (4.6.68), we obtain (4.6.60).

To conclude the proof, we note that

$$E|\hat{F}_I(t) - \hat{F}_N(t)| \le E|T_5| + E|T_6| + E|T_7| \tag{4.6.69}$$

$$\le \sqrt{E\{(T_5)^2\}} + E|T_6| + \sqrt{E\{(T_7)^2\}}, \tag{4.6.70}$$

and from equations (4.6.38), (4.6.47) and (4.6.60) we obtain $E|\hat{F}_I(t) - \hat{F}_N(t)| = o(1)$. The proof is complete.

Chapter 5

Propensity weighting for survey nonresponse through machine learning

The most common way to deal with unit nonresponse is through a weight adjustment procedure. The rationale behind this type of procedures is to eliminate the nonrespondents from the data file and to adjust the design (or basic) weights of the respondents, with the goal of reducing the nonresponse bias. Key to achieving a significant bias reduction is the availability of fully observed variables that are related to both the probability of response to the survey and the survey variables; e.g., Little and Vartivarian (2005) and Haziza and Beaumont (2017).

In practice, two types of weighting procedures are commonly used (Haziza and Lesage, 2016): in the first, the basic weights are multiplied by the inverse of the estimated response probabilities, whereas the second uses some form of calibration, that includes post-stratification and raking as special cases, for adjusting the basic weights. In this chapter, we focus on weight adjustment by the inverse of the estimated response probabilities.

Let $U = \{1, 2, ..., N\}$ be a finite population of size N. In most surveys conducted by statistical agencies, information is collected on a potentially large num-

ber of survey variables and the aim is to estimate many population parameters. This type of surveys is often referred to as multipurpose surveys. Let y be a generic survey variable. We are interested in estimating the finite population total, $t_y = \sum_{i \in U} y_i$, of the y-values. We select a sample S, of size n, according to a sampling design $p(S)$ with first-order inclusion probabilities π_i, $i = 1, \cdots, N$. In the absence of nonresponse, a design-unbiased estimator of t_y is the following expansion estimator:

$$\widehat{t}_{y,\pi} = \sum_{i \in s} w_i y_i, \tag{5.0.1}$$

where $w_i = 1/\pi_i$ denotes the basic weight attached to unit i.

In the presence of unit nonresponse, the survey variables are recorded for a subset S_r of the original sample S. This subset is often referred to as the set of respondents. Let r_i be a response indicator such that $r_i = 1$ if unit i is a respondent and $r_i = 0$, otherwise. We assume that the true probability of response associated with unit i is related to a certain vector of variables $\mathbf{x}_i$; that is, $p_i = P(r_i = 1 \mid S, \mathbf{x}_i)$. We assume that $0 < p_i \leq 1$ and that the response indicators are mutually independent. The latter assumption is generally not realistic in the context of multistage sampling designs because sample units within the same cluster (e.g., household) may not respond independently of one another; see Skinner and D'Arrigo (2011) and Kim et al. (2016) for a discussion of estimation procedures accounting for the possible intra-cluster correlation. If the vector $\mathbf{x}_i$ contains fully observed variables only, then the data are said to be Missing At Random (MAR). However, if the vector $\mathbf{x}_i$ includes variables that are subject to missingness, then the data are Not Missing At Random (NMAR); see Rubin (1976). In practice, it is not possible to determine whether or not the MAR assumption holds. However, the MAR assumption can be made more plausible by conditioning on fully observed variables that are related to both the probability of response and the survey variables; e.g., Little and Vartivarian (2005).

If the response probabilities p_i were known, an unbiased estimator of t_y would be the double expansion estimator (Särndal et al., 1992):

$$\widehat{t}_{y,DE} = \sum_{i \in S_r} \frac{w_i}{p_i} y_i. \tag{5.0.2}$$

In practice, the response probabilities p_i are not known and need to be estimated. To that end, a model for the response indicators r_i, called a nonresponse model, is assumed and the estimated probabilities $\widehat{p}_i$ are obtained using the assumed model (e.g., Särndal and Swensson, 1987; Ekholm and Laaksonen, 1991). This leads to the Propensity Score Adjusted (PSA) estimator:

$$\widehat{t}_{y,PSA} = \sum_{i \in S_r} \frac{w_i}{\widehat{p}_i} y_i, \tag{5.0.3}$$

where $\widehat{p}_i$ is an estimate of p_i. An alternative estimator of t_y is the so-called Hajek estimator:

$$\widehat{t}_{y,HAJ} = \frac{N}{\widehat{N}} \sum_{i \in S_r} \frac{w_i}{\widehat{p}_i} y_i, \tag{5.0.4}$$

where $\widehat{N} = \sum_{i \in S_r} \frac{w_i}{\widehat{p}_i}$ is an estimate of the population size N based on the respondents.

The estimated response probabilities in (5.0.3) or (5.0.4) may be obtained through parametric or nonparametric methods. In the context of parametric estimation, we assume that

$$p_i = f(\mathbf{x}_i, \boldsymbol{\alpha}), \tag{5.0.5}$$

for some function $f(\mathbf{x}_i, .)$, where $\boldsymbol{\alpha}$ is a vector of unknown parameters. The estimated response probabilities are given by

$$\widehat{p}_i = f(\mathbf{x}_i, \widehat{\boldsymbol{\alpha}}),$$

where $\widehat{\boldsymbol{\alpha}}$ is a suitable estimator (e.g., maximum likelihood estimator) of $\boldsymbol{\alpha}$. The class of parametric models (5.0.5) includes the popular linear logistic regression

model as a special case. It is given by

$$p_i = \frac{exp(\mathbf{x}_i^\top \boldsymbol{\alpha})}{1 + exp(1 + \mathbf{x}_i^\top \boldsymbol{\alpha})}.$$

There are several issues associated with the use of a parametric model: (i) they are not robust to the misspecification of the form of $f(\mathbf{x}_i, .)$; (ii) they can fail to account properly on local violations of the parametric assumption such as nonlinearities or interaction effects, both of which may not have been detected during model selection; (iii) they may yield very small estimated response probabilities, resulting in very large nonresponse adjustment factors $\widehat{p}_i^{-1}$, ultimately leading to potentially unstable estimates; e.g., Little and Vartivarian (2005) and Beaumont (2005).

In practice, nonparametric methods are usually preferred essentially because, unlike parametric methods, they protect against the misspecification of the nonresponse model. The class of nonparametric methods include kernel regression (Giommi, 1984 and Da Silva and Opsomer, 2006), local polynomial regression (Da Silva and Opsomer, 2009), weighting classes formed on the basis of preliminary estimated response probabilities (Little, 1986; Eltinge and Yansaneh, 1997 and Haziza and Beaumont, 2007), the CHi square Automatic Interaction Detection (CHAID) algorithm (Kass, 1980) and regression trees (Phipps and Toth, 2012).

In this chapter, we conduct an extensive simulation study to compare several methods for estimating the response probabilities in a finite population setting. For each method, we assess the performance of the propensity score estimator (5.0.3) and the Hajek estimator (5.0.4) in terms of relative bias and relative efficiency. In our study, we attempted to cover a wide range of (parametric and nonparametric) methods; see Hastie et al. (2009) for a comprehensive overview of machine learning methods.

5.1 Nonresponse modeling

Estimating the response probabilities is typically a supervised classification issue, in which the response variable is the two-class categorical response indicator r. However, whereas machine learning methods designed to address classification issues usually focus on optimizing prediction performance, we will less ambitiously restrict our attention to the estimation of the posterior class probabilities. For that issue, in some of the statistical learning methods presented below in the present section, it will be considered as a regression issue in which $r = 0, 1$ is treated as a numeric variable.

5.1.1 Nonparametric Discriminant analysis

Linear logistic regression is often compared to two-class Linear Discriminant Analysis (LDA) since they can both be thought of as different estimations of the same logit-linear regression model, either using maximum-likelihood for linear logistic regression or moment estimation for LDA. LDA originally relies on the assumption that the within-class distributions of the profile $\mathbf{x}$ of explanatory variable is normal with equal variance matrices. Extending LDA to the case of different within-class variance matrices leads to the Quadratic Discriminant Analysis (QDA, see McLachlan 2005). More generally, if $f_r(.)$ stands for the density function of the distribution of $\mathbf{x}$ with class r, for $r = 0,\ 1$, then it is deduced from Bayes' rule that:

$$p_i \ = \ \frac{f_1(x_i)P(r_i = 1)}{f(x_i)},$$

where $f(x) = (1 - P(r_i = 1))f_0(x) + P(r_i = 1)f_1(x_i)$ is the density function of the two-component mixture model with mixing coefficients $1 - P(r_i = 1)$ and $P(r_i = 1)$.

In a classification perspective, once the within-class distributions are estimated, the predicted class is 1 if the corresponding estimation of p_i exceeds a

threshold which is chosen to guarantee a low misclassification rate or a good compromise between true positive and true negative rates. Nonparametric discriminant analysis relies on a nonparametric estimation of group-specific probability densities. Either a kernel method or the k-nearest-neighbor method can be used to generate those nonparametric density estimates. Kernel density estimators were first introduced in the scientific literature for univariate data in the 1950s and 1960s (Rosenblatt 1956, Parzen 1962) and multivariate kernel density estimation appeared in the 1990s (Simonoff 1996). We used a kernel density estimation procedure with normal kernel function, which is the most widely used due to its convenient mathematical properties.

$$\mathcal{K}_r(\mathbf{x_i}) = \frac{1}{(2\pi)^{J/2} d^J |V_r|^{1/2}} exp(-\frac{1}{2d^2} \mathbf{x_i}^\top V_r^{-1} \mathbf{x_i})$$

where J is the number of explanatory variables, d is a fixed radius and V_k the within-group covariance matrix of group r, for $r = 0$ or 1.

5.1.2 Classification and Regression Tree (CART)

Unlike scoring methods such as logistic regression or discriminant analysis that provide a global decision rule in the range of data, decision trees are designed to search for subgroups of data for which the prediction rule is locally adapted. The CART decision tree (Breiman *et al.*, 1984) achieves this partitioning of the data using a binary recursive algorithm: each split of the learning sample is defined by a binary rule, consisting either in thresholding a quantitative variable or forming two sets of levels of a categorical variable. Decision trees have become very popular in machine learning issues because they can handle both continuous and nominal attributes as targets and predictors.

Once a criterion has been chosen to measure the so-called purity of a group of data, the whole learning dataset, viewed as the root of the decision tree, is optimally split into two children nodes (left and right), so that the sum of the purity indices of the two subgroups is as large as possible. Each of the children

node is in turn split following the same goal ... and so on until no further splits are possible due to lack of data. The tree is grown to a maximal size and then pruned back to the root with the method of cost-complexity pruning. Indeed, (Breiman *et al.*, 1984) show that pruning the largest optimal tree produces optimal subtrees of smaller size. Simple or cross-validation assessment of the predictive performance can be used to determine the right size for the decision tree. In order to be able to estimate class probabilities, we choose hereafter to consider $r = 0, 1$ as a numeric variable (which in fact sums up to the use of the Gini index as the impurity measure associated with a unit misclassification cost matrix, see Nakache and Confais, 2003).

Splitting criteria

For each node t which is not terminal, splitting t in two children nodes t_{left} and t_{right} is based on a binary classification rule involving one of the explanatory variables. For each explanatory variable, say x, the binary rule depends on the nature, categorical or numeric, of x. In the case x is nominal, the binary rule just consists in dividing the node t by choosing a group of x levels for t_{left} and the remaining x levels for t_{right}. In the case x is numeric or ordinal, the binary rule consists in a thresholding of x: if the value of x for a given item exceeds a threshold s, then the item goes to t_{left}, otherwise it goes to t_{right}. The best split is obtained by an exhaustive screening of the variables, and for each variable, by optimization of the binary decision rule. For example, if x is numeric, the optimal choice of the threshold s is achieved by minimizing the sum of within-children nodes sum-of-squared deviations to the mean:

$$\sum_{x_i<s}(r_i - \bar{r}_{t_{\text{left}}})^2 + \sum_{x_i \geq s}(r_i - \bar{r}_{t_{\text{right}}})^2$$

Finally, applying the sequence of recursive binary splitting rules to an item based on its values of the explanatory variables assigns this item to one of the terminal node, say t. The corresponding estimated probability that $r = 1$ is just the proportion $\bar{r}_t$ of respondents in t.

Pruning

Consistently with the above splitting algorithm, if $\mathcal{T}$ stands for the set of terminal nodes of a tree T, then the goodness-of-fit of T can be measured by sum-of-squared differences between the observed and fitted values, namely $C(T) = \sum_{t\in\mathcal{T}}(r_i - \bar{r}_t)^2$. The largest possible tree obtained by applying the recursive binary splitting rules until no further split is possible minimizes $C(T)$. This largest tree may overfit the data, which can be detrimental to its prediction performance. Therefore, it is recommended to prune the tree by minimizing the following goodness-of-fit criterion, penalized by the so-called size $|T|$ of the tree, namely the number of its terminal nodes:

$$C_\alpha(T) = \sum_{t\in\mathcal{T}}(r_i - \bar{r}_t)^2 + \alpha|T|$$

where $\alpha > 0$ is a penalty parameter.

For a given value of α, minimizing $C_\alpha(T)$ results in a unique smallest subtree $T_\alpha \subseteq T_0$. Consistently, progressively elevating α produces a sequence of subtrees $T_0 \supseteq T_1 \supseteq ... \supseteq T_L = t_0$, where t_0 is the complete set of items in the sample. The penalty parameter α is usually obtained by minimization of a cross-validated evaluation of the penalized goodness-of-fit criterion for all the subtrees in the sequence or, as suggested in Breiman *et al.*. (1984) to get more stable results, by taking the subtree which cost is one standard-error above the minimal cost.

Surrogate splits CART handles missing data among regressors with surrogate splits. Breiman proposes to define a measure of similarity between the best split of any node t and any other possible split of t built with a regressor taht is not involved in the best split definition. Surrogate splits are computed by searching for splits leading approximately to the same partition of the observations as the original best split.

In section 5.2.1, we will see that this way of choosing the optimal tree by pruning is not appropriate for our final purpose of estimating totals on variables of interest that are subject to missingness.

5.1.3 Conditional Inference Trees for simple and multitarget decision problems

Due to its exhaustive search algorithm for the optimal splitting rules, the above recursive partitioning algorithm has several drawbacks among which overfitting (if not pruned) and selection bias towards covariates with many possible splits. Conditional Inference Trees (Ctree, Hothorn *et al.* 2006) are designed to overcome those two drawbacks by improving the search of the best splitting rules using conditional inference procedures and permutation tests (see Strasser and Weber, 1999).

According to Hothorn *et al.* (2006), conditional inference trees keep the same flexibility as the original tree methods, since they can be applied to different kinds of decision problems, "including nominal, ordinal, numeric, censored **as well as multivariate response variables** and arbitrary measurement scales of the covariates".

Let us assume that, based on a model for the conditional distribution of the response indicator r given a $J-$vector of explanatory variables $\mathbf{x} = (x_1, ..., x_J)^\top$, test statistics can be derived for the significance of the relationship between the response and each of the explanatory variable. As for the standard tree method presented above, the Ctree algorithm to define the optimal splitting rule of a non-terminal node can be divided in two steps:

1. Variable selection: significance of the relationship between the response and each of the explanatory variables is tested, based on random permutations of the response values to obtain a nonparametric estimate of the null distribution of the test statistics. A multiple testing procedure controlling the Family-Wise Error Rate (FWER), such as the Bonferroni correction of the p-values, is then implemented for testing the global null hypothesis H_0 of independence between any of the covariates x_j and the response indicator r. The algorithm is stopped if H_0 cannot be rejected at a pre-specified FWER

control level α. Otherwise the covariate x_{j^*} with the strongest association to r is selected.

2. Optimal split: the best split point for x_{j^*} is also chosen using permutation tests for the significance of the difference between the response rates in the two children nodes.

In the above algorithm, the FWER control level α turns out to be the key parameter to determine the size of the final tree.

Predictions

As with CART, in each cell t which is a terminal node, $\hat{p}_i = \bar{r}_t$.

Missing values in regressors

CTree, as well as CART, handles missing data among regressors which is not the case with logistic regression. Surrogates splits are computed by searching for splits leading approximately to the same partition of the observations as the original best split.

5.1.4 Iterated Multivariate decision trees

Conditional inference trees, introduced in subsection 5.1.3, can also produce decisions rules with several targets at once (see De'ath G 2002 and 2014). Thus, they enable us to provide **groups of items that can be homogeneous regarding a $Q-$vector of target variables $\mathbf{y} = (y_1, \ldots, y_Q)'$ and the response indicator r.** This could be related with the concept of doubly robustness (Bang and Robins 2005, Haziza and Rao 2006).

In the present item nonresponse context, where all the target variables $y_1, ..., y_Q$ are missing for an item with the target $r = 0$, we propose to implement iteratively MultiVariate CTrees. This procedure can be viewed as an **estimation method of p_i, $i = 1...n$ based on successive steps of simultaneous y imputation.**

1. In the first step, the training sample of the multivariate Ctree is based on the sample of respondents only S_r. The targets are $\mathbf{y}$ and the response indicator r. The predictors are J covariates $x_1, ..., x_J$. In case of missing values among the covariates then surrogate rules can be used. Applying on the nonrespondents sample S_{nr} this first decision tree built on S_r, we get $\hat{\mathbf{y}}$ for non respondents sample S_{nr}.

2. In the second step, the training sample of multivariate Ctree contains all items (respondents and nonrespondents) with observed values of $\mathbf{y}$ for respondents and imputed values (from step one) for nonrespondents. We still use the observed values of the response indicator (not those predicted in step 1) to get new values $\hat{\mathbf{y}}$ for non respondents sample S_{nr}.

3. Step 2 is repeated iteratively until $\hat{\mathbf{y}}$ is stabilized. In our simulation study (section 5.3), few iterations have been necessary (less than ten). The final output is the n-vector of estimated response probabilities $\hat{p}_i$'s, $i = 1...n$ for each sample item, provided at the last iteration of multivariate Ctree as the response rate in the terminal node of each item.

This iterated method deals with different patterns of missingness: item nonresponse with imputation of $\mathbf{y}$, unit nonresponse with estimation of response probability and nonresponse among regressors with surrogates rules. It highlights the fact that missingness can be seen as a multivariate problem.

5.1.5 Bagging and Random Forests

Bootstrap aggregating, also called Bagging "is a method for generating multiple versions of a predictor and using these to get an aggregated predictor. The aggre-

gation averages over the versions when predicting a numerical outcome and does a plurality vote when predicting a class." (Breiman 1996).

This machine learning ensemble meta-algorithm is especially beneficial to the notoriously unstable decision tree methods. It is a special case of the model averaging approach, which aim is both to avoid overfitting and to improve the reproducibility and accuracy of machine learning algorithms.

In a general regression problem, bagging averages predictions over a set of bootstrap samples, thereby reducing the variance of a base estimator (e.g., a decision tree). For each bootstrap sample S_b, $b = 1, 2, ..., B$, drawn in the whole learning sample S_n, a model is fitted with a base estimator, giving prediction $\widehat{f}_b(x)$. The bagging estimate of the response probability p_i, $i \in S_n$ is defined by

$$\hat{p}_i = \widehat{f}_{bag}(\mathbf{x}_i) = \frac{1}{B}\sum_{b=1}^{B} \widehat{f}_b(\mathbf{x}_i)$$

Bagging takes advantage of the independence between base learners fitted on different bootstrap samples to reduce the estimation variance while keeping the bias unchanged. It performs best with strong and complex models (e.g., fully developed decision trees), in contrast with boosting methods (see next subsection) that usually work best with weak models (e.g., small decision trees).

Random Forest (Breiman, 2001) is an extension of Bagging applied to regression and classification tree methods, where the main difference with standard Bagging is the randomized covariate selection. Indeed, to optimize each splitting rule, the Random Forest method first randomly selects a subset of covariates, and then apply the usual split selection procedure within the subset of selected covariates. The former additional randomized feature selection is meant to lead to more independent base learners leading to a more efficient variance reduction, in comparison with Bagging. The Random Forest method usually has a worse starting point (when $b = 1$) than Bagging but converges to lower test errors as B increases (Zhou, 2012).

Note that we have chosen to aggregate within a family of learning algorithm, both in Bagging and Random Forest, and not in an overall perspective mixing different families - unlike in stacking (Wolpert 1992, Breiman 1996, Nocairi et al. 2016).

5.1.6 Gradient Boosting and Stochastic Gradient Boosting

Similarly as in the Bagging methods, Boosting aims at taking advantage of a set of classification methods, named learners, to improve the overall classification performance. The original learners are assumed to be just slightly better than random guessing: for this reason, we talk about weak learners. The basic principle of Boosting is to iteratively derive a performant classification rule by selecting a weak learner at each iteration and combine it with the learner derived at the preceding step in such a way that the items with largest prediction errors are especially targeted by the current update of the boosted learner. Boosting was first proposed in the computational learning theory literature (Shapire 1990, Freund 1995, Freund and Shapire 1997) and rapidly became popular since it can result in dramatic improvements in performance.

Friedman *et al.* (2000) give a more statistical perspective to boosting by using the principles of additive modeling and maximum likelihood. Hastie *et al.* (2009) argued that decision trees are ideal base learners for applications of boosting. This motivates our choice of boosting decision trees in our study.

One of the most famous family of boosting methods is Adapative Boosting (AdaBoost, Freund and Shapire, 1996). Hereafter, we present a variant of Adaboost, named Real Adaboost (Freund and Shapire 1996, Schapire and Singer 1999, Friedman et al. 2000), especially suited to the present purpose of estimating response probabilities rather than predicting the membership to the group of respondents. Indeed, at each iteration b, $b = 1, ...B$, the Real AdaBoost algorithm uses weighted class probability estimates $\hat{p}_b(x)$ to build real-value contributions

$f_b(x)$ to the final aggregated rule $F(x)$, i.e. to update the additive model. In the following, the base learners $h_\gamma^{(b)} : \ x \mapsto h_\gamma^{(b)}(x) = \pm 1$ (+ 1 for respondents and -1 for non-respondents) are B decision trees with a number γ of terminal nodes.

Real AdaBoost

Input: Learning sample S_n,
Base learning algorithms $h_\gamma^{(b)}$
Number of iterations B,

Process:

1: Initialize the boosted estimator $F^{(0)}(x) = 0$ and weights $w_i^{(0)} = \frac{1}{n}$, $i \in S_n$

2: **For** $b = 1$ to B **do**

a: Fit $\hat{h}_\gamma^{(b)}$ with the target $\tilde{r_i}$ (where $\tilde{r_i} = 1$ if $r_i = 1$ and $\tilde{r_i} = -1$ if $r = 0$) on the weighted items in the training samples, using weights $w_i^{(b)}$, in order to obtain class probability estimates $\hat{p}_b(x_i)$

c: Update

· $w_i^{(b+1)} = w_i^{(b)} \ exp\{-\tilde{r}_i f_b(x_i)\}$, $i \in S_n$, with $f_b(x_i) = 0.5 \ log\{\frac{\hat{p}_b(x_i)}{1-\hat{p}_b(x_i)}\}$ and renormalize so that $\sum_{i \in S_n} w_i^{(b+1)} = 1$

· $\hat{F}^{(b)}(x) = \hat{F}^{(b+1)}(x) + f_b(x)$

End for

Outputs:

· The classifier $sign[\hat{F}^{(B)}(x)]$ estimates the label

· The estimated probability $\hat{p}(\tilde{r} = 1|x) = \hat{p}(r = 1|x) = \frac{1}{1+exp(-2\hat{F}^{(B)}(x))}$

In our study, the more sophisticated Gradient Boosting and Stochastic Gradient Boosting versions (Friedman 2002, Culp et al. 2006) of Real AdaBoost are implemented.

Gradient Boosting is a mix of gradient descent optimization and boosting. Both Boosting and Gradient Boosting fit an additive model in a forward stage-wise manner. In each stage, they both introduce a weak learner to compensate the shortcomings of previous weak learners. However, Gradient Boosting especially focuses on the minimization of a loss function, here the exponential loss function derived from the maximum-likelihood estimation of a logistic regression model, by identifying those "shortcomings" using gradients, instead of the AdaBoost

weighting function: "Both high-weight data and gradients tells us how to improve the model", (Li 2016). In addition, a regularization parameter is introduced to control at each iteration the weight of the new learners in the current update of the boosted classification method.

The Stochastic Gradient boosting algorithm is referred to as a hybrid bagging and boosting algorithm (Friedman 2002), in the sense that it combines advantages of the two procedures: at each iteration, the new learner is not fitted on the whole learning sample but on a randomly drawn subsample.

5.1.7 The Suppport vector Machine

Support Vector Machines (SVM) are among the most famous machine learning methods in the statistical learning theory presented in Vapnik (1998). In the special case where the p-dimensional space of data points $(x_{i1}, \ldots, x_{ip})$, where x_{ij} is the observation of the jth explanatory variable on the ith sampling item, is fully separable into two subgroups, one with only respondents and one with only non-respondents, using a linear combination of the explanatory variables, then there exists two parallel hyperplanes separating the two subgoups, with maximal distance between those two hyperplanes: this maximal distance is named an hard margin. The maximal-margins hyperplanes contains data points that are called the support vectors. In this special case of separable groups of respondents and non-respondents, the linear SVM classifier consists of considering the position of a data point with respect to the hyperplane that lies in the middle of the maximal-margins hyperplanes to determine the class of an item.

In the general case where the space of data points $(x_{i1}, \ldots, x_{ip})$ is not fully separable, whatever the hyperplane and the margin chosen to separate the two subgroups, any linear classification rule defined as in the fully separable case by the position with respect to a separating hyperplane will result in misclassified data points. A so-called hinge loss function, very similar to the deviance loss function minimized in the maximum-likelihood estimation of a logistic regression

model, is introduced to measure the relevance of a linear classification rule in-between the two maximal-margin hyperplanes. For a given soft margin, finding the optimal hyperplane can be stated as minimizing the mean hinge loss over the learning sample, which is convex optimization issue. The SVM solution finally consists in choosing the best compromise between a low mean hinge loss over the learning sample and a wide margin.

One of the reason why SVM has become so popular is that it can easily be extended to non-linear classification rule, using the so-called "kernel trick" (Schölkopf and Smola 2002). Indeed, in the linear framework, both the mean hinge loss function and the squared inverse of the margin size involve standard scalar products $x_i.x_{i'}$ of data points i and i'. This standard scalar product can be replaced by $K(x_i, x_{i'})$, where K is a symmetric positive definite kernel function (Hastie *et al.*, 2009), that is intentionally introduced to define the similarity of two observations, after a nonlinear transformation of the explanatory variables: to each choice of K corresponds a nonlinear transformation φ such that $K(x_i, x_{i'}) = \varphi(x_i).\varphi(x_{i'})$. For example, the gaussian radial kernel, that is used in the following because it is a "general-purpose kernel used when there is no prior knowledge about the data" (Karatzoglou et al. 2006), is defined as follows:

$$K(\mathbf{x}_i, \mathbf{x}_{i'}) = exp(-\gamma \sum_{j=1}^{J} (x_{ij} - x_{i'j})^2)$$

where γ is a positive constant.

It can be shown that the SVM classifier can be expressed as the sign of a score function $\hat{f}(\mathbf{x})$ which is straightforward deduced from the hinge loss function. Since we are more interested in estimating class probabilities than in predicting class labels, we use Platt's a posteriori probabilities (see Platt, 2000):

$$\widehat{P}(r = 1|\hat{f}(\mathbf{x})) = \frac{1}{1 + exp(A\hat{f}(\mathbf{x}) + B)}$$

where A and B are estimated by minimizing the negative log-likelihood function.

5.2 Modifications of "raw" probabilities estimations

5.2.1 Homogeneous Response Groups (HRG)

The different methods listed above produce "raw" estimated probabilities. The survey weights may be then adjusted inversely to those raw estimated response probabilities. But in order to protect against model insufficiency, it is suggested that homogeneous response groups be formed, i.e. that units with the same characteristics and the same propensity to respond be grouped together (Eltinge and Yansaneh, 1997, Haziza and Beaumont, 2007, Little, 1986). That is why we computed, for each set of "raw" estimated probabilities, a corresponding Homogeneous Response Groups (HRG) version.

Defining HRG requires to partition the population into C groups. The design weight of respondents in group c is adjusted by multiplying it by the inverse of the observed response rate in class c, for $c = 1$ to C. Homogeneous groups are formed by using a clustering algorithm (k-means) on "raw" estimated probabilities. Finally, the probability of a unit in class c is estimated by the response rate observed in the same class.

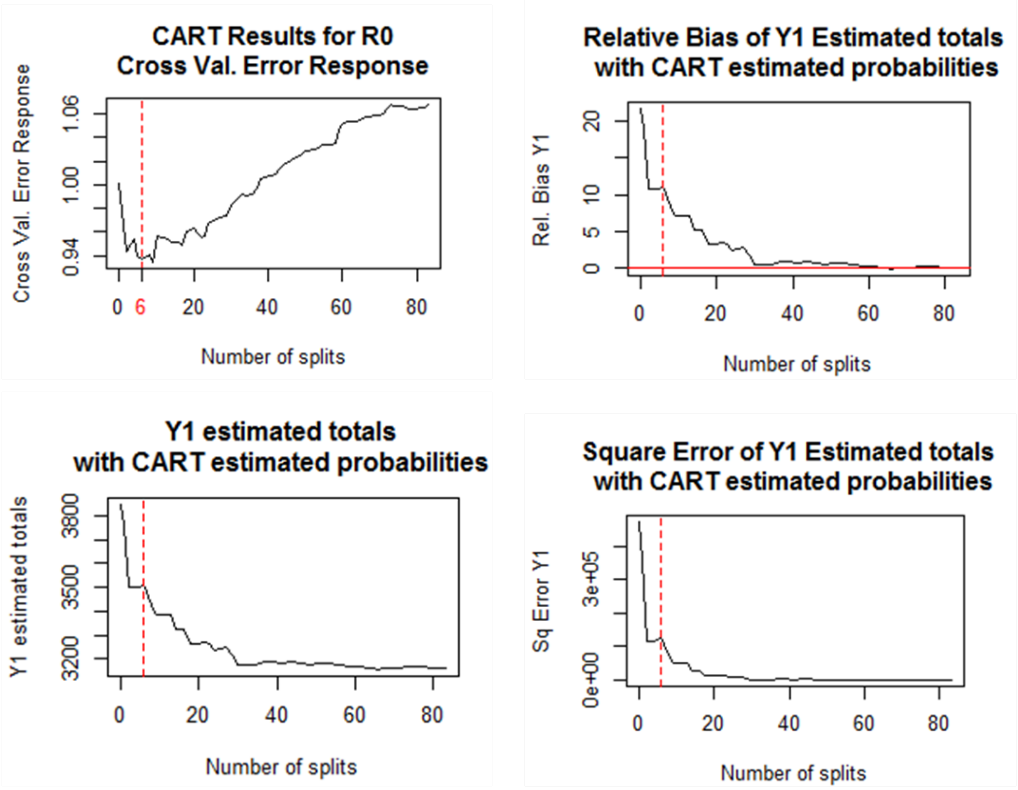
Example of HRG's usefulness with CART:
CART pruning consists in selecting a tree minimizing a cross-validated error (see section 5.1.2). Therefore, the way the learning method is optimized is not especially designed to match our final aim which is to minimize the following expected estimation error:

$$E(t_y - \widehat{t}_y)^2$$

Therefore, in the following simulation study, we propose to extract clusters of homogeneous estimated response probabilities calculated using unpruned trees.

Let us take as example, the variable of interest Y_1 and response mechanism $R0$ described bellow in the simulation study section 5.3. With this example, we

Figure 5.2.1: Performance of CART depending on the number of splits



measure a bias of 11% for the expansion estimation $\hat{t}_{y_{Exp}}$ of t_y in our simulation study with a default pruned CART leading to 6 splits but no bias with an unpruned tree (see figure 5.2.1 below). Furthermore, the SSE of $\hat{t}_{y_{Exp}}$ is much lower with 40 splits than with 6 splits.

5.2.2 Truncation of estimated probabilities

In order to prevent from too small weights, a lower bound has to be determined for the $\hat{p}_i$'s. In practice, the lower bound 0.02 is often used. However, some of our simulations show that the choice of the lower bound may have a certain impact depending on the machine learning method in use. For instance, in our simulations, the global performance of Ctree is robust to variations of the truncation

level, which is not the case with the Bagging version of Ctree (see appendix 5.6.3 for details).

5.3 Simulations study

5.3.1 Simulations set-up

We conduct an extensive simulation study to compare the different methods described in Section 2 in terms of bias and efficiency. We perform $K = 1000$ iterations of the following process: first, a finite population of size $N = 1500$ is generated from a given model. Then, from the realized population, we generate nonresponse according to a specific nonresponse mechanism. Below, we describe one iteration in further details.

We generate a finite population of size $N = 1500$ consisting of ten survey variables, y_j, $j = 1, ..., 10$ and five auxiliary variables x_1-x_5. First, the auxiliary variables were generated as follows. The x_1-values are generated from a standard normal distribution. The x_2-values are generated from a beta parameter with shape parameter equal to 3 and scale parameter equal to 1. The x_3-values are generated from a gamma distribution with shape parameter equal to 3 and scale parameter equal to 2. The x_4-values are generated from a Bernoulli distribution with probability equal to 0.7. Finally, the x_5-values are generated from a multinomial distribution with probabilities $(0.4, 0.3, 0.3)$. We standardize x_2 and x_3 so that their means equal zero and their variances equal one: without loss of generality, it allows us to have more readable coefficients in the definition of the models $M1$ to $M10$ and of response mechanisms $R0$ to $R6$ provided bellow.

Given the values of x_1 to x_5, the values of y_1-y_{10} were generated according to the following models:

$M1$: $y_{i1} = 2 + 2x_{i1} + x_{i2} + 3x_{i3} + \epsilon_{i1}$;

$M2$: $y_{i2} = 2 + 2x_{i1} + x_{i2} + 3x_{i3} + \epsilon_{i2}$;

$M3$: $y_{i3} = 2 + 2x_{i4} + 1.5\ \mathbf{1}_{(x_{i5}=1)} - 2\ \mathbf{1}_{(x_{i5}=2)} + \epsilon_{i3}$;

$M4$: $y_{i4} = 2 + 2x_{i1} + x_{i2} + 3x_{i3} + 2x_{i4} + 1.5\ \mathbf{1}_{(x_{i5}=1)} - 2\ \mathbf{1}_{(x_{i5}=2)} + \epsilon_{i4}$;

$M5$: $y_{i5} = 2 + 2x_{i1} + x_{i2} + 3x_{i3}x_{i4} + 1.5\ \mathbf{1}_{(x_{i5}=1)} - 2\ \mathbf{1}_{(x_{i5}=2)} + \epsilon_{i5}$;

$M6$: $y_{i6} = 2 + 2x_{i1} + x_{i2}^2 + 3x_{i3} + \epsilon_{i6}$;

$M7$: $y_{i7} = 2 + 2x_{i1}^3 + x_{i2}^2 + 3x_{i3}x_{i4} + 1.5\ \mathbf{1}_{(x_{i5}=1)} - 2\ \mathbf{1}_{(x_{i5}=2)} + \epsilon_{i7}$;

$M8$: $y_{i8} = 1 + exp(2x_{i1} + x_{i2} + 3x_{i3}) + \epsilon_{i8}$;

$M9$: $y_{i9} = 1 + x_{i4}exp(2x_{i1} + x_{i2} + 3x_{i3}) + \epsilon_{i9}$;

$M10$: $y_{i10} = 1 + 4cos(x_{i1}) + \epsilon_{i10}$.

As a first step away from our simplest linear model $M1$, for y_2 we only modify the errors: they are generated from a mixture of a standard normal distribution and a beta distribution with shape parameter equal to 3 and scale parameter equal to 1. For the other variables $y_3, ...y_{10}$, the models are more complicated in terms of relations between variables of interest and covariates but the errors ϵ_{ji} are generated from a standard normal distribution.

In order to focus on the nonresponse error, we consider the case of a census that is, $n = N = 1500$. In each population, response indicators are generated according to the following response mechanisms. The response mechanism $R0$ is a logistic model and constitutes the reference model in our empirical study. The other response mechanisms $R1$- $R5$ are expressed as the sum of p_0 and different terms that draw them away from the reference model. The response mechanism $R6$ is built as a regression tree decision rule.

In each population, seven sets of response indicators r_{id} are generated independently from a Bernoulli distribution with parameter p_{id} (i.e. response probabilities), $i = 1, \cdots, N$ and $d = 0, \cdots, 6$, which leads to seven sets of respondents.

$R0$: $p_{i0} = 1/[1 + exp\{-0.4(6.5 + 2x_{i1} + 2x_{i2} + 2x_{i3} - x_{i4} + 1.5\ \mathbf{1}_{(x_{i5}=1)}$
$- 2\ \mathbf{1}_{(x_{i5}=2)} - x_{i3}x_{i4})\}]$;

$R1$: $p_{i1} = 0.65p_{i0} + 0.007x_{i1}^2$;

$R2$: $p_{i2} = 0.5p_{i0} + 0.02 - 0.01x_{i2}^3$;

$R3$: $P_{i3} = 0.5p_{i0} + 0.1|x_{i1}|$;

$R4$: $p_{i4} = 0.5p_{i0} + 0.01 + exp(x_{i2})$;

$R5$: $p_{i5} = 0.5p_{i0} + 0.2 + 0.1\{(sin(x_{i1}) + cos(x_{i2})\}$;

$R6$: $p_{i6} = \mathbf{1}_{(x_{i1}<0)}(0.4 + 0.2x_{i4}) +$
$\mathbf{1}_{(x_{i1}\geq 0)}\mathbf{1}_{(x_{i2}<0.75)}\mathbf{1}_{(x_{i3}<6)}\{0.5\mathbf{1}_{(x_{i5}=1)} + 0.65\mathbf{1}_{(x_{i5}=2)} + 0.7\mathbf{1}_{(x_{i5}=3)}\}$
$+ 0.8\mathbf{1}_{(x_{i1}\geq 0)}\mathbf{1}_{(x_{i2}<0.75)}\mathbf{1}_{(x_{i3}\geq 6)} + 0.9\mathbf{1}_{(x_{i1}\geq 0)}\mathbf{1}_{(x_{i2}\geq 0.75)}$;

Figures presented in Appendix 5.6.1 show the distributions of the simulated values of response probabilities p_{id}, $d = 0, \cdots, 6$. Note that the resulting response rates are approximately 85% for $R0$, 56% for $R1$, 45% for $R2$, 51% for $R3$, 58% for $R4$, 69% for $R5$ and 61% for $R6$. Figures presented in Appendix 5.6.2 illustrate the possibility of non linear links between the response probabilities and the survey variables in our simulations: Hajek's estimator is expected to outperform the expansion estimator in such situations.

We use a truncation for $\hat{p}_i$ with a 0.02 lower bound for all the methods (with or without HRG). As a measure of bias of an estimator $\hat{t}_{y_{(m)}}$ of the finite population parameter t_y, using machine learning method m for response probabilities estimations, we compute the Monte Carlo percent relative bias

$$RB_{MC}(\hat{t}_{y_{(m)}}) = \frac{1}{K}\sum_{k=1}^{K}\frac{(\hat{t}_{y_{(m,k)}} - t_y)}{t_y} \times 100, \tag{5.3.1}$$

where $\hat{t}_{y_{(m,k)}}$ denotes the estimator of t_y in the k-th sample obtained with machine learning method m. As a measure of relative efficiency, we compute

$$RE_{MC}(\hat{t}_{y_{(m)}}) = \frac{MSE_{MC}(\hat{t}_{y_{(m)}})}{MSE_{MC}(\hat{t}_{y_{(HRG\ \ Reglog)}})}, \tag{5.3.2}$$

where $\hat{t}_{y_{(m)}}$ and $\hat{t}_{y_{(HRG\ \ Reglog)}}$ denote respectively the estimator of t_y obtained with method m and the estimator of t_y obtained with Homogenous Response Group applied to logistic regression estimated probabilities, and where

$$MSE_{MC}(\hat{t}_{y_{(m)}}) = \frac{1}{K}\sum_{k=1}^{K}(\hat{t}_{y_{(m,k)}} - t_y)^2.$$

Using $RB_{MC}(\hat{t}_{y_{(m)}})$ and $RE_{MC}(\hat{t}_{y_{(m)}})$ as measures of performance leads to a huge amounts of indicators. Indeed, we have to cross 7 response mechanisms, by 10 variables of interest, 30 methods (with and without HRG versions of 15 machine learning methods) and this for 2 types of estimators $\hat{t}_{y_{Exp}}$ and $\hat{t}_{y_{Haj}}$: 42000 performance indicators. We have to sum up all this information. In order to get a global ranking of the 30 methods for $\hat{t}_{y_{Exp}}$ and $\hat{t}_{y_{Haj}}$, we build two kind of global indicators: one to sum up the RB_{MC} tables and one to sum up the RE_{MC} tables of each machine learning method.

5.3.2 Relative Bias results

Global indicator of relative bias

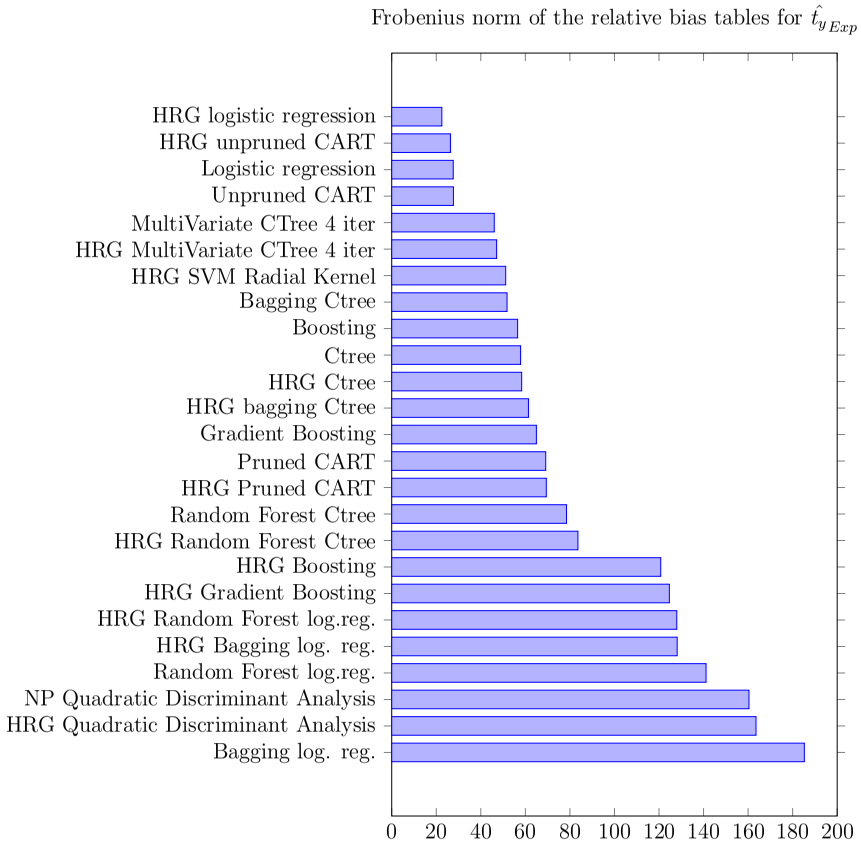
For each machine learning method, we have a RB_{MC} table containing **70 indicators** (10 rows for the 10 variables of interest and 7 columns for the 7 response mechanisms) that can be **summed up by one indicator : the Frobenius norm of the $\mathbf{RB_{MC}}$ table**. The definition of the Frobenius norm of a matrix T, with all its elements in $\mathbb{R}$, is $\|T\|_F = \sqrt{Tr(T^\top T)}$. We want to identify the methods with the lowest relative bias. Thus we look for the methods for which the Frobenius norm of relative bias tables are the smallest. Once we get the global

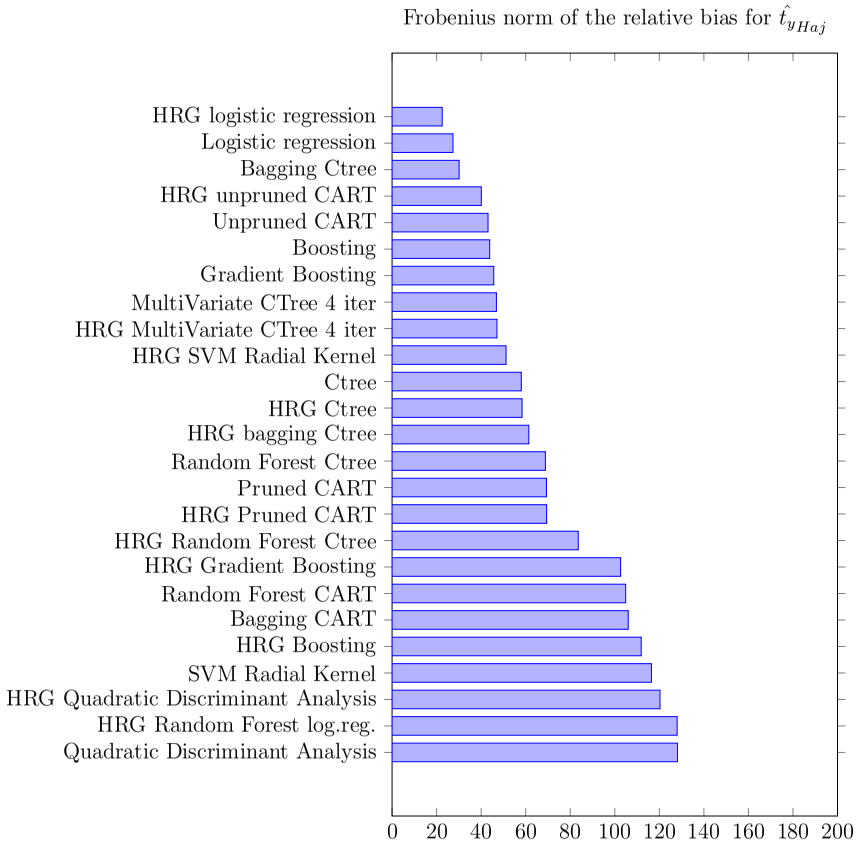
ranking of the methods based on this norm, we can go into more details for the best methods.

Global ranking results

In terms of relative bias results summed up with $\|RB_{MC}\|_F$ (figure 5.3.1), the best method is HRG logistic regression for both $\hat{t}_{y_{Exp}}$ and $\hat{t}_{y_{Haj}}$. However, among the methods that could handle missing values in predictors, HRG unpruned CART is good and performs better than unpruned CART (and much better than default pruned CART and than HRG prunned CART). Bagging Ctree (which also could handle missing values in predictors) performs also quite good but better for $\hat{t}_{y_{Haj}}$ than for $\hat{t}_{y_{Exp}}$. As shown in figure 5.3.1, the four best methods for $\hat{t}_{y_{Exp}}$ provide lower bias than the four best for $\hat{t}_{y_{Haj}}$. We also can see that applying HRG reduces bias for the very best methods (logistic regression and Unpruned CART) but it is not the case for all the methods (see for instance Bagging Ctree and MultiVariate CTrees). Note that in figure 5.3.1, the most extreme values have been removed for a better readability: only the 25 best methods (among 30) are provided.

Figure 5.3.1: Frobenius norm of the relative bias tables for $\hat{t}_{y_{Exp}}$ and $\hat{t}_{y_{Haj}}$





a. RB_{MC}**: Focus on the three best methods for** $\hat{t}_{y_{Exp}}$

a.1 ***HRG logistic regression (Table 5.3.1, Frobenious norm = 22.5)***

Among the 70 scenarios, 30 show unbiased $\hat{t}_{y_{Exp}}$ (bias $< 1\%$) and 9 scenarios exhibit bias above 4%. The best results occur with $R0$ (reference response mechanism i.e. logit link) and $R6$ (decision tree response mechanism). The worse results occur with $R2$ (reference response mechanism + a quadratic term) and $R4$ (reference response mechanism + an exponential term). The highest bias equals -7.7 with $Y7$ (model with quadratic, cubic and interaction terms) and $R5$ (reference response mechanism + sine and cosine terms).

a.2 ***HRG Unpruned CART (Table 5.3.2, Frobenious norm = 26.36)***

Among the 70 scenarios, 17 show unbiased $\hat{t}_{y_{Exp}}$ (bias $< 1\%$) and 13 scenarios exhibit bias above 4%. The best results occur with $R0$ (reference response mechanism i.e. logit link) and $R6$ (decision tree response mechanism). The worse results occur with $R2$ (reference response mechanism + a quadratic term) and $R3$ (reference response mechanism + an absolute value term). The highest bias equal -8.49% with $Y10$ and $R2$ (reference response mechanism + a quadratic term) and -7.22% with $Y10$ (model with a cosine term) and $R3$ (reference response mechanism + an absolute value term).

a.3 ***Logistic (Table 5.3.3, Frobenious norm = 27.62)***

Among the 70 scenarios, 27 show unbiased $\hat{t}_{y_{Exp}}$ (bias $< 1\%$) and 14 scenarios exhibit bias above 4%. The best results occur with $R0$ (reference response mechanism i.e. logit link) and $R6$ (decision tree response mechanism). The worse results occur with $R2$ (reference response mechanism + a quadratic term) and $R4$ (reference response mechanism + an exponential term). The highest relative bias equals 9.28% with $Y7$ (model quadratic, cubic and interaction terms) and $R4$ (reference response mechanism + an exponential term).

Table 5.3.1: Relative bias of $\hat{t}_{y_{Exp}}$ with HRG after logistic regression

Variable	R0	R1	R2	R3	R4	R5	R6
Y1	0.82	3.65	5.30	3.35	4.60	2.49	- 0.33
Y2	0.60	2.68	3.85	2.44	3.33	1.84	- 0.23
Y3	- 0.12	0.19	0.69	0.78	0.70	0.35	- 0.04
Y4	0.35	2.40	3.77	2.77	3.39	1.88	- 0.24
Y5	0.11	2.91	5.28	3.62	4.67	2.35	- 0.80
Y6	0.07	1.71	3.41	0.84	4.33	- 1.46	2.27
Y7	0.85	2.79	5.91	- 0.22	6.99	- 7.68	- 2.32
Y8	1.19	- 1.29	- 1.51	0.63	- 3.06	2.55	- 0.26
Y9	0.72	- 0.86	- 1.80	- 0.23	- 2.57	1.95	- 0.92
Y10	0.15	0.03	0.64	- 4.70	1.21	- 0.99	0.16

Table 5.3.2: Relative bias of $\hat{t}_{y_{Exp}}$ with HRG after unpruned CART

Variable	R0	R1	R2	R3	R4	R5	R6
Y1	2.53	2.33	- 2.13	- 0.95	2.61	2.29	- 0.88
Y2	1.86	0.75	- 4	- 2.37	0.58	1.39	- 1.26
Y3	0.36	- 2.81	- 7.56	- 5.07	- 3.66	- 0.23	- 1.94
Y4	1.82	0.91	- 3.70	- 2.06	0.65	1.62	- 1.11
Y5	2.61	3.32	- 1.13	0.09	3.66	2.76	- 1.36
Y6	0.77	- 2.05	- 4.68	- 4.81	- 4.03	- 0.34	- 1.10
Y7	3.41	2.12	0.70	- 2.87	0.47	1.21	- 0.43
Y8	3.78	3.49	- 1.02	- 1.38	5.32	5.18	2.44
Y9	3.14	5.23	- 4.77	- 0.36	3.61	4.37	3.68
Y10	- 0.02	- 3.72	- 8.49	- 7.22	- 4.80	- 0.71	- 2.57

Table 5.3.3: Relative bias of $\hat{t}_{y_{Exp}}$ with Logistic regression

Variable	R0	R1	R2	R3	R4	R5	R6
Y1	0.06	3.56	5.25	2.95	4.08	2.61	- 0.71
Y2	0.04	2.55	3.77	2.15	3.07	1.85	- 0.42
Y3	- 0.21	- 0.01	0.71	0.76	1.16	0.15	0.00
Y4	- 0.17	2.32	3.78	2.51	3.29	1.88	- 0.59
Y5	- 0.70	3.06	5.42	3.18	4.18	2.53	- 1.16
Y6	- 0.03	0.51	3.23	0.66	6.80	- 1.86	2.46
Y7	- 0.25	1.69	5.93	- 0.85	9.28	- 8.42	- 5.80
Y8	- 0.11	- 7.28	- 6.80	- 1.14	- 3.54	- 0.38	0.72
Y9	- 0.49	- 6.22	- 6.50	- 1.52	- 2.90	- 0.82	- 0.13
Y10	0.01	0.27	1.29	- 5.13	1.60	- 0.59	- 0.62

b. RB_{MC}: Focus on the three best methods for $\hat{t}_{y_{Haj}}$

b.1 ***HRG logistic regression (Table 5.3.4, Frobenious norm = 22.5)***

Among the 70 scenarios, 30 show unbiased $\hat{t}_{y_{Exp}}$ (bias $< 1\%$) and 9 scenarios exhibit bias above 4%. The best results occur with $R0$ (reference response mechanism i.e. logit link) and $R6$ (decision tree response mechanism). The worse results occur with $R2$ (reference response mechanism + a quadratic term) and $R4$ (reference response mechanism + an exponential term). The highest relative bias equals -7.68% with $Y7$ (model with quadratic, cubic and interaction terms) and $R5$ (reference response mechanism + sine and cosine terms).

b.2 ***Logistic regression (Table 5.3.5, Frobenious norm = 27.36)***

Among the 70 scenarios, 28 show unbiased $\hat{t}_{y_{Exp}}$ (bias $< 1\%$) and 12 scenarios exhibit bias above 4%. The best results occur with $R0$ (reference response mechanism i.e. logit link). The worse results occur with $R2$ (reference response mechanism + a quadratic term). The highest relative bias equals 8.81% with $Y7$ (model with quadratic, cubic and interaction terms) and $R4$ (reference response mechanism + an exponential term).

b.3 ***Bagging Ctree (Table 5.3.6, Frobenious norm = 30.11)***

Among the 70 scenarios, 23 show unbiased $\hat{t}_{y_{Exp}}$ (bias $< 1\%$) and 21 scenarios exhibit bias above 4%. The best results occur with $R0$ (reference response mechanism i.e. logit link) and $R6$ (decision tree response mechanism). The worse results occur with $R2$ (reference response mechanism + a quadratic term). The highest relative bias equals -8.62% with $Y10$ (model with a cosine term) and $R2$ (reference response mechanism + a quadratic term).

Table 5.3.4: Relative bias of $\hat{t}_{y_{Haj}}$ with HRG after logistic regression

Variable	R0	R1	R2	R3	R4	R5	R6
Y1	0.82	3.65	5.30	3.35	4.60	2.49	-0.33
Y2	0.60	2.68	3.85	2.44	3.33	1.84	-0.23
Y3	-0.12	0.19	0.69	0.78	0.70	0.35	-0.04
Y4	0.35	2.40	3.77	2.77	3.39	1.88	-0.24
Y5	0.11	2.91	5.28	3.62	4.67	2.35	-0.80
Y6	0.07	1.71	3.41	0.84	4.33	-1.46	2.27
Y7	0.85	2.79	5.91	-0.22	6.99	-7.68	-2.32
Y8	1.19	-1.29	-1.51	0.63	-3.06	2.55	-0.26
Y9	0.72	-0.86	-1.80	-0.23	-2.57	1.95	-0.92
Y10	0.15	0.03	0.64	-4.70	1.21	-0.99	0.16

Table 5.3.5: Relative bias of $\hat{t}_{y_{Haj}}$ with Logistic regression

Variable	R0	R1	R2	R3	R4	R5	R6
Y1	0.08	3.85	5.33	2.93	3.64	2.85	-0.98
Y2	0.06	2.84	3.85	2.14	2.63	2.09	-0.70
Y3	-0.20	0.27	0.78	0.74	0.73	0.39	-0.28
Y4	-0.15	2.61	3.86	2.50	2.85	2.11	-0.87
Y5	-0.68	3.35	5.50	3.16	3.74	2.77	-1.43
Y6	-0.02	0.78	3.30	0.65	6.34	-1.64	2.18
Y7	-0.23	1.98	6.00	-0.87	8.81	-8.22	-6.06
Y8	-0.09	-7.02	-6.75	-1.16	-3.96	-0.15	0.44
Y9	-0.47	-5.96	-6.44	-1.55	-3.33	-0.59	-0.40
Y10	0.03	0.55	1.36	-5.15	1.17	-0.36	-0.90

Table 5.3.6: Relative bias of $\hat{t}_{y_{Haj}}$ with Ctree Bagging

Variable	R0	R1	R2	R3	R4	R5	R6
Y1	1.46	5.26	6.30	5.78	5.44	1.85	0.67
Y2	1.13	3.90	4.54	4.22	3.99	1.36	0.50
Y3	-0.05	0.73	0.81	0.89	0.97	0.59	-0.47
Y4	0.84	3.86	4.45	4.31	4.20	1.75	-0.13
Y5	0.16	5.16	6.39	6.20	5.81	2.03	-0.11
Y6	2.32	3.67	4.29	3.45	4.13	2.09	0.54
Y7	2.81	7.65	9.13	7.08	9.34	3.01	0.01
Y8	2.86	0.67	1.32	4.84	2.22	4.04	2.16
Y9	2.41	1.19	0.83	3.02	2.56	3.71	1.35
Y10	-0.88	-0.02	0.39	-3.02	0.38	0.03	-0.11

5.3.3 Relative Efficiency results

Global indicator of relative efficiency

In the definition of relative efficiency RE_{MC} (equation 5.3.2), we explicitly use the logistic regression combined with HRG as the reference method. It is not the case in the definition of relative bias RB_{MC} (equation 5.3.1). That is why we propose a different global indicator of performance, normalized to 1 for the logistic regression combined with HRG.

Let us denote $RE_{MC}(e, m)$ the table computed for:
- e the estimator type of t_y's, $e \in \{\hat{t_{y}}_{Exp}, \hat{t_{y}}_{Haj}\}$,
- m the machine learning method used to estimate response probabilities.

Note that the model m can either be a machine learning used alone to estimate probabilities or a machine learning method associated to the Homogeneous Response Group creation (see section 5.2.1).

We compute the following normalized indicator (based on the Frobenius norm):

$$NREF_{(e,m)} = \|RE_{MC}(e, m)/RE_{MC}(e, \text{HRG logistic regression})\|_F/8.3666$$

where $RE_{MC}(e, m)/RE_{MC}(e, \text{HRG logistic regression})$ is a term by term division of $RE_{MC}(e, m)$ by $RE_{MC}(e, \text{HRG logistic regression})$. The denominator 8.3666 is the Frobenius norm of a 10×7 matrix filled with $1's$: it is the Frobenius norm of the table $RE_{MC}(e, \text{HRG logistic regression})/RE_{MC}(e, \text{HRG logistic regression})$.

Global ranking results

In the Bar plot (Figure 5.3.2) the most extreme values of $NREF$ have been removed for a better readability. The best methods are HRG logistic regression, Logistic regression HRG Unpruned CART and Unpruned CART for both $\hat{t}_{y_{Exp}}$ and $\hat{t}_{y_{Haj}}$. However among the methods that could handle missing values in predictors, MultiVariate CTree with four iterations is not far, particularly for $\hat{t}_{y_{Exp}}$.

a. RE_{MC}: Focus on the three best methods for $\hat{t}_{y_{Exp}}$

HRG logistic regression is a common used method and appears to the best rank among all the machine learning methods we used. That is why we used it as the reference: data table 5.3.7 is used as denominator in RE_{MC} computation for all the other methods. Consequently, it's RE_{MC} table is filled with 1's only which leads to a Frobenius norm equal to 3.87 and a Normalized Frobenius norm equal to 1. Thus we rather provide here the MSE_{MC} table. In the following table, we darkened the worse cases for each variable of interest (the maximal value in each row). It shows for each variable of interest, on which response mechanism HRG logistic regression performs the best (always $R0$ i.e. the reference response mechanism with logit link)) and the worse ($R2$ i.e. the reference response mechanism + a quadratic term for $Y1$ to $Y5$, $R3$ for $Y8$ to $Y10$ for instance).

Let us focus now on the two other best methods in terms of RE_{MC}.

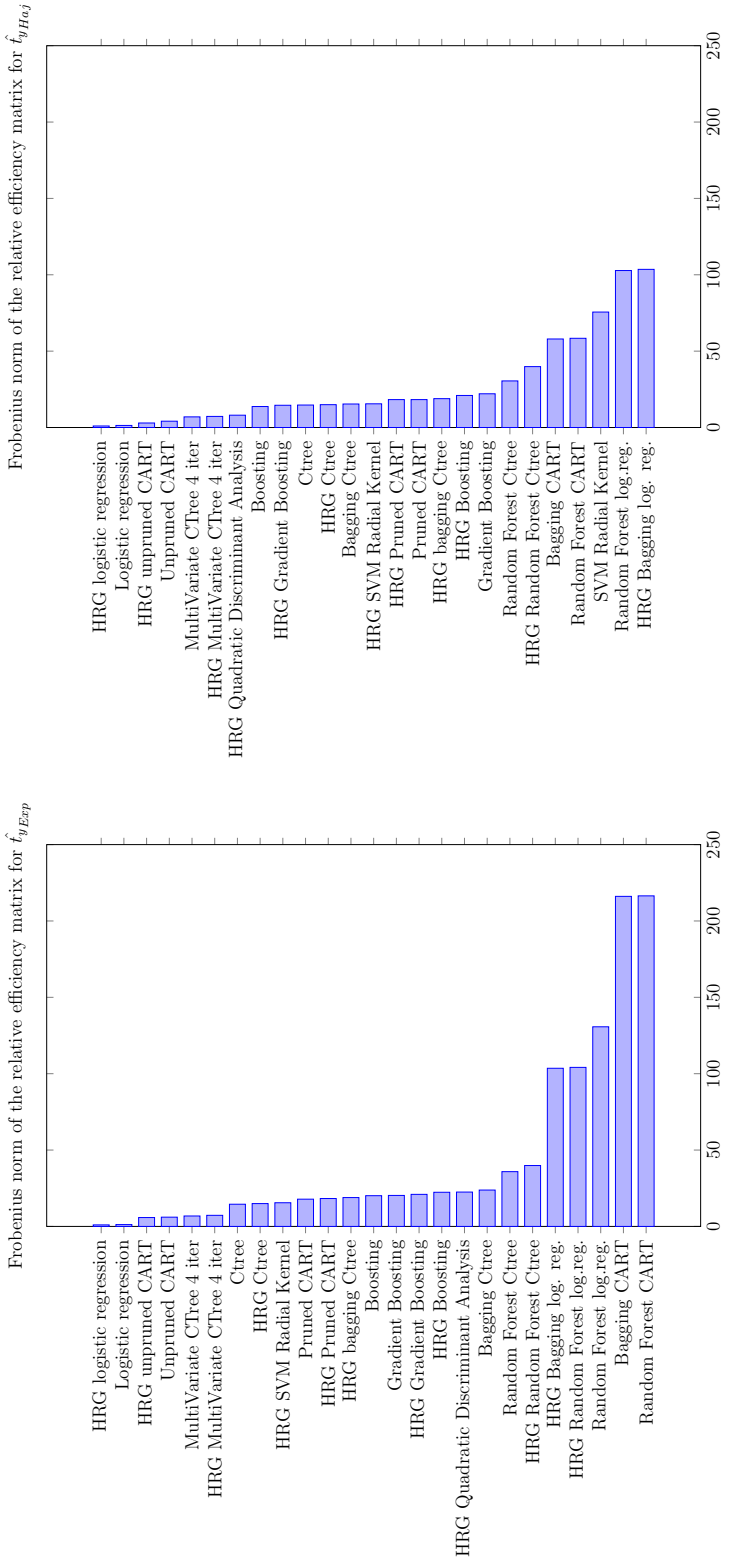


Figure 5.3.2: Normalized Frobenius norm of the relative efficiency tables for $\hat{t}_{y_{Exp}}$ and $\hat{t}_{y_{Haj}}$

a.1 ***Logistic regression (Table 5.3.8, Normalized Frobenious norm = 1.2)***

Among the 70 scenarios, Logistic regression outperforms HRG Logistic regression in 36 scenarios ($RE_{MC} < 1$)) and is much worse in 3 scenarios with $RE_{MC} > 2$. The relative best outperformances of Logistic regression occur with $R1$ (logit response mechanism with non normal residuals) and $R2$ (reference response mechanism + a quadratic term). The worse underperformances occur with $Y5$ and $R0$ (reference response mechanism i.e. logit link): $RE_{MC} = 3.41$ which means that the MSE of Logistic regression is more than three times the one of HRG Logistic regression.

a.2 ***HRG Unpruned CART (Table 5.3.9, Normalized Frobenius norm = 5.8)***

Among the 70 scenarios, HRG Unpruned CART outperforms HRG Logistic regression in 18 scenarios ($RE_{MC} < 1$) and is much worse in 25 scenarios with a RE_{MC} higher than 2. Relative underperformances occur with $R0$ to $R6$. The highest RE_{MC} equals 29.41 with $Y10$ and $R2$ (reference response mechanism + a quadratic term) and 20.69 with $Y3$ and $R2$.

Table 5.3.7: MSE_{MC} for $\hat{t}_{y_{Exp}}$ with HRG logistic regression

Variable	R0	R1	R2	R3	R4	R5	R6
Y1	1.40E+03	2.14E+04	3.86E+04	2.30E+04	3.07E+04	1.12E+04	8.99E+03
Y2	1.42E+03	2.17E+04	3.94E+04	2.30E+04	3.02E+04	1.19E+04	8.98E+03
Y3	6.51E+02	4.14E+03	8.10E+03	6.62E+03	5.71E+03	3.27E+03	2.27E+03
Y4	1.33E+03	2.66E+04	5.33E+04	3.50E+04	4.27E+04	1.67E+04	9.75E+03
Y5	1.10E+03	1.70E+04	4.00E+04	2.59E+04	3.12E+04	1.10E+04	8.06E+03
Y6	2.72E+03	2.01E+04	4.05E+04	1.80E+04	5.16E+04	1.29E+04	2.15E+04
Y7	3.30E+04	9.16E+04	1.48E+05	1.01E+05	1.57E+05	1.79E+05	5.90E+04
Y8	1.96E+17	3.08E+20	3.94E+20	4.38E+20	2.04E+20	3.42E+19	1.79E+20
Y9	1.68E+17	2.83E+20	3.60E+20	4.23E+20	2.00E+20	3.27E+19	1.75E+20
Y10	1.54E+03	5.79E+03	7.11E+03	6.51E+04	9.55E+03	5.22E+03	4.11E+03

Table 5.3.8: Relative efficiency for $\hat{t}_{y_{Exp}}$ with logistic regression

Variable	R0	R1	R2	R3	R4	R5	R6
Y1	1.52	0.91	0.97	0.89	0.90	1.03	1.05
Y2	1.43	0.89	0.96	0.90	0.93	0.97	1.02
Y3	1.33	0.86	0.93	1.03	1.57	0.87	1.03
Y4	1.97	0.90	1.00	0.88	0.99	0.97	1.10
Y5	3.41	0.97	1.01	0.88	0.92	1.05	1.03
Y6	1.23	0.63	0.92	0.93	2.33	1.13	1.11
Y7	1.80	0.70	0.98	1.05	1.65	1.19	2.67
Y8	0.01	1.25	0.56	1.03	0.88	0.48	0.97
Y9	0.01	1.28	0.53	1.05	0.88	0.48	0.97
Y10	1.04	0.97	1.72	1.16	1.36	0.79	1.43

Table 5.3.9: Relative efficiency for $\hat{t}_{y_{Exp}}$ with HRG after unpruned CART

Variable	R0	R1	R2	R3	R4	R5	R6
Y1	7.02	1.05	0.70	0.95	0.90	1.51	2.27
Y2	7.23	0.85	1.30	1.35	0.72	1.31	2.73
Y3	3.30	8.12	20.69	12.52	8.86	3.17	10.67
Y4	10.16	0.89	1.25	1.16	0.65	1.36	3.16
Y5	9.40	1.70	0.65	0.88	1.15	1.92	2.57
Y6	3.09	1.74	1.80	3.98	1.27	1.60	1.25
Y7	2.20	1.52	1.00	1.06	0.89	0.66	2.11
Y8	0.08	0.09	0.57	1.84	1.89	1.97	0.63
Y9	0.05	0.02	0.53	1.88	1.72	2.04	0.63
Y10	1.56	8.78	29.41	2.37	8.00	1.87	7.87

b. RE_{MC}: Focus on the three best methods for $\hat{t}_{y_{Haj}}$

Here again, HRG logistic regression is used as the reference (denominator in RE_{MC} computation). In the following table, we darkened the worse cases for each variable of interest (the maximal value in each row). It shows that HRG logistic regression performs the best with $R0$ (reference response mechanism) and the worse with $R2$ (reference response mechanism + a quadratic term) for $Y1$ to $Y6$, $R3$ for $Y8$ to $Y10$.

Let us focus now on the two other best methods in terms of RE_{MC}.

b.1 ***Logistic regression (Normalized Frobenius norm = 1.4)***

Among the 70 scenarios, logistic regression outperforms HRG Logistic regression in 31 scenarios ($RE_{MC} < 1$) and is much worse in 5 scenarios with a RE_{MC} higher than 2. The relative best outperformances occur with $R2$ (reference response mechanism + a quadratic term) and the worse underperformances with $R0$ (reference response mechanism). The highest RE_{MC} equals 4.57 with $Y5$ and $R0$ (reference response mechanism).

b.2 ***HRG Unpruned CART (Normalized Frobenius norm = 2.9)***

Among the 70 scenarios, HRG Unpruned CART outperforms HRG Logistic regression in 12 scenarios ($RE_{MC} < 1$) and is much worse in 42 scenarios with a RE_{MC} higher than 2. The relative best outperformances occur with $Y8$ and $Y9$ and worse underperformances occur with $R0$ (reference response mechanism). The highest RE_{MC} equals 10.37 with $Y5$ and $R0$ (reference response mechanism).

Table 5.3.10: MSE_{MC} for $\hat{t}_{y_{Haj}}$ with HRG logistic regression

Variable	R0	R1	R2	R3	R4	R5	R6
Y1	1.40E+03	2.15E+04	3.86E+04	2.30E+04	3.08E+04	1.12E+04	9.00E+03
Y2	1.42E+03	2.17E+04	3.95E+04	2.30E+04	3.03E+04	1.19E+04	8.98E+03
Y3	6.52E+02	4.15E+03	8.11E+03	6.62E+03	5.72E+03	3.27E+03	2.28E+03
Y4	1.33E+03	2.66E+04	5.33E+04	3.50E+04	4.27E+04	1.67E+04	9.76E+03
Y5	1.10E+03	1.70E+04	4.00E+04	2.59E+04	3.13E+04	1.11E+04	8.07E+03
Y6	2.73E+03	2.02E+04	4.05E+04	1.80E+04	5.16E+04	1.29E+04	2.15E+04
Y7	3.30E+04	9.17E+04	1.49E+05	1.01E+05	1.58E+05	1.80E+05	5.90E+04
Y8	1.97E+17	3.08E+20	3.95E+20	4.39E+20	2.04E+20	3.43E+19	1.79E+20
Y9	1.68E+17	2.83E+20	3.60E+20	4.24E+20	2.00E+20	3.27E+19	1.75E+20
Y10	1.54E+03	5.80E+03	7.11E+03	6.51E+04	9.56E+03	5.22E+03	4.12E+03

Table 5.3.11: Relative efficiency of $\hat{t}_{y_{Haj}}$ with logistic regression

A	R0	R1	R2	R3	R4	R5	R6
Y1	2.30	1.01	0.99	0.90	0.81	1.16	1.14
Y2	2.47	1.02	0.98	0.91	0.80	1.14	1.13
Y3	1.33	0.88	0.95	0.99	1.07	0.92	1.09
Y4	3.57	1.04	1.02	0.88	0.83	1.14	1.27
Y5	4.57	1.08	1.03	0.89	0.83	1.18	1.14
Y6	1.12	0.64	0.94	0.91	2.00	1.00	0.99
Y7	1.89	0.74	0.99	1.07	1.51	1.16	2.81
Y8	0.03	1.27	0.56	1.03	0.88	0.49	0.97
Y9	0.03	1.30	0.53	1.05	0.87	0.49	0.97
Y10	1.30	1.14	1.83	1.17	1.01	0.82	1.90

Table 5.3.12: Relative efficiency of $\hat{t}_{y_{Haj}}$ with HRG after unpruned CART

Variable	R0	R1	R2	R3	R4	R5	R6
Y1	7.14	2.82	1.97	2.28	2.66	2.13	2.41
Y2	7.38	2.83	1.90	2.27	2.67	2.04	2.46
Y3	3.24	2.69	2.03	2.03	2.21	2.27	3.74
Y4	10.37	3.48	2.19	2.35	2.75	2.13	3.01
Y5	9.53	4.42	2.40	2.59	3.26	2.65	2.35
Y6	3.12	1.64	1.77	1.56	0.59	1.49	1.15
Y7	2.21	2.46	2.72	1.27	1.37	0.72	2.19
Y8	0.08	0.10	0.67	2.09	2.10	1.97	0.71
Y9	0.05	0.03	0.63	2.12	1.89	2.04	0.70
Y10	1.53	1.43	1.61	0.17	0.91	1.12	1.86

5.4 Discussion

In this chapter, we conducted a comprehensive simulation study, aiming at a global ranking of different machine learning methods in totals t_y estimation performance through response probabilities estimation. In our simulation set-up with a census context, the best method in terms of MSE is the logistic regression associated with Homogeneous Response Groups creation. This is true both for the expansion estimator and for the Hajek estimator. One drawback of this method is that it does'nt handle missing data among regressors. Unpruned CART associated with Homogeneous Response Groups creation appear among the methods with good performance and that could handle missing values among regressors, particularly with the expansion estimator. Note that those two first methods turn out to be very robust against changes in lower bound truncation of estimated probabilities. Bagging Ctree (which also could handle missing values among regressors) outperforms Unpruned CART associated to Homogeneous Response Groups creation with the Hajek estimator. However, it seems to require a higher level of truncation than the usual 0.02 value.

In further researches, we would like to study deeper our proposed iterated version of multivariate Ctree whose performances are quite good. For instance, which variables of interest pattern makes the Iterated MultiVariate CTrees work or fail ? Furthermore, this method could maybe prove useful in a context of imputation. Another interesting field would be evaluating the performance of the different machine learning methods with missing data among the regressors. We could also enlarge the set of model aggregation with stacking for instance (Wolpert 1992, Breiman 1996, Nocairi et al. 2016). And lastly, evaluating the methods with different complex sampling designs could bring useful information.

5.5 References

Agresti, A. (2013). *Categorical Data Analysis*. New York: Wiley-Interscience.

Bang, H., Robins, J. M. (2005). Doubly Robust Estimation in Missing Data and Causal Inference Models. *Biometrics*, 61, 962–973.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2), 123–140.

Breiman, L. (1996). Stacked Regression, *Machine Learning*, 24(1), 49-64.

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5-32.

Breiman, L., Friedman, J. H., Olshen, R. A., Stone, C. J. (1984). *Classification and regression trees*. Wadsworth & Brooks/Cole Advanced Books & Software.

Chang, T., Kott, P. (2008). Using Calibration Weighting to Adjust for Nonresponse under a Plausible Model. *Biometrika*, 95(3), 555-571.

Cortes, C., Vapnik, V. (1995). Support-Vector networks. *Machine Learning*, 20, 273-297.

De'ath, G., (2002). Multivariate Regression Trees: A New Technique for Modeling Species- Environment Relationships. *Ecology*, 83(4), 1105-1117.

De'ath G (2014). *mvpart: Multivariate Partitioning*. R package version 1.6-2, URL http: //CRAN.R-project.org/package=mvpart.

Breiman, L. (1996). Bagging Predictors. *Machine Learning*, 24, 123-140

Cortes, C., Vapnik, V. (1995). Support-vector networks. *Machine Learning*. 20(3), 273-297.

Culp, M., Johnson, K., Michailidis, G. (2006). ada: An R Package for Stochastic Boosting. *Journal of Statistical Software*, 17.

Freund, Y. (1995). Boosting a weak learning algorithm by majority. *Information and computation*, 121, 256-285.

Freund, Y., Schapire, R.E. (1996). Experiments with a new boosting algorithm. In *Machine Learning*: Proceedings of the thirteenth International Conference, 148-156. Morgan Kaufman, San Francisco.

Freund, Y., Schapire, R.E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, 55(1), 119-139

Friedman, J., Hastie, T., Tibshirani, R. (2000). Additive logistic regression: a statistical view of Boosting. *The annals of statistics*, 28(2), 337-407.

Friedman, J. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189-1232.

Friedman, J. (2002). Stochastic Gradient Boosting, *Computational Statistics and Data Analysis*, 38(4), 367-378.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, second edition.

Haziza, D. (2009). Imputation and inference in the presence of missing data. *Handbook of statistics*, 29, 215-246.

Haziza, D., Beaumont, J.F. (2007.) On the construction of imputation classes in surveys. *International Statistical Review*, 75(1),25-43.

Haziza, D. and Rao, J. N. K. (2006). A nonresponse model approach to inference under imputation for missing survey data. *Survey Methodology*, 32(4), 53-64.

Ho, T.K. (1995). Random Decision Forests. *Proceedings of the 3rd International Conference on Document Analysis and Recognition*, Montreal, QC, 14-16, 278-282.

Hosmer, D.W., Lemeshow, S. (2000). *Applied logistic regression.* Wiley Series in Probability and Mathematical Statistics.

Hothorn, T., Hornik, K., Zeileis, A. (2006). Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics*, 15(3), 651-674.

Geoffrey, J. McLachlan (2005). *Discriminant Analysis and Statistical Pattern Recognition.* Wiley.

Karatzoglou, A., Meyer, D., Hornik, K. (2006). Support Vector Machines in R. *Journal of Statistical Software*, 15(9).

Li, C. (2016). *A Gentle Introduction to Gradient Boosting.*

URL: http://www.ccs.neu.edu/home/vip/teach/MLcourse/4_ boosting/slides/gradient_boosting.pdf.

Nakache, J.P., Confais, J. (2003). *Statistique explicative appliquée*, Technip, 206-211.

Niculescu-Mizil, A., Caruana, R. (2005). Obtaining Calibrated Probabilities from Boosting. *Uncertainty in Artificial Intelligence.*

Niculescu-Mizil, A., Caruana, R. (2005). Predicting Good Probabilities with Supervised Learning. *ICML.*

Phipps, P., Toth, D. (2012). Analyzing establishment nonresponse using an interpretable regression tree model with linked administrative data. *Annals of Applied Statistics*, 6(2), 772-794.

Nocairi, H., Gomes, C., Thomas, M., Saporta, G. (2016) Improving Stacking Methodology for Combining Classifiers; Applications to Cosmetic Industry. *Electronic Journal of Applied Statistical Analysis*, 9(2), 340-361.

Parzen, E. (1962). On estimation of a probability density function and mode. *Annals of Mathematical Statistics*, 33, 1065–1076.

Platt, J. (2000). Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. In A. Smola, P. Bartlett, B. Schölkopf, and D. Schuurmans (Eds.), *Advances in large margin classifiers*. Cambridge: MIT Press.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics*, 27, 832–837.

Rubin, D.B. (1976). Inference and Missing Data. *Biometrika*, 63, 581-590.

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

Schapire, R.E. (1990). The Strength of Weak Learnability. *Machine Learning*, Boston, MA: Kluwer Academic Publishers, 5 (2), 197-227.

Schapire, R.E., Singer, Y. (1999). Improved boosting algorithms using confidence-rated predictions. *Machine learning*, 37(3), 297-336.

Simonoff, J.S. (1996). *Smoothing Methods in Statistics*. Springer.

Schölkopf B., Smola A. (2002). *Learning with Kernels*. MIT Press.

Strasser, H., Weber, C. (1999). On the Asymptotic Theory of Permutation Statistics. *Mathematical Methods of Statistics*, 8, 220–250.

Vapnik, V. (1998). *Statistical Learning Theory*. Wiley.

Wolpert, D. (1992). Stacked Generalization. *Neural Networks*, 5, 41-259

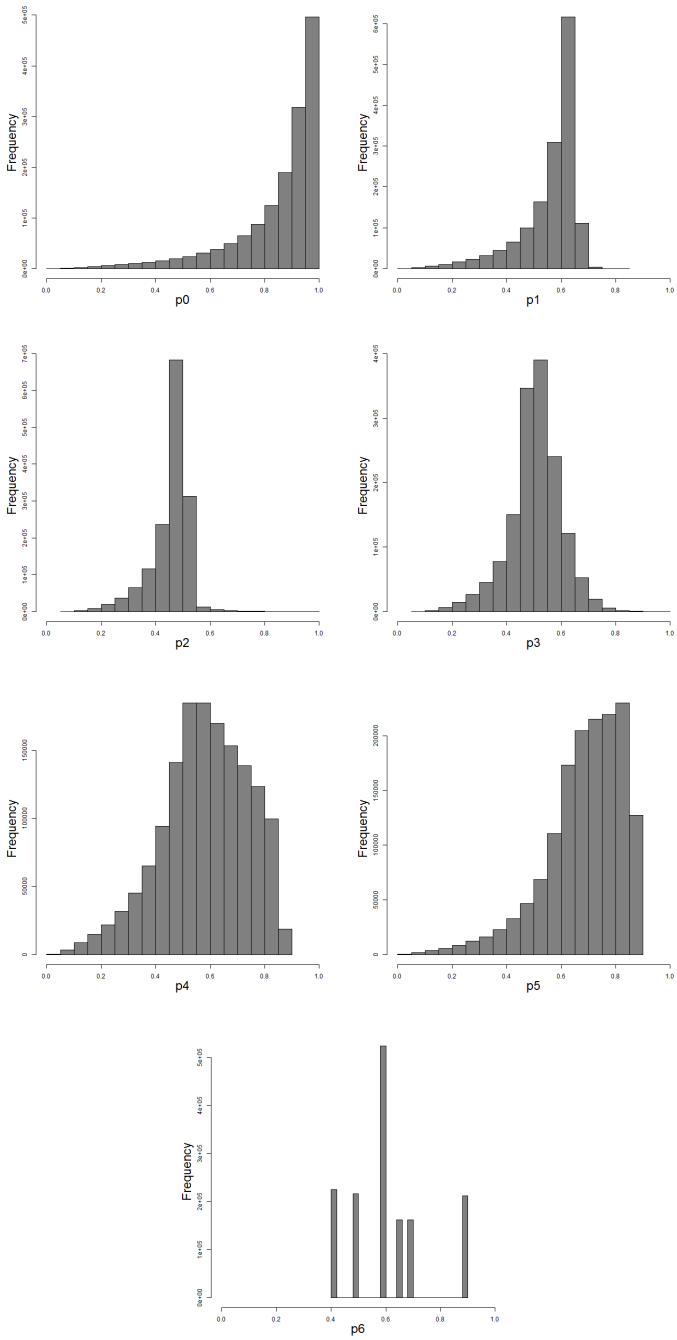
Zadrozny, B., Elkan, C. (2001). Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. *ICML*, 1, 609-616.

Zhou, Z.-H., (2012). *Ensemble Methods: Foundations and Algorithms*. Chapman & Hall/CRC.

5.6 Appendix

5.6.1 Distributions of the generated response probabilities

Distributions on the $K \times N = 1000 \times 1500$ units for the seven response mechanisms



5.6.2 Plots between response probabilities ($p0$ to $p6$) and variables of interest ($Y1$ to $Y10$)

Figure 5.6.1: Scatter plots of $p0$ and variables of interest

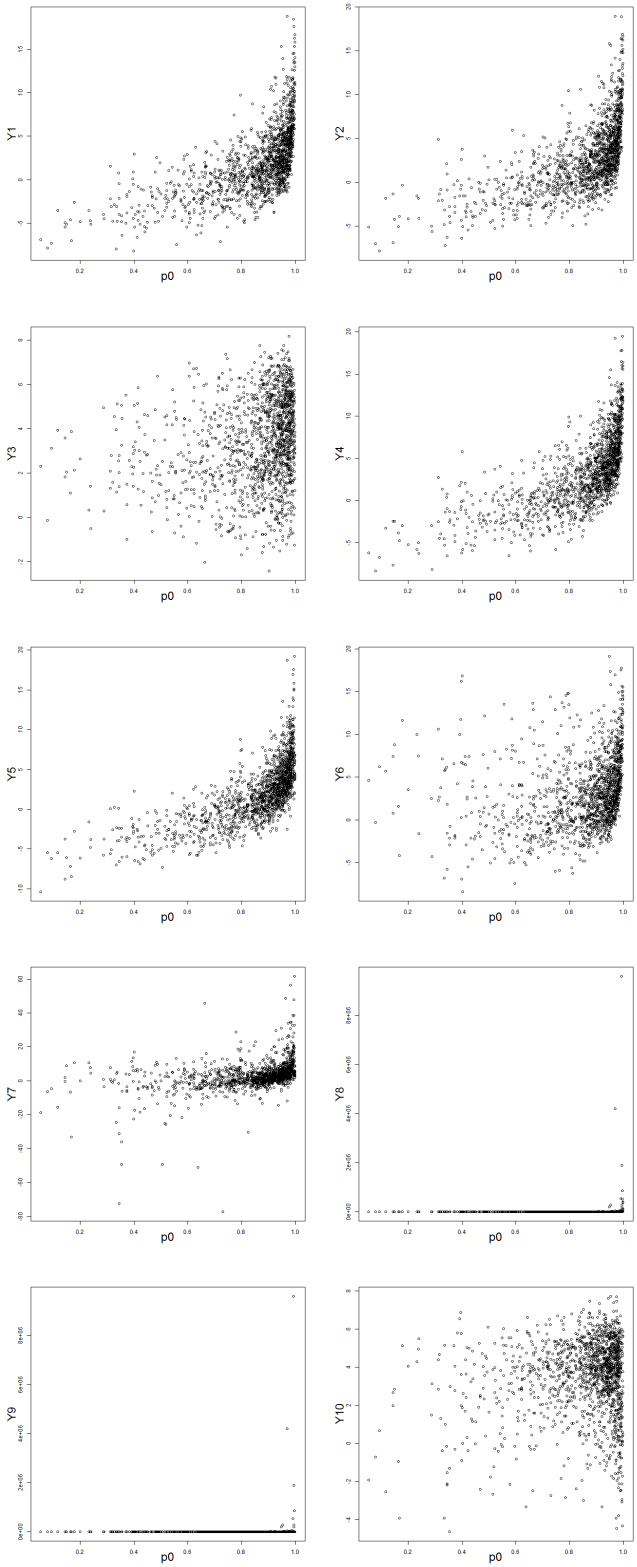


Figure 5.6.2: Scatter plots of $p1$ and variables of interest

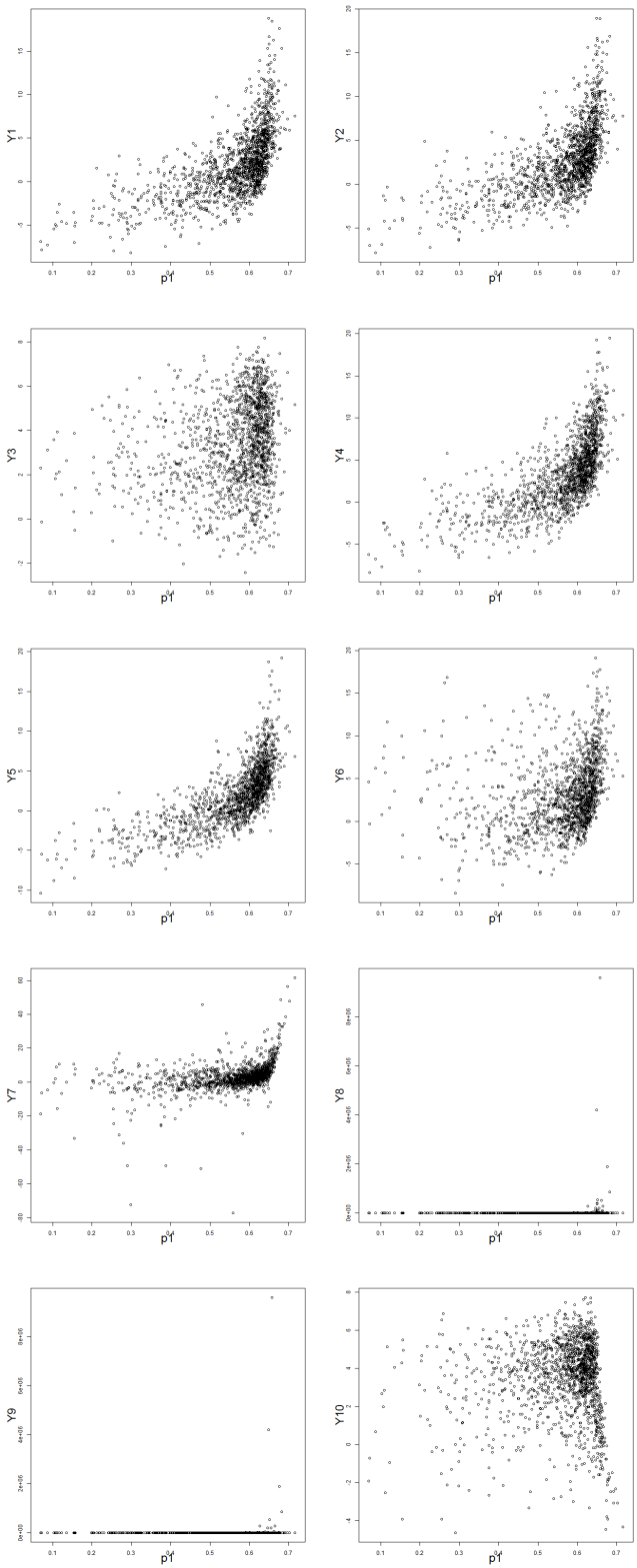


Figure 5.6.3: Scatter plots of $p2$ and variables of interest

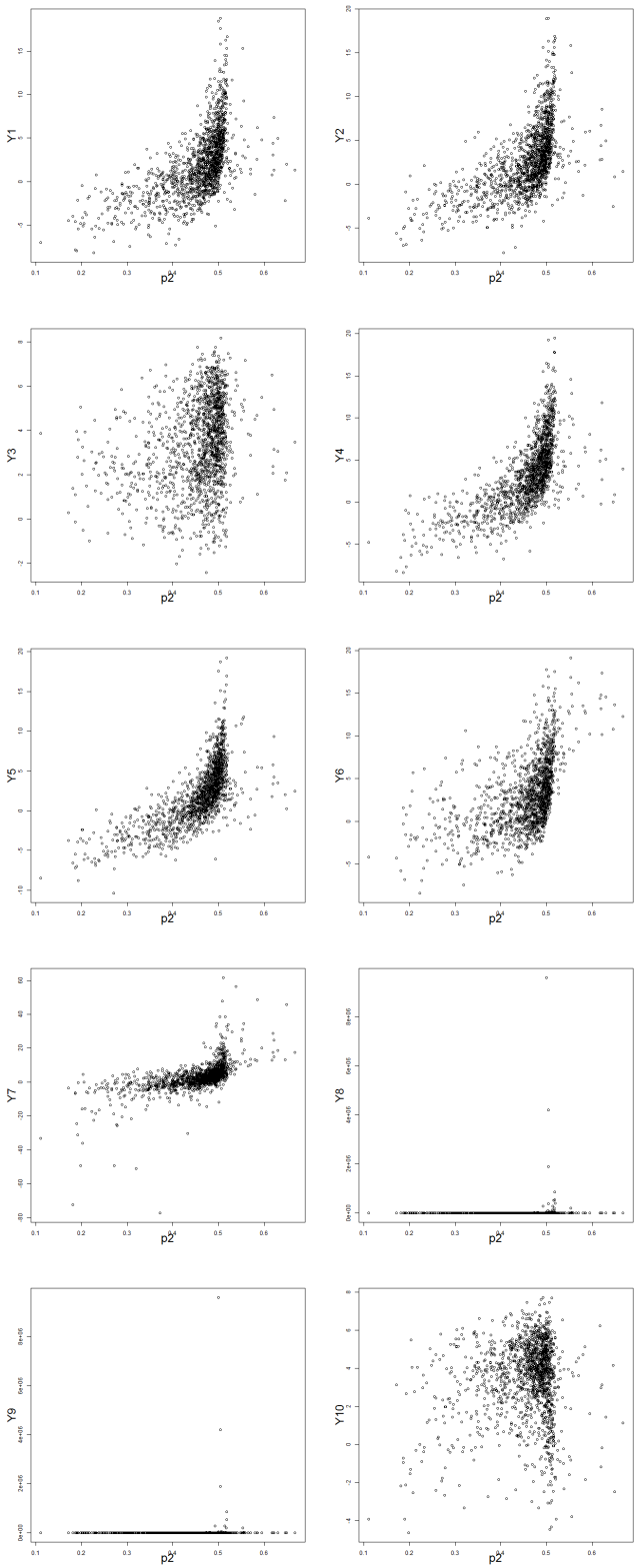


Figure 5.6.4: Scatter plots of $p3$ and variables of interest

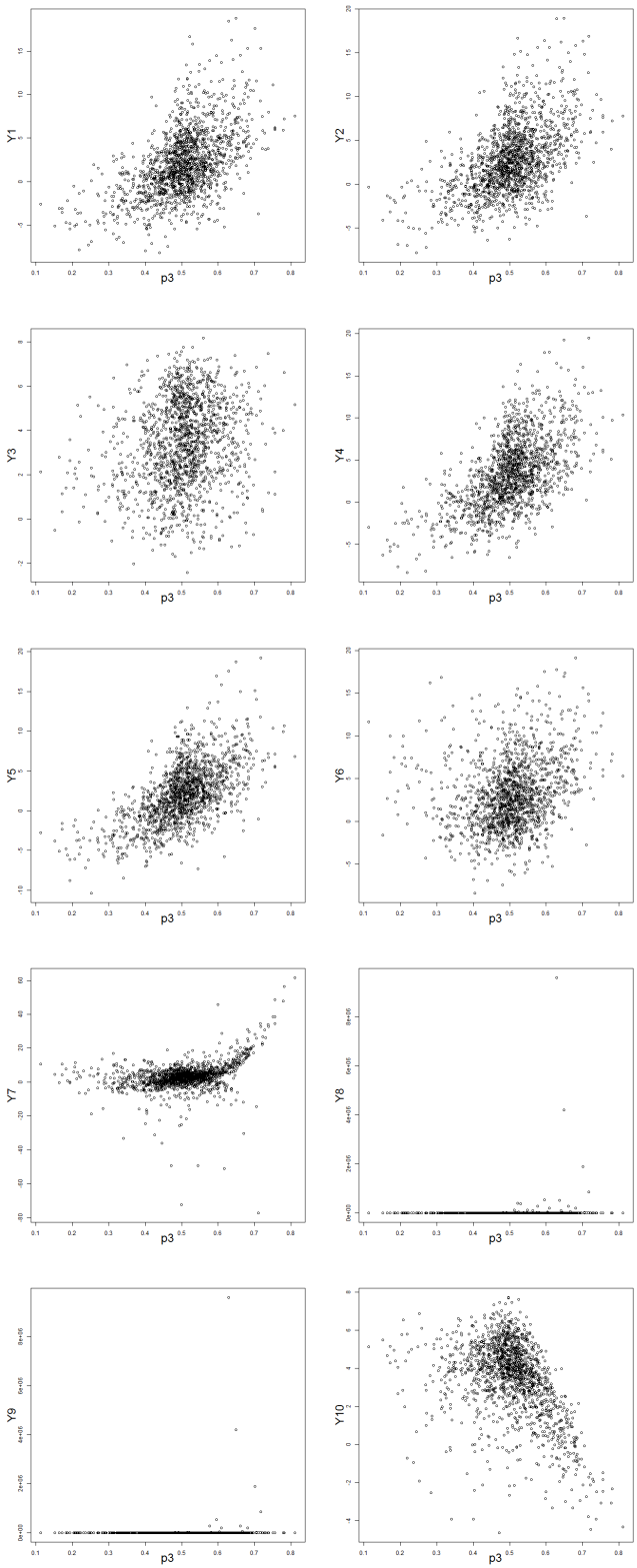


Figure 5.6.5: Scatter plots of $p4$ and variables of interest

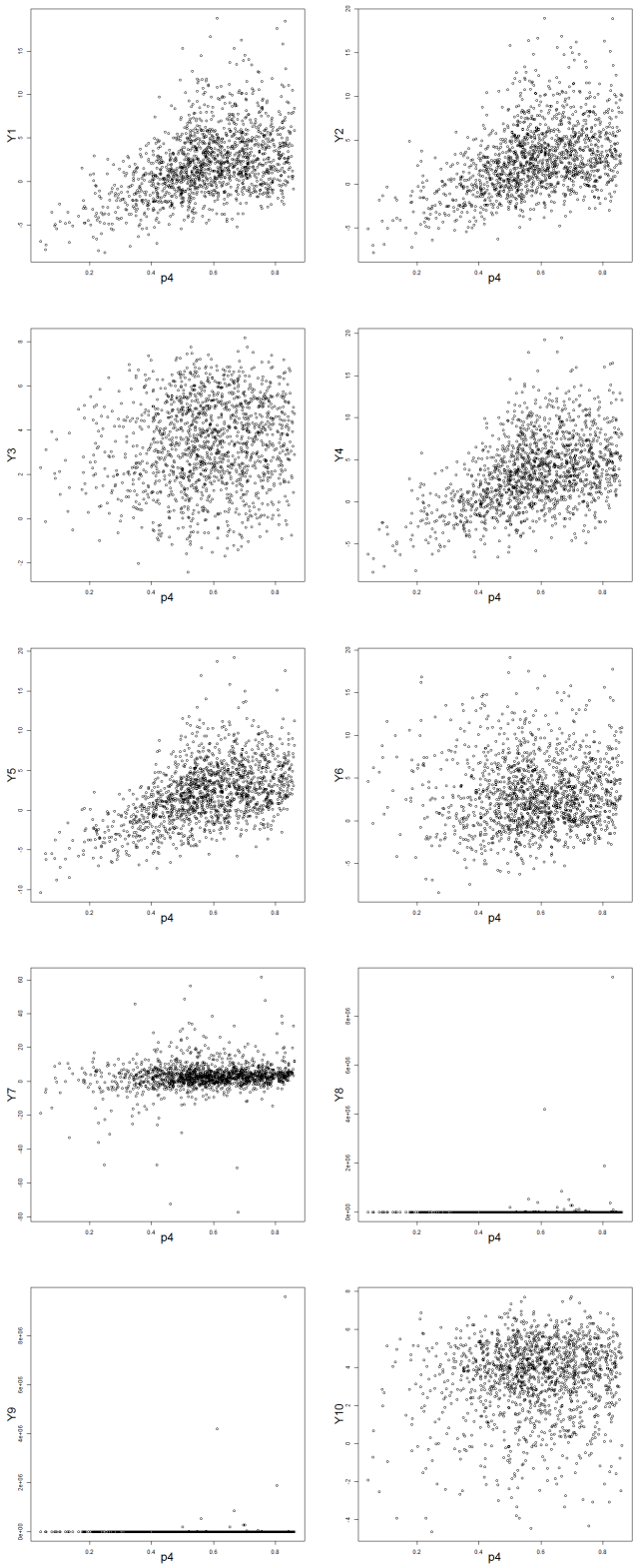


Figure 5.6.6: Scatter plots of $p5$ and variables of interest

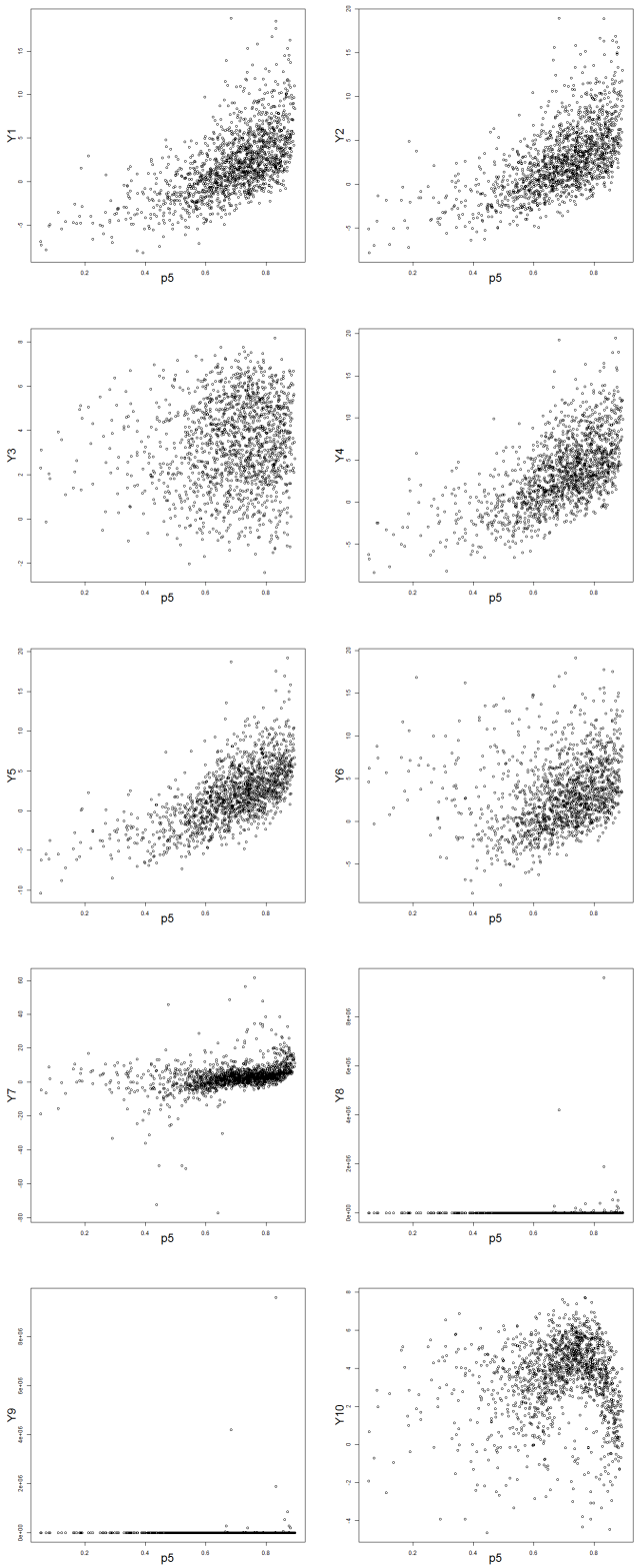
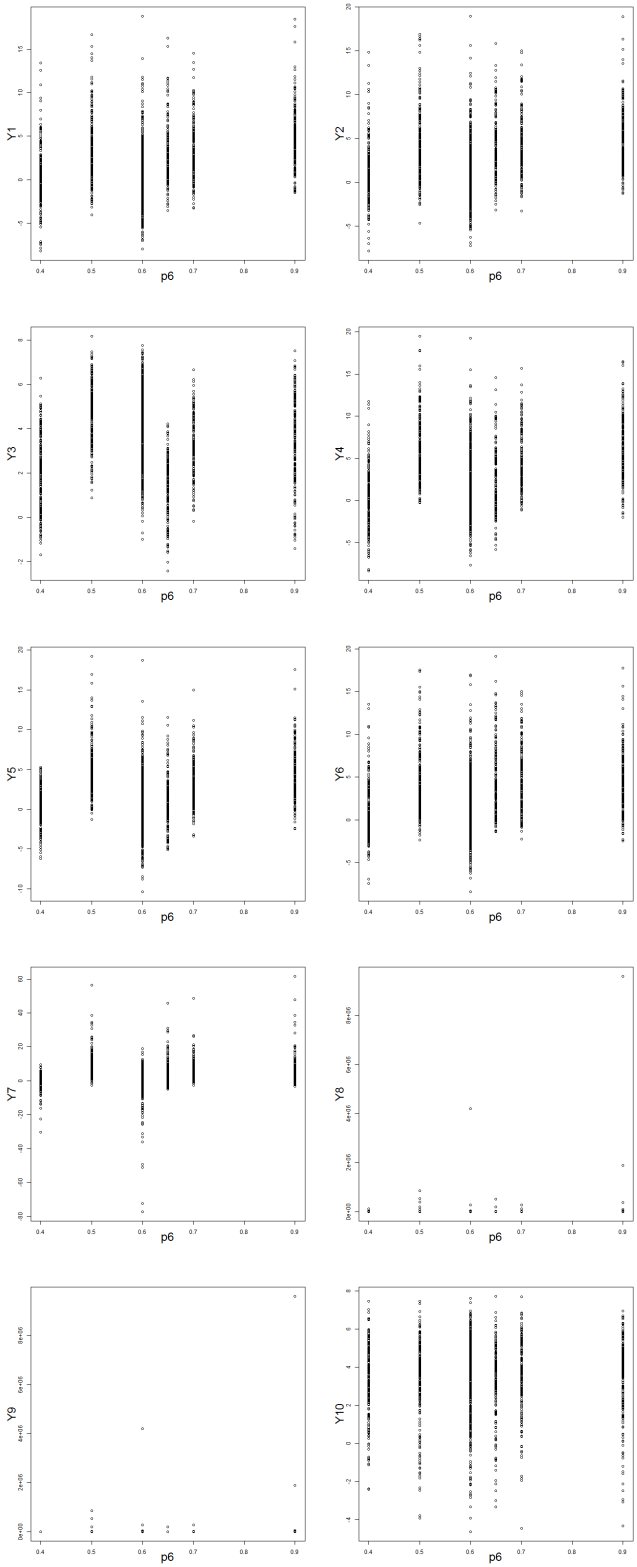


Figure 5.6.7: Scatter plots of $p6$ and variables of interest



5.6.3 Impact of estimated probabilities' truncation

In order to avoid too small values for $\hat{p}_i$, the common practice is to implement truncation with a lower bound t for $\hat{p}_i$'s. A usually implemented lower bound is $t = 0.02$. We want to check how much the choice of a different value in t could change the final performance in terms of MSE for $\hat{t}_y$ built on different machine learning methods. Let us denote $TMSE_{(e,m,t)}$ an MSE table computed for:
- e the estimator type of t_y's, $e \in \{\hat{t_{y}}_{Exp}, \hat{t_{y}}_{Haj}\}$,
- m the machine learning method used to estimate response probabilities,
- t the lower bound used for truncation.

Note that the model m can either be a machine learning used alone to estimate probabilities or a machine learning method associated to the Homogeneous Response Group creation (see section 5.2.1).

In our simulation study (see section 5.3), for each combination $e \times m \times t$, we have 70 indicators of MSE for $\hat{t}_y$ (10 variables of interest $\times$ 7 response mechanisms) - see for instance table 5.6.1. Thus we need a global indicator to sum up the overall modification of the 70 MSE's induced by a change in t. The Frobenius norm $\|TMSE_{(e,m,t)}\|_F = \sqrt{Tr(TMSE_{(e,m,t)}^{\top} TMSE_{(e,m,t)})}$ of the $TMSE_{(e,m,t)}$'s could provide this global measure of performance, and help evaluating the impact of a change in t. Indeed, the lower the MSE's are, the better the combination $e \times m \times t$ is. Thus, given e and m, the best value for t is the one that provides the lowest $\|TMSE_{(e,m,t)}\|_F$.

However, for an easier analysis of the results, we rather compute the following normalized indicator (still based on the computation of a Frobenius norm):

$$NF_{(e,m,t)} = \|TMSE_{(e,m,t)}/TMSE_{(e,m,t=0.02)}\|_F/8.3666$$

where $TMSE_{(e,m,t)}/TMSE_{(e,m,t=0.02)}$ is a term by term division of $TMSE_{(e,m,t)}$ by $TMSE_{(e,m,t=0.02)}$. The reference value for t is 0.02. The denominator 8.3666 is the Frobenius norm of a 10×7 matrix filled with $1's$: it is the NF value in case of TMSE's global stability when t=0.02 is replaced by an other value of t.

Table 5.6.1: $TMSE_{(\hat{t_{y_{Haj}}}, \text{ HRG after logistic regression, } 0.02)}$ for the 10 Variables of interest and the 7 response mechanisms

Variable	R0	R1	R2	R3	R4	R5	R6
Y1	1.4021e+03	2.1458e+04	3.8609e+04	2.2985e+04	3.0776e+04	1.1202e+04	8.9972e+03
Y2	1.4213e+03	2.1680e+04	3.9453e+04	2.3009e+04	3.0271e+04	1.1876e+04	8.9842e+03
Y3	6.5187e+02	4.1476e+03	8.1098e+03	6.6249e+03	5.7192e+03	3.2749e+03	2.2757e+03
Y4	1.3316e+03	2.6639e+04	5.3325e+04	3.5028e+04	4.2719e+04	1.6692e+04	9.7642e+03
Y5	1.0984e+03	1.6990e+04	4.0012e+04	2.5905e+04	3.1258e+04	1.1059e+04	8.0650e+03
Y6	2.7269e+03	2.0168e+04	4.0525e+04	1.8004e+04	5.1633e+04	1.2902e+04	2.1496e+04
Y7	3.3016e+04	9.1681e+04	1.4850e+05	1.0149e+05	1.5752e+05	1.7955e+05	5.9044e+04
Y8	1.9665e+17	3.0832e+20	3.9469e+20	4.3861e+20	2.0440e+20	3.4252e+19	1.7940e+20
Y9	1.6805e+17	2.8333e+20	3.6039e+20	4.2388e+20	2.0022e+20	3.2704e+19	1.7510e+20
Y10	1.5409e+03	5.7964e+03	7.1124e+03	6.5140e+04	9.5585e+03	5.2240e+03	4.1191e+03

For instance in table 5.6.2, we can examine in detail the 70 ratios of

$$TMSE_{(\hat{t_{y_{Haj}}}\text{, HRG after logistic regression, 0.06})}/TMSE_{(\hat{t_{y_{Haj}}}\text{, HRG after logistic regression, 0.02})}$$

In this example, a change in truncation bound from t=0.02 to 0.06 has very little impacts (only 3 cases in bold font where the ratios are slightly different from 1). The corresponding indicator $NF_{(\hat{t_{y_{Haj}}}\text{,HRG after logistic regression, 0.06})}$ is 1 (see table 5.6.4).

Table 5.6.2: Ratios of TMSE with truncation 0.06 / TMSE with truncation 0.02

$$TMSE_{(\hat{t_{y_{Haj}}}\text{, HRG after logistic regression, 0.06})}/TMSE_{(\hat{t_{y_{Haj}}}\text{, HRG after logistic regression, 0.02})}$$

Variable	R0	R1	R2	R3	R4	R5	R6
Y1	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Y2	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Y3	1.00	1.00	1.00	1.00	***0.99***	1.00	1.00
Y4	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Y5	1.00	***1.01***	1.00	1.00	1.00	1.00	1.00
Y6	1.00	1.00	1.00	1.00	***0.99***	1.00	1.00
Y7	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Y8	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Y9	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Y10	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Let us focus on two of the best methods in terms of MSE's (see section 5.3.3). NF indicators table 5.6.3 for $\hat{t_{y_{Exp}}}$ and table 5.6.4 for $\hat{t_{y_{Haj}}}$, show that HRG after logistic regression is robust in terms of MSE: we can see that NF indicators are always equal to 1, with m = HRG after logistic regression. HRG after Unpruned CART is quite robust but exhibits better global performance in terms of MSE with $t = 0.06$ both for $\hat{t_{y_{Exp}}}$ and for $\hat{t_{y_{Haj}}}$.

Table 5.6.3: NF indicator for $\hat{t_{y_{Exp}}}$ with different lower bounds truncation of $\hat{p}_i$

Method	**Lower bound 0.06**	**Lower bound 0.08**	**Lower bound 0.10**	**Lower bound 0.14**
Logistic regression	0.99	0.98	0.97	0.97
Logistic regression Bagging	1.00	1.00	1.00	1.00
Logistic Random Forest	1.00	1.00	1.00	1.00
Quadratic nonparametric discriminant analysis	1.00	1.00	1.00	1.00
Default pruned CART	1.00	1.00	1.00	1.00
Unpruned CART	1.00	1.01	1.05	1.05
CART Bagging	1.00	1.00	1.00	1.00
CART Random Forest	1.00	1.00	1.00	1.00
CART Boosting	1.02	1.11	1.27	1.27
CART Gradient Boosting	1.00	1.00	1.00	1.00
Ctree	1.00	1.00	1.00	1.00
Ctree Bagging	0.80	0.78	0.76	0.76
Ctree Random Forest	0.93	0.92	0.91	0.91
MultiVariate CTrees	1.00	1.00	1.00	1.00
Radial Kernel SVM	1.00	1.00	1.00	1.00
HRG after Logistic regression	***1.00***	***1.00***	***1.00***	***1.00***
HRG after Logistic regression Bagging	1.00	1.00	1.00	1.00
HRG after Logistic Random forest	1.00	1.00	1.00	1.00
HRG after Quadratic nonparametric HRG after discriminant analysis	1.04	1.09	1.16	1.16
HRG after Default pruned CART	1.00	1.00	1.00	1.00
HRG after Unpruned CART	***0.94***	***0.96***	***1.00***	***1.00***
HRG after CART Bagging	1.00	1.00	1.00	1.00
HRG after CART Random Forest	1.00	1.00	1.00	1.00
HRG after CART Boosting	1.02	1.11	1.27	1.27
HRG after CART Gradient Boosting	1.01	1.08	1.17	1.17
HRG after Ctree	1.00	1.00	1.00	1.00
HRG after Ctree Bagging	1.00	1.00	1.00	1.00
HRG after Ctree random Forest	1.00	1.00	1.00	1.00
HRG after MultiVariate CTrees	1.00	1.00	1.00	1.00
HRG after SVM	1.00	1.00	1.00	1.00

Table 5.6.4: NF indicator for $\hat{t}_{y_{Haj}}$ with different lower bounds truncation of $\hat{p}_i$

Method	**Lower bound 0.06**	**Lower bound 0.08**	**Lower bound 0.10**	**Lower bound 0.14**
Logistic regression	0.98	0.97	0.96	0.97
Logistic regression Bagging	1.00	1.00	1.00	1.00
Logistic Random Forest	1.00	1.00	1.00	1.00
Quadratic nonparametric discriminant analysis	1.00	1.00	1.00	1.00
Default pruned CART	1.00	1.00	1.00	1.00
Unpruned CART	1.00	1.01	1.03	1.01
CART Bagging	1.00	1.00	1.00	1.00
CART Random Forest	1.00	1.00	1.00	1.00
CART Boosting	1.00	1.00	1.00	1.00
CART Gradient Boosting	1.00	1.00	1.00	1.00
Ctree	1.00	1.00	1.00	1.00
Ctree Bagging	0.79	0.78	0.77	0.78
Ctree Random Forest	0.95	0.95	0.94	0.95
MultiVariate CTrees	1.00	1.00	1.01	1.00
Radial Kernel SVM	1.00	1.00	1.00	1.00
HRG after Logistic regression	***1.00***	***1.00***	***1.00***	***1.00***
HRG after Logistic regression Bagging	1.00	1.00	1.00	1.00
HRG after Logistic Random forest	1.00	1.00	1.00	1.00
HRG after Quadratic nonparametric HRG after discriminant analysis	0.85	0.83	0.84	0.83
HRG after Default pruned CART	1.00	1.00	1.00	1.00
HRG after Unpruned CART	***0.95***	***0.97***	***0.99***	***0.97***
HRG after CART Bagging	1.00	1.00	1.00	1.00
HRG after CART Random Forest	1.00	1.00	1.00	1.00
HRG after CART Boosting	1.39	5.03	11.21	5.03
HRG after CART Gradient Boosting	0.74	0.69	0.67	0.69
HRG after Ctree	1.00	1.00	1.00	1.00
HRG after Ctree Bagging	1.00	1.00	1.00	1.00
HRG after Ctree random Forest	1.00	1.00	1.00	1.00
HRG after MultiVariate CTrees	1.00	1.00	1.00	1.00
HRG after SVM	1.00	1.00	1.01	1.00

Chapter 6

Conclusion and prospect

In chapter 3, we dealt with item nonresponse through imputation. In this chapter, we proposed a modification of the Shao-Wang joint procedure, where initial imputed values obtained using this method, are modified so as to satisfy calibration constraints, which corresponds to MIVQUE estimators of model parameters. When the underlying distribution of the variables being imputed is symmetric or exhibits a low degree of asymmetry, our proposed procedure is significantly more efficient than the Shao-Wang procedure in terms of mean squared error. To go further, we could investigate the preservation of relationships of more than two items. Futhermore, in the presence of imputed values, variance estimators computed by treating the imputed as observed values are prone to underestimate the true variance of point estimators. That is why, considering the complexity of our proposed procedure, we could develop variance estimation with bootstrap techniques.

In chapter 4, we proposed imputation methods adapted to a study variable containing a large number of zeros. Motivated by a mixture regression model, we proposed two imputation procedures for such data and studied their properties in terms of bias and efficiency. We showed that these procedures preserve the distribution function if the imputation model is well specified. The results of a simulation study illustrate the good performance of the proposed methods in terms of bias and mean square error, as compared to alternative methods proposed

by Haziza et al. (2014). A motivating sequel for this research subject could be the construction of an imputation procedure preserving quantiles.

In chapter 5, we estimated response probabilities in the context of weighting for unit nonresponse. We conducted a comprehensive simulation study, aiming at a global ranking of different machine learning methods in totals estimation performance through response probabilities estimation. The best method in terms of mean squared error was the logistic regression associated with Homogeneous Response Groups creation, both for the expansion estimator and for the Hajek estimator. Unpruned CART associated with Homogeneous Response Groups creation appear among the methods with good performance and that could handle missing values among regressors, particularly with the expansion estimator. Those two methods turned out to be very robust against changes in lower bound truncation of estimated probabilities. To go further, we could also enlarge the set of machine learning to compare with stacking for instance.

The central topic of this PhD thesis was item and unit nonresponse handling in survey sampling theory. In all this work, either with Imputation Model approach or with Nonresponse Model approach, we had to rely on auxiliary information. Indeed, this one has been used in imputation procedures to deal with item nonresponse, aiming at the preservation of some finite population preservation: correlation coefficient among two variables of interest to be imputed (chapter 3) and the finite population distribution function in case of zero inflated variable of interest (chapter 4). Auxiliary information also intervenes in unit nonresponse handling through response probabilities estimation (chapter 5). However, in the previous mentioned chapters, we did not deal with nonresponse in auxiliary variables. This could be an interesting field for further researches - using MIVQUE (chapter 3), CART or Conditional inference trees (chapter 5).