UNIVERSITÉ DE NEUCHÂTEL

FACULTÉ DES SCIENCES ÉCONOMIQUES

IMPUTATION OF INCOME VARIABLES IN A SURVEY CONTEXT AND ESTIMATION OF VARIANCE FOR INDICATORS OF POVERTY AND SOCIAL EXCLUSION

PhD Thesis submitted to the Faculty of Economics and Business

Institute of Statistics

University of Neuchâtel

For the degree of PhD in Statistics

by

Eric GRAF

Accepted by the dissertation committee:

Prof. Yves TILLÉ, University of Neuchâtel, thesis director

Prof. Catalin STARICA, University of Neuchâtel, jury president

Dr. Camelia GOGA, HDR, University of Bourgogne, France

Dr. Jean-Pierre RENFER, Swiss Federal Office of Statistics

Prof. Louis-Paul RIVEST, University of Laval, Canada

Defended on 3rd of November 2014

Faculté des Sciences Economiques
Avenue du 1er-Mars 26
CH-2000 Neuchâtel
www.unine.ch/seco

LETTRES ET SCIENCES HUMAINES
SCIENCES
DROIT
SCIENCES ECONOMIQUES
THEOLOGIE

UNIVERSITÉ DE
NEUCHÂTEL

IMPRIMATUR POUR LA THÈSE

Imputation of income variables in a survey context and estimation of variance for indicators of poverty and social exclusion

Eric GRAF

UNIVERSITÉ DE NEUCHÂTEL

FACULTÉ DES SCIENCES ÉCONOMIQUES

La Faculté des sciences économiques,
sur le rapport des membres du jury

Prof. Yves Tillé (directeur de thèse, Université de Neuchâtel)
Prof. Catalin Starica (président du jury, Université de Neuchâtel)
Dr. Camelia Goga (Université de Bourgogne)
Dr. Jean-Pierre Renfer (Office Fédéral de la Statistique),
Prof. Louis-Paul Rivest (Université de Laval, Québec)

Autorise l'impression de la présente thèse.

Neuchâtel, le 20 novembre 2014

Le doyen

Jean-Marie Grether

■ Téléphone : +41 32 718 15 00 ■ Fax : +41 32 718 15 01 ■ E-mail : secretariat.seco@unine.ch ■ www.unine.ch

To my family.

Acknowledgements

This work is funded and was carried out under a productive collaboration agreement between the Swiss Federal Office of Statistics (SFSO) and the Institute of Statistics (ISTAT) at the University of Neuchâtel. Let both parties being warmly thanked here. Let Pr. Yves Tillé be thanked here for the opportunity, the privilege and also the pleasure of working with him. I always found great support over the past three years whenever I needed it. Let the Section of Methods (METH) and the Section of Income, Consumption and Living Conditions (EKL) of the SFSO also be thanked. Particularly I want to express my great gratitude to Dr. Philippe Eichenberger and Dr. Jean-Pierre Renfer from METH because they agreed and supported me in quitting my methodologist position at the SFSO in favour of transfer to ISTAT in order to produce this thesis. I have extensively benefited from my working experience of nearly 9 years at the SFSO and also at the Swiss Household Panel (SHP) for the time it was still based at the University of Neuchâtel. The problems I encountered during these years of service and how they were solved really gave birth to my inspiration to pursue this thesis. I sincerely thank Pr. Catalin Starica who accepted to be the president of this jury, Pr. Louis-Paul RIVEST, Dr. Camelia Goga and Dr. Jean-Pierre Renfer who accepted to review this document. It is a great honour for me. Thanks also go to my colleagues and former colleagues, at the University of Neuchâtel, the SFSO and the SHP. I always had good contacts with them and I learnt a lot with them. Finally, I wish to thank my whole family and my friends who gave me support and encouragement.

Neuchâtel, 28 August 2014.

Contents

List of Figures

List of Tables

Abstracts

Imputation of income variables in a survey context and estimation of variance for indicators of poverty and social exclusion

Keywords : imputation, linearization, weighting, variance, generalized calibration, GB2.

Abstract

This Phd thesis proposes to develop a method of imputation for income variables allowing direct analysis of the distribution of such data, particularly the estimation of complex statistics such as indicators for poverty and social exclusion as well as the estimation of their precision.

In an introduction chapter we present the Swiss Survey on Income and Living Conditions (SILC) which we extensively used to illustrate our research.

In a first article accepted for publication, co-written with Dr. Lionel Qualité, we present an overview of the production methods at the Swiss Federal Office of Statistics (SFSO). Samples are selected with coordination so as to spread the survey burden over the population. We present the computation of extrapolation weights adapted to different cases and needs with its main steps. The SFSO relies on international recommendations for data editing and imputation, and contributes to their elaboration. The precision of estimators is consistently evaluated, according to the different treatments and methods involved in their construction.

In a second published article, co-written with Pr. Yves Tillé, we have used the generalized linearization technique based on the concept of influence function, as Osier (2009) has done, to estimate the variance of complex statistics such as Laeken indicators. Through simulations, we show that the use of Gaussian kernel estimation to estimate an income density function results in a strongly biased variance estimate. We propose two other density estimation methods that significantly reduce the observed bias.

In a working paper, we resume the idea presented by Deville & Särndal (1994) which consists in constructing an unbiased estimator of the variance of a total based solely on the information at our disposal (i.e. on the selected sample and the subset of respondents) in the case of regression imputation. While these authors dealt with a conventional total of a variable of interest, we reproduce a similar development in the case where the considered total is one of the linearized variable of quantiles. We show by means of simulations on real survey data that regression imputation can have an important impact on the bias and variance estimations of social inequality indicators. This leads us to a method capable of taking into account the variance due to imputation in addition to the one due to the sampling design in the cases of quantiles.

In a submitted article, we present our new imputation method for income variables. Empirical studies have shown that the generalized beta distribution of the second kind (GB2) fits income data very well. We present a parametric method of imputation relying on weights stemming from generalized calibration. A GB2 distribution is fitted on the income distribution in order to determine whether these weights can compensate even for nonignorable nonresponse that affects the variable of interest. The success of the operation greatly depends on the choice of auxiliary and instrumental variables used for calibration, which we discuss. We validate our imputation system on SILC data and compare it to imputations performed through the use of IVEware software. We have made great efforts to estimate variances through linearization, taking all the steps of our procedure into account.

The last part of this Phd thesis discusses additional material which we could not include in the other chapters. Namely we give some more insights into the GB2 distribution, study the possibility of using Durbin-Wu-Hausman tests in the framework of generalized calibration and present a way of forming imputation classes for an income variable.

Imputation de variables de revenu provenant d'enquête et estimation de variance pour des indices de pauvreté et d'exclusion sociale

Mots-clés : imputation, linéarisation, pondération, variance, calage généralisé, GB2.

Résumé

Cette thèse développe une méthode d'imputation pour des données de revenus permettant des analyses directes sur la distribution de ces variables et également l'estimation de statistiques complexes telles que des indices de pauvreté et d'exclusions sociale ainsi que l'estimation de leur précision.

Dans un chapitre introductif, nous présentons l'enquête sur les revenus et conditions de vie (SILC) dont les données sont utilisées à plusieurs reprises pour illustrer nos recherches.

Dans un premier article accepté pour publication, co-écrit avec Dr. Lionel Qualité, nous présentons un aperçu des méthodes actuellement utilisées à l'Office Fédéral de la Statistique (OFS). Les échantillons sont sélectionnés de manière coordonnée afin de répartir au mieux la charge d'enquête sur les ménages et les personnes. Le calcul des pondérations, dont on présente les principales étapes, est adapté aux différents besoins et aux différentes situations rencontrées. L'Office se base sur les recommandations internationales, dont il participe à l'élaboration, pour le traitement des données d'enquête et les imputations. La précision des estimateurs est systématiquement évaluée en tenant compte des traitements réalisés.

Dans un deuxième article publié, co-écrit avec le Pr. Yves Tillé, nous avons mis en oeuvre la technique de linéarisation généralisée reposant sur le concept de fonction d'influence, tout comme l'a fait Osier (2009),

pour estimer la variance de statistiques complexes telles que les indices de Laeken. Des simulations montrent que, pour les cas où l'on a recours à une estimation par noyau gaussien de la fonction de densité des revenus considérés, on obtient un fort biais pour la valeur estimée de la variance. On propose deux autres méthodes pour estimer la densité qui diminuent fortement le biais constaté.

Dans un rapport de recherche, nous résumons l'idée proposée par Deville & Särndal (1994) consistant à construire un estimateur non biaisé de la variance d'un total basé uniquement sur l'information à disposition (c'est-à-dire l'échantillon sélectionné et le sous-ensemble des répondants) dans le cas d'une imputation par régression. Alors que ces auteurs ont traité le total conventionnel d'une variable d'intérêt, nous reproduisons un développement similaire dans le cas où le total considéré est celui de la variable linéarisée d'un quantile. Nous montrons à l'aide de simulations sur des données d'enquête réelles que l'imputation par régression peut avoir un impact important sur le biais de la variance estimée pour des indicateurs d'inégalité sociale. Cela nous mène à une méthode capable de prendre en compte la variance due à l'imputation, en plus de celle du plan dans le cas de quantiles.

Dans un article soumis, nous présentons notre nouvelle méthode d'imputation pour des variables de revenus. Des études empiriques ont montré que la loi bêta généralisée de seconde espèce (GB2) s'ajuste très bien à des données monétaires. Nous présentons une méthode d'imputation paramétrique reposant sur l'utilisation de poids issus d'un calage généralisé. Une loi GB2 est ajustée sur la distribution des revenus pour valider ces poids capables de compenser même pour de la non-réponse non-ignorable. Le succès de l'opération dépend grandement du choix, qui est discuté, des variables auxiliaires et instrumentales utilisées pour le calage. Nous validons notre système d'imputation sur les données SILC et comparons les résultats avec ceux obtenus par des imputations réalisées avec le logiciel IVEware. Nous avons investi de gros efforts pour estimer les variances par linéarisation, en prenant toutes les étapes de la procédure en compte.

La dernière partie de la thèse discute du matériel additionnel qui n'a pas pu être inclus dans les autres chapitres. Nous donnons notamment quelques détails supplémentaires sur la distribution GB2, étudions la possibilité d'utiliser des tests de Durbin-Wu-Hausman dans le cadre du calage généralisé et présentons une façon de former des classes d'imputation pour une variable de revenu.

Introduction

Undoubtedly, knowledge about income distribution of the population is of vital interest to all studies of economic markets in order to govern economic and social decision-making. In economics and social statistics, the study of income distribution is at the heart of inequality measures and more generally of evaluations of social welfare.

In official national sample surveys of households and individuals, nonresponse in income variables is often a difficult and important problem in the sense that the phenomenon affects the image of the reality that the survey reflects. Without proper treatment, the outcomes of political or social decision-making based on the results of the survey may be wrong.

The initial motivation for this thesis was to attempt to improve the methodology of imputations used at the SFSO for income variables in two of its main surveys, namely the Survey on Income and Living Conditions (SILC) and the Swiss Labour Force Survey (SLFS). Indeed the author of this dissertation played an active role for almost three years developing and adapting acceptable imputations for these surveys. Most of the well-known methods were tested and the main concern was systematically how to adapt the theory to practice in a defensible, reasonable, documentable and reproducible manner.

Typically, the survey methodologist is given a survey-data file containing missing values. His mission is to compute survey-weights, imputations or set up a variance estimation methodology for some given statistics. A conscientious person cannot be content with accomplishing such a task without obtaining information about what has been done at all stages of the Data Preparation Process (DPP, see Section 2.5.1) for this particular study. For example, we had noticed that the number of phone calls needed during the fieldwork to establish a first contact with a household in Computer Assisted Telephone Interview (CATI) surveys was a very important piece of information when modelling nonresponse at several levels of the data collection as well as when imputing missing data. Another example is that the treatment of *outliers* (in the sense of Chambers, 1986; Hulliger, 1999) or extreme values in the initial phase of the DPP (see Figure 2.8, Section 2.5.1) can have a major impact on computed weights or imputations afterwards. To illustrate this point, sometimes policy decisions have to be met: if the richest person of Switzerland is randomly selected for part of the sample, that person's fortune (over 35 billion of Swiss Francs[1]) is thousands of times the median fortune of the Swiss individual or household (around 350,000 SFR, see Müller, 2014). In the fortune distribution, the

[1]See BILAN journal, http://www.bilan.ch/300plusriches.

person will be a *representative outlier* as defined by Chambers (1986); Hulliger (1999). Such an observation can have a huge and undesired effect on a regression model for example, although it is a true and correctly measured unit. In such cases, the usual treatment is to put aside the influence of that observation or to reduce it, nevertheless without erasing it from the data file.

Red path and motivation

My thesis has been driven by the desire to explore the potential of a new imputation procedure for income variables which would perform better than IVEware (Raghunathan et al., 2001) software that I extensively used for imputations tasks in my work at the SFSO. In this sense, Chapter 5 represents the heart of my work.

Chapter 1 is a short introduction to the Swiss SILC 2009 survey data survey since the following chapters often use an extract of the data originating from this survey to illustrate our different research. It allows the reader to gain more insights about the nature of the dataset. We also explain how the data stemming from the CATI survey could be matched with a register providing exceptional possibilities for methodological research.

Chapter 2, which is co-written with Dr. Lionel Qualité, reproduces an article which was reviewed and accepted for publication in a special issue of the *Journal de la Société Française de Statistiques*. We present an overview of the production methods at the SFSO. Samples are selected with coordination so as to spread the survey burden over the population. We present the computation of extrapolation weights adapted to different cases and needs with its main steps. The SFSO relies on international recommendations for data editing and imputation, and contributes to their elaboration. The precision of estimators is consistently evaluated, according to the different treatments and methods involved in their construction. This chapter can be considered as an extension of the introduction since it describes the framework in which this thesis is rooted. In the following chapters the reader is considered to be familiar with the notations and issues presented in Chapter 2.

Chapter 3, co-written with Pr. Yves Tillé, reproduces an article which was published in the Canadian journal *Survey Methodology*. In order to evaluate the results of my works, I was led to estimate the variance of estimators relying on imputed data. Having participated in the European project AMELI (Advanced Methodology for European Laeken Indicators, see European Union, 2011a), and having implemented the whole weighting and imputation procedures for SILC at the SFSO, I examined non-linear statistics as the inequality indicators (formerly called Laeken indicators, see Eurostat, 2003, 2005a). We have used the generalized linearization technique based on the concept of influence function, as Osier (2009) has done, to estimate the variance of complex statistics such as Laeken indicators. In this area we have brought some substantial improvement in estimating the variance through the generalized linearization method

(Deville, 1999b; Demnati & Rao, 2004b; Osier, 2009). Through simulations, we show that the use of Gaussian kernel estimation to estimate an income density function results in a strongly biased variance estimate. We propose two other density estimation methods that significantly reduce the observed bias.

Chapter 4 studies the possibility of taking the imputation into account in the linearization procedure. Indeed, due to lack of time or of expertise, in practice, at the SFSO, the fact that a part of the data has been imputed is often neglected when it comes to estimating various statistics of interest. The files are delivered to the public with imputation flags allowing the user to identify which observations are real and which ones have been produced by imputation. Most of the time imputation is avoided and missing values are compensated by re-weighting. Or, if the missing rate is low (say $< 5\%$), imputation is neglected and all the values are considered as true measured values. Nevertheless, in the case of the SILC survey, EUROSTAT asks for files without missing values, which implies that some data have to be imputed. We resume the idea presented by Deville & Särndal (1994) which consists in constructing an unbiased estimator of the variance of a total based solely on the information at our disposal (on the selected sample and the subset of respondents) in the case of regression imputation. While these authors dealt with a conventional total of an interest variable, we reproduce a similar development in the case where the considered total is one of the linearized variable of quantiles. We show by means of simulations on real survey data that regression imputation can have a major impact on the bias and variance estimations of inequality indicators. This leads us to a method capable of taking into account the variance due to imputation in addition to the one due to the sampling design in the cases of quantiles. Some more work should be done in order to achieve variance estimating expressions for the other Laeken indicators. Of course, another means of estimating the variance due to imputations is to produce multiple imputations (see Section 2.6.2).

Chapter 5 reproduces a submitted article presenting a new imputation method for income variables. Empirical studies have shown that the generalized beta distribution of the second kind (GB2) fits income data very well. We use a parametric method of imputation relying on weights obtained by generalized calibration. A GB2 distribution is fitted on the income distribution in order to determine whether these weights can compensate even for nonignorable nonresponse that affects the variable of interest. The success of the operation greatly depends on the choice of auxiliary and instrumental variables used for calibration, which we discuss. We validate our imputation system on data from the Swiss Survey on Income and Living Conditions (SILC) and compare it to imputations performed through the use of IVEware software running on SAS®. We have made great efforts to estimate variances through linearization, taking all the steps of our procedure into account.

Finally, Chapter 6 is last part of this Phd thesis and discusses additional considerations which we could not include in the different articles because of length restrictions. Namely we give some more insights on the GB2

distribution, study the possibility of using Durbin-Wu-Hausman tests in the framework of generalized calibration.

A few appendices complete the document providing more details on some parts of the text.

The SILC survey in Switzerland

1

This survey aims to study poverty, social exclusion and living conditions using comparable indicators at European level. The main indicators are those described in the Chapters 3 and 6. Multidimensional micro data, updated and comparable, on income, housing, work, education and health are collected annually. Two types of data are produced: cross-sectional data (covering a period or a specific time, typically the current year) and longitudinal data (relating to changes at the individual level, observed over a period up to four years). In Switzerland, as in most of the other participating countries, it is designed as a rotating panel over four years (as Figure 2.6 in Section 2.4.4 is illustrating it). More precisely, our analyses were conducted on the SILC 2009 data. The year 2009 was of special methodological interest because the survey data could be matched for the first time with a register. This paved the way for many studies and comparisons to assess the quality of collected CATI-data. The SFSO delivered two files: one file of household-level data containing 7372 observations; and a file containing data on all individuals living in those households counting 17,561 rows.

1.1 Sampling design and weighting scheme

This section aims to give minimal but sufficient information about the survey weights produced and delivered by the SFSO each year with the SILC data. We felt it necessary to give the reader some basic information about these weights, since they are used in almost all our investigations with the SILC 2009 data.

In Switzerland, the SILC survey was launched in 2007 for its first wave. It was preceded by three years of testing, in particular, in 2004 and 2005 a pilot study was conducted in close collaboration with the Swiss Household Panel (SHP). The sampling design of each wave of the SILC survey as well as the methodology of the weighting system have both been strongly inspired by the weighting scheme in place at the SHP. And the weighting scheme of the SHP was developed in the years 1999 and 2000 in collaboration between the SHP, the SFSO and Statistics Canada. In Graf (2006, 2008a) we detailed the sampling design and weighting methodology for the swiss SILC. Very briefly the survey functions as follow. Every year,

a stratified simple random sample of households (in fact telephone numbers) is drawn out of the sampling frame of the SFSO. That sample is then surveyed for four consecutive years and then dropped out of the survey. Complications arise from the fact that the survey has many stages: first a "grid" questionnaire collects very general information on the household and its members and then every individual has to answer an individual questionnaire. Moreover, one of these individuals called the "reference person" has to answer an additional household questionnaire. At every stage and for the four consecutive years, nonresponse may appear with more or less damaging effects. In the end, to simplify, three conceptually different sets of weights are produced annually:

- cross-sectional individual weights,
- cross-sectional household weights,
- longitudinal individual weights, themselves of three different kinds:
  - longitudinal individual weights for analyses of change over a two-years period,
  - longitudinal individual weights for analyses of change over a three-years period,
  - longitudinal individual weights for analyses of change over a four-years period.

The cross-sectional weights allow us to make inference on the Swiss population of the currently surveyed year. That means, for the SILC 2009 data, cross-sectional weights allow us to make inference on the Swiss resident population of 2009. These are the survey weights we used as initial weights in the surveys presented in the following chapters.

The main steps applied to compute the weights are:

- sampling or initial weights, corresponding to the inverse of the sample inclusion probabilities,
- nonresponse correction factors are computed at various levels and each time a refusal could happen (to the grid questionnaire, to the individual questionnaire and to the household questionnaire, as well as from one year to the next for the panellised units),
- from the second year on, weight sharing is used to attribute weights to new members of surveyed households,
- panel combination: combine the different samples which are one to four years old respectively,
- calibration on known population totals.

These steps are presented succinctly in Section 2.4 and described in detail in Graf (2008a) and, in an important article with regards to our work, Massiani (2013b) detailed the procedure of how to take them into account while estimating the variance of a total through linearization.

1.2 Variable of interest for development of an imputation strategy

We aimed to develop an imputation procedure for income variables, in particular the variable containing the individual employed or salaried income. In the original data, it is the variable $y_{CATI} = P09I57G$ which was, as its name suggests, collected by CATI. This variable is subject to high non-response, Table 1.1 displays some details.

The SILC data could be matched with the register of the Central Compensation Office (CCO)[1] allowing, among other things, to have access to the declared salary of the same individuals. These data are of good quality for employees, which is why we have limited our study to these cases. Indeed, there are also similar data available for self-employed persons; however, experience has shown that the registered values are often below the actual income. On the other hand, data on self-employed take longer to be completed, which means that they are often not yet final when the SFSO wants to make the survey data available. Thus the salaried income of the CCO register is of good quality, still there subsists a small fraction (3%) of missing values due to unmatched records (see Table 1.1).

Table 1.1 – *Detailed nonresponse affecting employee personal income, CATI-variable. (NR = nonresponse, I.Q. = individual questionnaire, F_cati = nonresponse flag for variable y_{CATI}). Nonresponse affecting the corresponding register variable (due to the failure of matching) is also detailed (F_cco = nonresponse flag for variable y_{cco}).*

		F_cati		F_cco	
n	%	Reason for NR	no match	income= 0	income> 0
2967	16.9	not eligible to I.Q.	74	2893	0
321	1.8	proxy	29	238	54
1371	7.8	Total NR to I.Q.	110	475	786
97	0.6	NR to filter variable	2	12	83
4544	25.9	Filtered question	86	3938	520
317	1.8	NR	12	40	265
206	1.2	don't know	5	31	170
7738	44.1	> 0	195	659	6884
17561	100	Total	513	8286	8762

We decided to restrict the study to the cases highlighted by the shaded area of Table 1.1. These correspond to observations with a strictly positive value to the CCO-variable and where, simultaneously, the CATI-variable showed item nonresponse (the categories corresponding to the first three rows of the Table have been excluded since they correspond to total nonresponse, which is treated by reweighting). Individuals of interest are those whose CCO-income is reported in Table 1.2.

Our framework is the following: we aim to develop an imputation method for the variable $y_{CATI} = P09I57G\_cati$ which is subject to an un-

[1]The CCO is based in Geneva, it is a public institution active in the field of social insurance 1st (AVS / AI / APG), http://www.zas.admin.ch/cdc/.

known nonresponse mechanism. Luckily, we also have the complete variable $y_{CCO} = P09I57G\_cdc$. We apply the nonresponse observed for y_{CATI} to y_{CCO} and develop the imputation method for y_{CCO}. As we know the true values of y_{CCO}, we can evaluate the performance of the method.

By observing Table 1.2, we immediately notice that the nature of nonresponse to y_{CATI} is dependent on the level of income of the person. It is clearly a case of nonignorable nonresponse (NMAR) (*Non Missing At Random*, in the sense of Little & Rubin, 2002), see Chapter 5 for more details).

Table 1.2 – *Descriptives statistics of employee's income, CCO-variable, depending on the reasons for nonresponse to CATI.*

n	%	F_cati	Mean	Median	SD	Min	Max
83	1.0	NR to filter variable	12'165	8'126	13'462	278	69'502
520	6.6	Filtered question	23'580	9'214	47'760	16	511'775
265	3.3	NR	79'050	69'200	74'406	322	559'013
170	2.1	don't know	59'729	31'239	116'090	320	1'132'000
6884	86.9	> 0	70'250	61'936	63'718	100	1'758'845
7922	100.0						

Note that variables y_{CATI} and y_{CCO} do not behave in quite the same way. Indeed, the information collected by the telephone survey is not systematically very close to the one in the register. However, the two variables correlate to 84%. Figure 1.1 gives a graphical view of the relationship between the two variables.

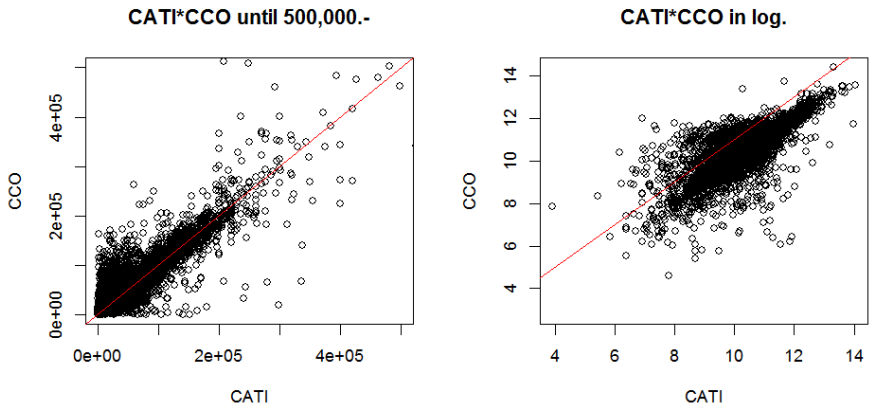


Figure 1.1 – *Employee's income collected by CATI versus CCO-register value.*

As will be recalled in Chapters 3 and 5, we further restricted the number of observations regarding the imputation model on units for which two other continuous variables were available, namely the housing costs and the occupation rate for the surveyed individuals. The number of full cases amounts to 6,188 out of 7,922 individuals, which means 21.89% nonresponse.

2

Sondage dans des registres de population et de ménages en Suisse : coordination d'enquêtes, pondération et imputation

Abstract

L'Office Fédéral de la Statistique harmonise ses enquêtes par échantillonnage auprès des personnes et des ménages en Suisse. Dans cet article, nous présentons un aperçu des méthodes actuellement utilisées. Les échantillons sont sélectionnés de manière coordonnée afin de répartir au mieux la charge d'enquête sur les ménages et les personnes. Le calcul des pondérations, dont on présente les principales étapes, est adapté aux différents besoins et aux différentes situations rencontrées. L'Office se base sur les recommandations internationales, dont il participe à l'élaboration, pour le traitement des données d'enquête et les imputations. La précision des estimateurs est systématiquement évaluée en tenant compte des traitements réalisés. [1]

The Swiss Federal Statistical Office is currently harmonizing its population and household survey operations. In this paper, we give an overview of its production methods. Samples are selected with coordination so as to spread the survey burden over the population. The computation of extrapolation weights adapted to different cases and needs is presented with its main steps. The Office relies on international recommendations for data editing and imputation, and contributes to their elaboration. The precision of estimators is consistently evaluated, according to the different treatments and methods involved in their construction.

Keywords: calage, variance, segmentation, valeurs aberrantes, census, calibration, variance, segmentation, outliers

[1] This chapter is a an authorized reprint of: Graf, E. & Qualité, L. (2014). Sondage dans des registres de population et de ménages en Suisse : coordination d'enquêtes, pondération et imputation. *Journal de la Société Française de Statistique* **155** No.4, 95–133.

2.1 Introduction

L'Office Fédéral de la Statistique (OFS) achève d'harmoniser ses enquêtes par échantillonnage auprès des personnes et des ménages en Suisse. Dans cet article, nous présentons un aperçu des méthodes actuellement utilisées par l'office. Il ne s'agit en aucun cas d'un manuel de bonnes pratiques, mais de l'exposé de solutions et compromis qui ont été réalisés pour les besoins spécifiques de l'OFS. L'uniformisation des méthodes d'échantillonnage, de pondération, et de traitement des données d'enquête est l'occasion de réévaluer toutes ces pratiques pour en dégager un corpus de méthodes solide et applicable aux divers cas rencontrés par l'office.

Dans une première partie, nous décrivons les enquêtes auprès de la population et auprès des ménages réalisées par l'OFS. On y présente les pendants suisses de grandes enquêtes européennes ainsi que le nouveau système de recensement, en partie basé sur des enquêtes annuelles par échantillonnage, qui remplace le recensement décennal de la population.

Dans une deuxième partie, nous présentons la méthode de tirage avec coordination utilisée pour les enquêtes du nouveau système de recensement et qui sera utilisée à terme pour toutes les enquêtes de l'OFS. Le tirage coordonné d'échantillons permet de mieux répartir la charge d'enquête sur la population.

Dans la troisième partie, nous présentons les méthodes d'estimation et de pondération utilisées pour les enquêtes de l'OFS. Les estimateurs utilisés sont en principe des estimateurs par dilatation, c'est-à-dire des sommes pondérées de valeurs observées (ou imputées). Nous présentons les étapes-clés de la construction des pondérations : calcul des probabilités d'inclusion dans l'échantillon, estimation des probabilités de réponse, partage des poids, combinaison d'échantillons, calages.

Dans la quatrième partie, nous présentons le principe de traitement des données et d'imputation appliqué à l'OFS. L'Office suit les recommandations issues de plusieurs grands projets européens quant au traitement des valeurs manquantes, aberrantes, erronées ou extrêmes. Nous mentionnons les principales techniques d'imputation utilisées en production.

Dans la cinquième partie, nous traitons les méthodes d'estimation de la variance utilisées à l'OFS. Celles-ci sont pensées pour tenir le mieux possible compte des principaux facteurs influençant la variance des estimateurs produits : le plan de sondage, la non-réponse et son traitement, les partages ou combinaisons de poids, enfin les calages et imputations. Dans la pratique, un compromis doit être réalisé car la prise en compte de tous ces aspects rend le problème extrêmement complexe. Des approximations simples sont donc utilisées.

Dans la sixième partie, nous donnons un aperçu sur les évolutions planifiées dans un futur proche : le traitement de la non-réponse par calage généralisé, l'estimation de la variance due à l'imputation ainsi que les dernières améliorations prévues quant à la coordination des échantillons.

2.2 Les différentes enquêtes auprès de la population des personnes et des ménages résidant en Suisse

2.2.1 Le nouveau système de recensement de la population suisse

Le traditionnel recensement décennal de la population suisse est remplacé, depuis 2010, par la création et l'exploitation par l'OFS d'un cadre d'échantillonnage[2] de population et de ménages appelé SRPH (Stichprobenrahmen für Personen und Haushaltserhebungen), et un système d'enquêtes annuelles. L'office ne gère pas de registre centralisé de la population, mais utilise différents registres pour créer le fichier SRPH. Celui-ci est essentiellement construit en agrégeant les registres administratifs du contrôle des habitants des communes et des cantons. L'inscription dans le registre de la commune de résidence est en effet obligatoire en Suisse, à quelques exceptions près, parmi lesquelles on peut citer les personnes ayant le statut de fonctionnaire international. Le fichier ainsi obtenu est encore enrichi par l'utilisation de registres fédéraux : le registre des demandeurs d'asile, le registre central des étrangers, le registre des fonctionnaires internationaux et le registre informatisé de l'état civil Infostar. La livraison des données par les communes et cantons, et la création du cadre d'échantillonnage, sont effectuées quatre fois par an, en dates de valeur des 31 décembre, 31 mars, 30 juin et 30 septembre. Les données sont disponibles environ six semaines après ces dates de référence.

Le fichier SRPH contient des informations démographiques basiques sur les personnes habitant en Suisse : sexe, date de naissance, état civil, permis de séjour, nationalité, date d'arrivée en Suisse et au lieu de résidence actuel, ainsi que l'adresse, le statut de résidence (résidence principale dans un ménage privé, résidence secondaire, ménage collectif, etc.). L'OFS a la conviction que ce répertoire offre une très bonne couverture de la population résidant de manière permanente en Suisse, l'inscription au registre de la commune étant nécessaire pour bénéficier d'un certain nombre de prestations, et déterminante pour le recouvrement de l'impôt sur le revenu. La pertinence de ce cadre de sondage pour les opérations statistiques est garantie par l'utilisation permanente du numéro unique de sécurité sociale comme identifiant à toutes les étapes de sa création, ainsi que par des procédures de contrôle de la qualité, depuis la livraison des données par les communes jusqu'à la constitution finale du fichier SRPH.

Depuis la fin décembre 2012, les communes ont l'obligation légale de fournir le numéro de bâtiment et de logement de chaque personne. Ces numéros, qui permettent de connaître la constitution des ménages, sont extraits du registre des bâtiments et logements entretenu à l'OFS. Cette obligation était déjà largement respectée depuis 2010, mais avec une certaine hétérogénéité selon les cantons et communes.

Le premier élément du nouveau système de recensement est le résultat de l'exploitation, chaque année, du répertoire de population du 31 décembre précédent. Les données utilisées pour constituer le fichier SRPH du

[2]La locution "cadre d'échantillonnage" utilisée en Suisse est équivalente à "base de sondage". Il s'agit de la traduction du mot allemand "Stichprobenrahmen".

31 décembre font l'objet de traitements supplémentaires pour produire la statistique de la population et des ménages (Statpop). Cette exploitation permet de fournir une première série de résultats, et également la population de référence pour les enquêtes complémentaires du système de recensement. Les données démographiques individuelles sont utilisées. Le taux de couverture de cette statistique sera évaluée sur la base d'une enquête réalisée en 2013. La constitution des ménages n'a pas été exploitée pour ces premiers résultats de comptages, tant que l'obligation légale de renseigner n'était pas en vigueur. Elle est par contre utilisée dans les processus de production de l'OFS au moment de l'échantillonnage.

Le deuxième élément du nouveau système de recensement est une enquête annuelle, par échantillonnage aléatoire, de grande ampleur : l'enquête structurelle. L'échantillon de base, financé par la confédération helvétique, est dimensionné de manière à obtenir approximativement 200'000 questionnaires renseignés par des personnes ayant leur résidence permanente en Suisse, et âgées de 15 ans ou plus. Cela représente un taux de sondage de 3% environ. Les cantons et communes peuvent financer une augmentation de l'échantillon les concernant, jusqu'à le doubler. Ponctuellement, en 2010, ils avaient la possibilité de quadrupler l'échantillon. Le questionnaire de cette enquête reprend en grande partie les questionnaires des recensements précédents en omettant les informations disponibles dans le cadre d'échantillonnage SRPH. Les thèmes abordés sont, entre autres, le niveau d'éducation, le statut d'activité et la branche d'activité, les trajets domicile-travail, la langue employée, la composition du ménage et les relations familiales, ainsi que le statut d'occupation du logement et le loyer s'il y a lieu. Le taux de réponse observé à cette enquête obligatoire est de 90% environ. Le mode de collecte est en principe le questionnaire papier à retourner ou la télédéclaration par internet, le choix étant laissé au répondant.

Le troisième élément du système de recensement suisse est une enquête thématique annuelle, avec un échantillon de base de 10'000 à 40'000 répondants selon le thème. Chacun des cinq thèmes suivants est traité à son tour tous les cinq ans : la mobilité et les transports, la formation initiale et continue, la santé, la famille, les langues les religions et les cultures. L'échantillon de ces enquêtes est constitué de 3 à 4 blocs sélectionnés dans les différents cadres d'échantillonnage de l'année d'enquête. La collecte est principalement réalisée par entretien téléphonique assisté par ordinateur. Enfin, le quatrième élément du système de recensement est une enquête téléphonique annuelle auprès d'environ 3'000 répondants, appelée enquête Omnibus, sur des thèmes qui sont décidés chaque année.

Tous ces échantillons sont sélectionnés dans le cadre d'échantillonnage le plus récent. L'OFS fournit également des échantillons tirés dans le SRPH pour les enquêtes menées par d'autres institutions sur des thèmes d'intérêt national, dont en particulier les projets soutenus par le Fonds National de la Recherche Scientifique.

L'apparition de l'enquête structurelle, qui touche chaque année une part non négligeable de la population suisse, a motivé l'utilisation d'une procédure pour coordonner les échantillons de l'OFS. En effet, si l'on ne

prenait aucune mesure pour l'éviter, des dizaines de milliers de personnes seraient sélectionnées à deux enquêtes structurelles successives ou à plusieurs enquêtes de l'OFS en même temps, sans que cela n'ait une utilité statistique. Cela pourrait conduire à une baisse des taux de réponse, à une dégradation de l'image de l'office et à une surcharge de travail liée au traitement des plaintes venant du public. On essaye donc d'éviter de solliciter de manière répétée les personnes et les ménages lorsque cela n'est pas utile. La procédure de sélection coordonnée d'échantillons utilisée à l'OFS pour répondre à ce besoin est décrite en Section 2.3.

2.2.2 Les enquêtes sur les conditions de vie, le budget des ménages et la population active

Outre les enquêtes du système de recensement, l'OFS réalise chaque année trois grandes enquêtes par échantillonnage auprès des ménages et des personnes résidant de manière permanente en Suisse dans des ménages privés. Il s'agit de l'enquête sur les revenus et conditions de vie SILC (Survey on Income and Living Conditions), l'Enquête sur le Budget des Ménages (EBM), et l'Enquête Suisse sur la Population Active (Espa).

L'enquête SILC est la source de référence pour les comparaisons statistiques en matière de distribution de revenu et d'exclusion sociale pour l'Union Européenne. Elle est, depuis 2007, le pendant suisse de l'enquête européenne EU-SILC réalisée dans plus de 30 pays d'Europe. Elle doit de ce fait satisfaire aux exigences et recommandations données par Eurostat dans de nombreux domaines, en particulier sur la manière de calculer les pondérations (Eurostat, 2004b, 2005b; Graf, 2008a), sur les tailles minimales des échantillons de répondants, ménages et individus (Graf, 2006), et enfin sur la précision des indicateurs clés qui sont livrés chaque année à Eurostat.

L'enquête SILC utilise un panel rotatif de répondants : chaque année, un quart de l'échantillon est renouvelé, et chaque ménage enquêté est théoriquement sollicité pendant quatre années. Elle permet ainsi de produire des résultats pour une année donnée, mais aussi d'estimer des évolutions avec une bonne précision. SILC est réalisée annuellement par téléphone auprès d'environ 7'000 ménages comprenant 17'000 personnes.

L'objectif principal de l'enquête Espa est de fournir des données sur la structure de la population active et sur les comportements en matière d'activité professionnelle. Grâce à l'application stricte de définitions internationales, les données de la Suisse peuvent être comparées avec celles des pays de l'OCDE et de l'Union européenne. Il s'agit d'une enquête auprès des personnes, qui est réalisée chaque année depuis 1991. L'organisation de l'enquête a été revue en 2010 pour permettre d'obtenir des estimations trimestrielles, mais aussi des estimations d'évolutions trimestrielles et annuelles précises. Les unités enquêtées sont ré-interrogées après 3, 12 et 15 mois, avant de sortir de l'échantillon. À un trimestre donné, l'échantillon est constitué de quatre blocs. L'un est interrogé pour la première fois, et les autres respectivement pour la deuxième, troisième et quatrième fois (Renfer, 2009; OFS, 2012). L'échantillon de l'Espa est complété par un

échantillon de personnes de nationalité étrangère sélectionné dans le registre central des étrangers (Système d'information central sur la migration - Symic). L'Espa est réalisée auprès d'un échantillon annuel d'environ 105'000 personnes, augmenté d'un sur-échantillon annuel de 21'000 personnes étrangères.

Les résultats de l'EBM permettent d'adapter chaque année la composition du panier-type pour le calcul de l'indice suisse des prix à la consommation, ainsi que d'étudier la structure et l'évolution des revenus et des dépenses des ménages. Cette enquête se base sur des fondements et définitions méthodologiques en accord avec les directives internationales, notamment la nomenclature des fonctions de consommation des ménages COICOP (Classification Of Individual COnsumption by Purpose) définie par le Bureau International du Travail (BIT). L'EBM comporte environ 3'000 ménages interviewés chaque année. L'échantillon de l'EBM est partitionné en douze sous-échantillons qui sont interrogés chacun leur tour durant un mois. Cette enquête est réalisée annuellement à l'OFS depuis 2000.

Hormis l'échantillon spécifique d'étrangers de l'Espa, les échantillons de ces trois enquêtes sont tous actuellement sélectionnés dans le registre de numéros de téléphone Castem (Cadre de sondage pour le tirage d'échantillons de ménages) constitué à partir des livraisons des opérateurs téléphoniques actifs en Suisse. La collecte de données est en principe réalisée par entretien téléphonique assisté par ordinateur, éventuellement complétée par des questionnaires papiers ou électroniques. Afin de limiter la charge d'enquête demandée à la population, les numéros de téléphone qui ont été sélectionnés pour une enquête sont écartés pour une durée déterminée de la liste des numéros sélectionnables. Cette pratique est, à l'OFS, appelée "historisation". Cette procédure n'est acceptable sur le plan méthodologique que lorsque les échantillons représentent une part négligeable et non spécifique de la population. D'ici à la fin 2014, les échantillons de ces trois enquêtes seront désormais sélectionnés dans le cadre d'échantillonnage SRPH. Elles pourront ainsi être coordonnées avec les enquêtes du système de recensement. La procédure "d'historisation" des numéros de téléphone déjà sélectionnés sera abandonnée au profit de la méthode de coordination développée pour le système de recensement.

2.3 Enquêtes coordonnées dans le fichier SRPH

La méthode décrite dans cette section permet la sélection d'échantillons positivement ou négativement coordonnés, c'est à dire d'échantillons dont l'intersection est volontairement importante ou au contraire petite, voire vide. Une coordination négative est utile pour répartir la charge d'enquête équitablement sur la population en évitant que les mêmes unités soient sélectionnées dans plusieurs échantillons lorsque cela n'est pas nécessaire. Elle devient particulièrement importante lorsqu'un grand nombre d'enquêtes sont réalisées, dont certaines ont de forts taux de sondage. Une coordination positive est souhaitable lorsque l'on veut mettre à jour un échantillon de panel, ou lorsque l'on veut pouvoir comparer avec précision

les résultats d'une nouvelle enquête avec ceux d'une enquête précédente. La méthode proposée est utilisable avec des probabilités d'inclusion inégales, dans une population dynamique avec des naissances, ou entrées, et des décès, ou sorties, d'unités.

La coordination des échantillons est un problème qui a et qui continue d'être largement étudié, et pour lequel de nombreuses solutions ont été développées en particulier par les Instituts Nationaux de Statistique (INS). On peut entre autres se référer à Patterson (1950); Keyfitz (1951); Kish & Scott (1971); Rosén (1997a,b); Brewer et al. (1972); De Ree (1983); Van Huis et al. (1994a,b); Cotton & Hesse (1992a,b); Rivière (1998, 1999, 2001a,b); Ohlsson (1995); Ernst (1996) et Kröger et al. (1999). Les méthodes de coordination développées dans les INS sont essentiellement prévues pour fonctionner avec des plans de sondage aléatoires simples ou stratifiés. Il est toutefois apparu très difficile de développer une méthode de tirage coordonnée qui respecte exactement ces plans de sondage lorsque la population et la définition des strates changent entre les enquêtes. A l'exception de l'enquête Espa qui contient un sur-échantillon d'étrangers, les plans de sondage pour les enquêtes de l'OFS ont tous des spécifications très semblables : les probabilités d'inclusion sont uniformes au sein des cantons ou des communes pour toutes les personnes résidentes permanentes âgées de 15 ans ou plus. Cependant, les échantillons que l'OFS fournit à ses partenaires pour leurs propres enquêtes peuvent avoir des caractéristiques très diverses. Ces instituts peuvent par exemple souhaiter cibler des sous-populations spécifiques identifiables grâce aux données contenues dans le SRPH. Il a donc été nécessaire de développer un système très souple qui permette de se départir d'une stratification fixe.

Brewer et al. (1972) ont inventé l'approche dite des numéros aléatoires permanents. Ce sont des méthodes qui reposent sur l'utilisation d'un même jeu de nombres aléatoires pour sélectionner les différents échantillons au cours du temps. Dans certains cas, dont celui qui est présenté ici, un nombre aléatoire entre 0 et 1 est attribué à chaque unité lorsqu'elle entre dans la population ou dans le cadre de sondage, et ce nombre lui reste attaché jusqu'à ce qu'elle sorte de la population ou du cadre. D'autres méthodes nécessitent de permuter les nombres aléatoires entre unités de la population, suivant les tirages passés (Cotton & Hesse, 1992a; Rivière, 2001a). Une des caractéristiques de la méthode de Brewer et coll. Brewer et al. (1972) est qu'elle permet de choisir librement les probabilités d'inclusion de chaque unité de la population à chaque enquête. Elle ne requiert donc pas de stratification fixe de la population et possède la flexibilité nécessaire à l'OFS. Une présentation des méthodes à nombres aléatoires permanent se trouve dans Ohlsson (1995).

2.3.1 Coordination d'échantillons

Il existe différentes définitions de ce qu'est la coordination d'échantillons. Le fait que les processus aléatoires qui permettent de sélectionner les échantillons de deux enquêtes soient ou non indépendants n'est pas ce qui est réellement intéressant. Une définition utile conduit à minimiser

ou bien à maximiser, et généralement à contrôler le nombre d'unités communes à deux échantillons. La notion que nous utilisons ici est légèrement différente, et plus adaptée au cas des enquêtes à probabilités inégales : nous travaillons sur les probabilités, pour chaque unité, d'appartenir simultanément à deux ou une certaine collection d'enquêtes. Considérons un plan de sondage pour T enquêtes, c'est à dire une loi de probabilité

$$\mathrm{P}(s^1, s^2, \ldots, s^T),$$

sur les T échantillons sélectionnables $s^1, s^2, \ldots, s^T$. Ces échantillons ne sont pas nécessairement tous sélectionnés dans la même population. Les distributions marginales d'ordre 1 de $\mathrm{P}(\cdot)$ sont les plans de sondages (appelés transversaux) des T enquêtes considérées. La probabilité qu'une unité k soit sélectionnée dans un échantillon s^i est désignée par

$$\pi_k^i = \mathrm{P}(s^i \ni k),$$

et la probabilité qu'une unité k soit sélectionnée dans l'échantillon s^i et dans l'échantillon s^j est notée

$$\pi_k^{i,j} = \mathrm{P}(s^i \ni k \ \& \ s^j \ni k).$$

Definition 2.1 *On dit qu'il y a une coordination positive pour l'unité k entre les enquêtes i et j si*

$$\pi_k^{i,j} > \pi_k^i \cdot \pi_k^j,$$

et qu'il y a une coordination négative si au contraire

$$\pi_k^{i,j} < \pi_k^i \cdot \pi_k^j.$$

La coordination, positive ou négative, est dite optimale pour l'unité k si le maximum ou minimum théorique de $\pi_k^{i,j}$*, qui vaut respectivement* $\min(\pi_k^i, \pi_k^j)$ *et* $\max(0, \pi_k^i + \pi_k^j - 1)$*, est atteint.*

Si la coordination est optimale et positive (resp. négative) entre les enquêtes i et j, au sens de la définition 2.1, pour toutes les unités, alors la taille espérée de l'intersection des échantillons s^i et s^j,

$$n^{i,j} = \mathrm{E}\left[\#(s^i \cap s^j)\right] = \sum_k \pi_k^{i,j},$$

est maximale (resp. minimale). Cela signifie que parmi toutes les distributions de probabilité $\mathrm{P}(\cdot)$ sur les couples d'échantillons, ayant des marginales fixées $\mathrm{P}(s^i)$ et $\mathrm{P}(s^j)$, la valeur de $n^{i,j}$ est maximale (resp. minimale) lorsque $\mathrm{P}(\cdot)$ fournit une coordination positive optimale (resp. négative optimale) pour toutes les unités k selon la définition 2.1. Cela n'implique pas que la taille $\#(s^i \cap s^j)$ de l'intersection des échantillons s^i et s^j est constante, maximale ou minimale, et égale à $n^{i,j}$.

2.3.2 Algorithme d'échantillonnage coordonné

La méthode que nous présentons ici est une extension de la méthode de Brewer et al. (1972). Chaque unité de la population y est traitée indépendamment des autres unités. Les plans de sondage transversaux, pour chaque enquête prise séparément, sont donc des plans dits de Poisson. L'échantillonnage de Poisson, voir par exemple Tillé (2006), est le plan a probabilités inégales qui est obtenu en sélectionnant ou pas chaque unité k de la population dans l'échantillon avec une probabilité π_k, indépendamment des autres unités de la population. La loi de probabilité correspondante est donnée par

$$\mathrm{P}(s) = \prod_{k \in s} \pi_k \prod_{k \in U \setminus s} (1 - \pi_k).$$

Les sélections étant indépendantes, la taille de l'échantillon ne peut pas être imposée. Comme écrit dans Brewer et al. (1984), le plan de sondage de Poisson a trois défauts apparents par rapport aux plans de sondage de taille fixe. Le premier défaut, selon les auteurs, est que la probabilité de sélectionner l'échantillon vide n'est pas nulle, le second est que l'estimateur de Horvitz-Thompson (Horvitz & Thompson, 1952) a souvent une variance plus grande avec ce plan de sondage, et le troisième défaut est que l'allocation optimale de l'échantillon entre différents domaines ou strates ne peut pas être obtenue car la taille de l'échantillon dans chaque domaine est aléatoire. On peut ajouter à ces défauts que la variabilité de la taille d'échantillon ne permet pas d'avoir un contrôle total sur le budget d'enquête.

Ces défauts sont en fait présents pour tous les plans de sondage auprès des ménages et des personnes lorsque la non-réponse est possible. Le premier et le troisième point ne sont pas pertinents pour les enquêtes de l'OFS. En effet, l'échantillon est alloué, en espérance, proportionnellement à la taille de la population dans le canton ou la zone administrative. Les tailles espérées d'échantillon sont parfois augmentées dans certains cantons pour permettre d'obtenir des résultats détaillés pour ceux-ci. Aucune optimisation n'est recherchée. La publication des résultats n'est envisagée que pour des domaines où la taille espérée de l'échantillon est tellement grande que la probabilité de sélectionner l'échantillon vide est infime, et où la variabilité de la taille d'échantillon due au plan de sondage est négligeable par rapport à celle due à la non-réponse. L'aléa induit sur les coûts de collecte est également négligeable pour les enquêtes de l'OFS. Le second défaut relevé par Brewer et al. (1984) est pallié par l'utilisation d'estimateurs par le ratio (Strand, 1979; Hájek, 1981) ou d'estimateurs calés (Deville & Särndal, 1992). En effet, la non-réponse conduit systématiquement l'OFS à utiliser une méthode de calage ou apparentée. Les institutions partenaires qui commandent des échantillons à l'OFS sont informées de ces possibles problèmes et encouragées à prévoir leurs tailles d'échantillons espérées dans les domaines en fonction de ceux-ci. Il convient de noter que le principal risque pour l'enquête, que ce soit pour son budget ou pour sa qualité, est lié à l'anticipation des taux de réponse qui peut-être mauvaise et non à l'aléa contrôlé dû à la méthode de sélection.

Une présentation complète de l'algorithme de tirage coordonné utilisé à l'OFS est donnée dans Qualité (2009). La méthode peut être présentée en quelques points. Chaque unité k reçoit un nombre aléatoire permanent u_k lorsqu'elle apparaît dans la population. Ce nombre aléatoire est généré selon une loi uniforme sur $[0,1]$, indépendamment des nombres attribués aux autres unités de la population. Ces nombres aléatoires sont ensuite utilisés pour tous les tirages d'échantillons tant que l'unité fait partie de la population. Si la probabilité d'inclusion de l'unité k à l'enquête t est notée π_k^t, cette unité est sélectionnée pour cette enquête lorsque u_k est inclus dans un sous-ensemble mesurable de $[0,1]$ (en fait pour nos besoins dans une union finie d'intervalles), appelé zone de sélection, de longueur totale égale à π_k^t. De cette manière, les probabilités d'inclusion sont respectées, dans la mesure où le choix de la zone de sélection n'est pas fonction de la valeur de u_k. La coordination entre les enquêtes est alors obtenue en choisissant des zones de sélection dont l'intersection est contrôlée pour les différentes enquêtes. On obtient une coordination positive maximale entre deux enquêtes lorsque les zones de sélection correspondantes ont un recouvrement maximal, c'est-à-dire lorsqu'elles sont incluses l'une dans l'autre. On obtient au contraire une coordination négative maximale lorsque les zones de sélection se recouvrent le moins possible, c'est-à-dire, dans la mesure du possible, où elles sont disjointes. Ainsi, la construction de ces zones de sélection pour chaque enquête est l'élément essentiel de notre méthode d'échantillonnage coordonné. Son fonctionnement est aisément compris en considérant un exemple. Comme les unités de la population sont traitées de manière indépendante, il suffit de considérer une unité générique.

Supposons que cette unité a des probabilités d'inclusions égales à π_k^1, π_k^2, π_k^3, et π_k^t respectivement à la première, deuxième, troisième et $t-$ième enquête.

1. Pour la première enquête, on choisit naturellement l'intervalle $[0, \pi_k^1[$ comme zone de sélection (voir figure 2.1).

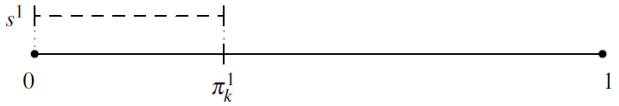


Figure 2.1 – *Zone de sélection pour la première enquête.*

2. La deuxième enquête peut être soit positivement soit négativement coordonnée avec la première. Nous n'envisageons que des cas où l'on va chercher à avoir une coordination maximale. Si la coordination voulue est positive, la zone de sélection choisie pour la deuxième enquête est l'intervalle $[0, \pi_k^2[$ (voir par exemple, si $\pi_k^2 \leq \pi_k^1$, la figure 2.2). Dans la situation représentée par la figure 2.2, si u_k est dans l'intervalle $[0, \pi_k^2[$, alors l'unité k est sélectionnée à la première et à la deuxième enquête. Si ce nombre est entre π_k^2 et π_k^1, alors k est sélectionnée à la première mais pas à la deuxième enquête, et si u_k est plus grand que π_k^1, alors k n'est pas sélectionnée. La probabilité

Figure 2.2 – *Coordination positive, avec* $\pi_k^2 \leq \pi_k^1$.

de sélectionner k dans les deux enquêtes est donc égale au minimum de π_k^1 et π_k^2, ce qui est le maximum théorique.

Si au contraire, la coordination souhaitée est négative, deux cas sont possibles :

- Si la probabilité d'inclusion π_k^2 est telle que $\pi_k^1 + \pi_k^2 \leq 1$, alors il est possible de faire en sorte que l'unité k ne soit prise que dans un seul des deux échantillons. Pour cela, on choisit comme zone de sélection à la deuxième enquête l'intervalle $[\pi_k^1, \pi_k^1 + \pi_k^2[$ (voir figure 2.3).

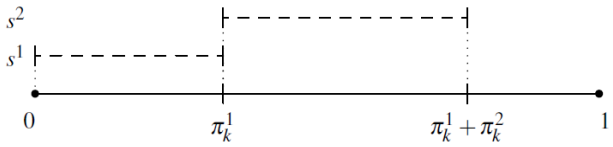


Figure 2.3 – *Coordination négative lorsque* $\pi_k^1 + \pi_k^2 \leq 1$.

- Si $\pi_k^1 + \pi_k^2 > 1$, on ne pourra exiger que l'unité k ne soit jamais sélectionnée dans les deux échantillons. La probabilité que k soit prise conjointement dans les deux échantillons est alors au minimum égale à $\pi_k^1 + \pi_k^2 - 1$. Ce minimum est obtenu en utilisant la zone de sélection $[\pi_k^1, 1] \cup [0, \pi_k^1 + \pi_k^2 - 1[$ à la deuxième enquête (voir figure 2.4).

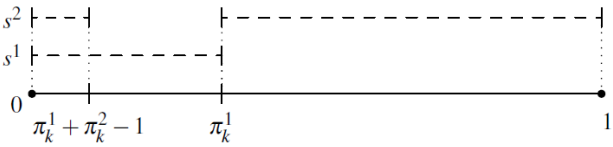


Figure 2.4 – *Coordination négative lorsque* $\pi_k^1 + \pi_k^2 \geq 1$.

3. La troisième enquête, quant à elle, peut être positivement ou négativement coordonnée avec la première enquête, et également positivement ou négativement coordonnée avec la deuxième enquête. Ces exigences ne peuvent pas toujours être toutes satisfaites. Par exemple, si les deux premières enquêtes sont coordonnées positivement entre elles, et que l'on souhaite que la troisième soit coordonnée positivement avec l'une des deux, et négativement avec l'autre, on ne peut espérer un très bon résultat pour ces deux coordinations. Un autre exemple est le cas où $\pi_k^1 = \pi_k^2 = \pi_k^3 = 0.5$, et où l'on veut

que les enquêtes soient négativement coordonnées. On pourra bien faire en sorte que k ne puisse être sélectionné simultanément dans le premier et dans le deuxième échantillon. Le recouvrement entre le troisième échantillon et l'un au moins des deux premiers est par contre inévitable. Il faut donc choisir quelle coordination l'on veut privilégier.

La méthode utilisée à l'OFS repose sur un choix de l'ordre de priorité pour les coordinations entre une nouvelle enquête et les enquêtes passées. Un exemple suffit à comprendre son fonctionnement : supposons que les deux premières enquêtes étaient positivement coordonnées, comme dans la figure 2.2, et que l'on souhaite que la troisième enquête soit positivement coordonnée avec la seconde, puis, avec une priorité moindre, négativement coordonnée avec la première. Supposons de plus que $\pi_k^3 > \pi_k^2$. Le premier objectif est atteint si la zone de sélection pour la troisième enquête a le plus grand recouvrement possible avec la zone de sélection pour la deuxième enquête. Puisque $\pi_k^3 > \pi_k^2$, on va choisir d'inclure l'intervalle $[0, \pi_k^2[$ dans cette zone de sélection. Puis, il faut compléter cette zone de sélection de manière à ce que sa longueur totale soit égale à π_k^3, et à respecter au mieux les coordinations voulues. On va naturellement choisir d'ajouter l'intervalle $[\pi_k^1, \pi_k^3 + \pi_k^1 - \pi_k^2[$ comme représenté sur la figure 2.5.

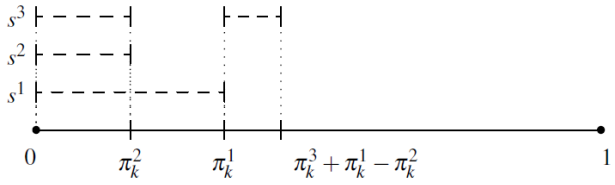


Figure 2.5 – *Coordination d'un troisième échantillon.*

4. Dans le cas général, après $t-1$ enquêtes, le segment $[0,1]$ est divisé en t intervalles, qui sont les intersections des zones de sélection des précédentes enquêtes. Les intervalles correspondent aux séquences de sélection possibles pour l'unité k aux $t-1$ enquêtes passées. On veut sélectionner un $t-$ième échantillon qui doit être négativement coordonné avec certaines enquêtes passées et positivement coordonné avec d'autres enquêtes passées. Les exigences de coordination pour la nouvelle enquête permettent de définir un ordre sur les t intervalles. Si les coordinations avec toutes les enquêtes passées sont spécifiées alors cet ordre est total, et permet de classer les intervalles, et séquences de sélection correspondantes par ordre de compatibilité avec les règles de coordination demandées.

   Une manière de calculer cet ordre est présentée dans (Qualité, 2009, p.105) : pour chaque unité k (indice omis par la suite) et chaque intervalle a_i, $i = 1, \ldots, t$, on calcule un score $S(a_i)$. Ce score est

défini par :

$$S(a_i) = \sum_{j=1}^{t-1} 2^{j-1} c_{t-j}(a_i),$$

où $c_{t-j}(a_i) = 1$ si l'intervalle a_i est compatible avec la coordination souhaitée pour l'enquête qui a la $t-j$-ième priorité, et 0 sinon. Par exemple, si l'on veut une coordination positive avec l'enquête qui a la priorité 1 (la priorité la plus élevée), on définit $c_1(a_i) = 1$ si l'intervalle a_i correspond à une sélection à cette enquête, et 0 si a_i correspond à une non-sélection à cette enquête. Si au contraire on veut une coordination négative, on définit $c_1(a_i) = 0$ si l'intervalle a_i correspond à une sélection à cette enquête, et 1 si a_i correspond à une non-sélection à cette enquête. De cette manière, les intervalles a_i tels que $c_1(a_i) = 1$ auront tous un score $S(a_i)$ supérieur ou égal à 2^{t-2} et ceux tels que $c_1(a_i) = 0$ auront tous un score $S(a_i)$ strictement inférieur à 2^{t-2}. On peut montrer que la fonction score $S(\cdot)$ ainsi définie est une injection dans $\{0, \ldots, 2^{t-1} - 1\}$, et que l'ordre induit par ce score est compatible avec l'ordre de priorité voulu sur les coordinations. En pratique, cette méthode de calcul est assez rapidement problématique puisque les scores calculés peuvent être très grands. Il sera donc utile de trouver une méthode adaptée au langage informatique utilisé.

La zone de sélection pour la nouvelle enquête est obtenue en incluant les intervalles les mieux classés, jusqu'à dépasser une longueur totale de π_k^t. Le dernier intervalle inclus est ensuite découpé, et seulement une partie est conservée dans la zone de sélection, de manière à ce que celle-ci ait une longueur totale égale à π_k^t. Le segment $[0, 1]$ est alors divisé en $t+1$ intervalles qui correspondent aux $t+1$ séquences de sélections possibles. La croissance linéaire du nombre d'intervalles à considérer est un des aspects de la méthode qui permet d'envisager son utilisation sur une période assez longue. Plus de 60 enquêtes auprès des environ 8 millions de personnes résidentes en Suisse ont déjà été sélectionnées en utilisant une implémentation assez grossière de cet algorithme.

Exemple 2.1 *Considérons une unité dans une population qui subit trois enquêtes: une enquête par panel s^1, une enquête unique s^2, un autre panel s^3, puis une mise à jour s^4 du premier panel et enfin une mise à jour s^5 du deuxième panel. Les probabilités d'inclusion de l'unité dans ces différents échantillons sont respectivement notées $\pi^1, \ldots, \pi^5$. Les deux panels et l'enquête ponctuelle sont, entre eux, coordonnés négativement avec des priorités définies par l'ancienneté des enquêtes. Par souci de simplicité, on suppose que la probabilité de tirage à l'ensemble de ces enquêtes est strictement inférieure à 1, que $\pi^4 > \pi^1$ et $\pi^5 < \pi^3$. Les intervalles et échantillons longitudinaux correspondants sont :*

1. *lors du premier tirage,*
   - $a_1 = [0, \pi^1]$ *avec* $s = 1$ *et*
   - $a_2 =]\pi^1, 1]$ *avec* $s = 0$,

2. *lors du deuxième tirage,*

   - $a_1 = [0, \pi^1]$ *avec* $s = (1, 0)$,
   - $a_2 =]\pi^1, \pi^1 + \pi^2]$ *avec* $s = (0, 1)$ *et*
   - $a_3 =]\pi^1 + \pi^2, 1]$ *avec* $s = (0, 0)$,

3. *lors du troisième tirage,*

   - $a_1 = [0, \pi^1]$ *avec* $s = (1, 0, 0)$,
   - $a_2 =]\pi^1, \pi^1 + \pi^2]$ *avec* $s = (0, 1, 0)$,
   - $a_3 =]\pi^1 + \pi^2, \pi^1 + \pi^2 + \pi^3]$ *avec* $s = (0, 0, 1)$ *et*
   - $a_4 =]\pi^1 + \pi^2 + \pi^3, 1]$ *avec* $s = (0, 0, 0)$,

4. *lors du quatrième tirage,*

   - $a_1 = [0, \pi^1]$ *avec* $s = (1, 0, 0, 1)$,
   - $a_2 =]\pi^1, \pi^1 + \pi^2]$ *avec* $s = (0, 1, 0, 0)$,
   - $a_3 =]\pi^1 + \pi^2, \pi^1 + \pi^2 + \pi^3]$ *avec* $s = (0, 0, 1, 0)$,
   - $a_4 =]\pi^1 + \pi^2 + \pi^3, \pi^2 + \pi^3 + \pi^4]$ *avec* $s = (0, 0, 0, 1)$ *et*
   - $a_5 =]\pi^2 + \pi^3 + \pi^4, 1]$ *avec* $s = (0, 0, 0, 0)$,

5. *et lors du cinquième tirage,*

   - $a_1 = [0, \pi^1]$ *avec* $s = (1, 0, 0, 1, 0)$,
   - $a_2 =]\pi^1, \pi^1 + \pi^2]$ *avec* $s = (0, 1, 0, 0, 0)$,
   - $a_3 =]\pi^1 + \pi^2, \pi^1 + \pi^2 + \pi^5]$ *avec* $s = (0, 0, 1, 0, 1)$,
   - $a_4 =]\pi^1 + \pi^2 + \pi^5, \pi^1 + \pi^2 + \pi^3]$ *avec* $s = (0, 0, 1, 0, 0)$,
   - $a_5 =]\pi^1 + \pi^2 + \pi^3, \pi^2 + \pi^3 + \pi^4]$ *avec* $s = (0, 0, 0, 1, 0)$ *et*
   - $a_6 =]\pi^2 + \pi^3 + \pi^4, 1]$ *avec* $s = (0, 0, 0, 0, 0)$.

Les priorités définies sur les coordinations prennent toute leur importance lorsque certaines coordinations sont contradictoires (cas que nous n'avons jamais rencontré à l'OFS), ou lorsque la probabilité d'inclusion des unités dans l'ensemble des échantillons atteint 1. Considérons que l'on sélectionne un échantillon pour une sixième enquête, avec une coordination négative par ordre chronologique avec les anciennes enquêtes. L'ordre de préférence pour définir la zone de tirage dans s^6 *est:* a_6 *(aucune sélection antérieure) puis* a_2 *(dernière sélection dans* s^2*) puis* a_4 *(dernière sélection dans* s^3*) puis* a_5 *et* a_1 *(dernière sélection dans* s^4*, mais une sélection de plus pour* a_1*), et enfin* a_3 *(dernière sélection dans* s^5*). Si* $\pi^6 > 1 - \pi^2 - \pi^3 - \pi^4$*, cet ordre de préférence permet de choisir lesquels des intervalles* $a_1, \ldots, a_5$ *sont inclus ou partiellement inclus dans la zone de sélection pour* s^6 *bien qu'ils ne respectent pas parfaitement toutes les coordinations voulues.*

2.3.3 Sélection d'une seule personne par ménage

Afin d'obtenir un bon taux de réponse, il a été décidé à l'OFS de ne pas interroger plus d'une personne par ménage à une enquête donnée lorsque cela n'est pas nécessaire pour obtenir l'information nécessaire aux estimations. Les exceptions notables à l'OFS sont l'enquête sur le budget des ménages (EBM) et l'enquête sur les revenus et conditions de vie (SILC), où, au contraire, tous les membres des ménages sélectionnés sont interrogés. Jusqu'à fin 2013, SILC et l'EBM n'étaient pas sélectionnées dans le fichier SRPH et nous n'avons pas été confrontés à ce dernier problème.

La méthode de tirage avec coordination décrite dans la section 2.3.2 ne permet pas d'obtenir directement ce résultat. En effet, elle repose sur des sélections indépendantes entre les unités, et ne permet donc pas d'exclure la sélection simultanée de deux personnes d'un même ménage, ou au contraire de forcer la sélection simultanée de toutes les personnes d'un même ménage. Les enquêtes pour lesquelles on s'est interdit de sélectionner plusieurs personnes par ménage sont l'enquête structurelle, l'enquête Omnibus et les enquêtes thématiques. Les échantillons des deux premières sont sélectionnés en une fois dans un seul cadre de sondage tandis que les enquêtes thématiques sont sélectionnées par blocs dans trois ou quatre fichiers SRPH successifs. Pour ces dernières, il faut ainsi considérer la difficulté supplémentaire que la composition des ménages peut changer entre les tirages des différents blocs.

Pour obtenir le résultat voulu, on a choisi comme principe de sélectionner les échantillons selon une procédure de tirage en plusieurs phases. La première phase consiste en la sélection coordonnée d'un échantillon de personnes, selon la méthode de la section 2.3.2, et avec des probabilités d'inclusion à déterminer. Les phases suivantes consistent à éliminer les cas indésirables en :

- ne retenant qu'une unité dans chaque ménage où plusieurs ont été sélectionnées en première phase,
- et pour les enquêtes thématiques en éliminant toutes les sélections dans des ménages dont un membre participe déjà à l'enquête.

Afin de rendre ce procédé le plus simple possible, le choix de la personne retenue lors de la deuxième phase est effectué à probabilités égales parmi les unités d'un même ménage présélectionnées en première phase. Les probabilités d'inclusion à la première phase sont calculées de manière à obtenir les probabilités d'inclusion finales souhaitées sur l'ensemble des deux ou trois phases.

Enquête structurelle et enquête Omnibus

Les personnes d'un même ménage qui sont sélectionnables à ces enquêtes reçoivent des probabilités de sélection égales. En effet, les probabilités d'inclusion dans les enquêtes de l'OFS ne sont fonctions que du canton ou de la commune de résidence, et de l'appartenance à la population cible de l'enquête.

Le calcul explicite des probabilités de sélection de première phase en fonction des probabilités d'inclusion finales souhaitées est alors possible. Notons p la probabilité d'inclusion de première phase de chacune des m personnes sélectionnables d'un ménage. Si la deuxième phase consiste en la sélection avec probabilités égales d'une seule personne parmi celles sélectionnées en première phase, on obtient que la probabilité d'inclusion finale π d'une de ces personnes est égale à

$$\pi = \frac{1}{m}\left[1 - (1-p)^m\right]. \tag{2.1}$$

En effet, la probabilité de sélectionner au moins une personne dans le ménage est égale à $1-(1-p)^m$. Les probabilités conditionnelles de sélectionner l'un ou l'autre individu en deuxième phase sachant que le ménage contient des unités présélectionnées en première phase sont égales (les individus sont échangeables), et la somme de ces probabilités conditionnelles vaut 1 (on retient exactement une unité par ménage dans lequel des unités ont été présélectionnées). Chacune de ces probabilités conditionnelles vaut donc $1/m$. On obtient l'équation 2.1 en multipliant ces deux termes.

On peut déduire le paramètre p de la probabilité d'inclusion π souhaitée, en inversant l'équation 2.1, pour autant que $m\pi \leq 1$. La probabilité d'inclusion en première phase vaut alors

$$p = 1 - (1 - m\pi)^{\frac{1}{m}}. \tag{2.2}$$

Pour les très grands ménages dans les cantons qui ont choisi de financer une augmentation de leur échantillon pour l'enquête structurelle, la valeur de $m\pi$ peut dépasser 1. En effet, le taux de sondage de base des personnes π est, dans ces cantons, proche de 8%. Or, dans quelques ménages, m est supérieur ou égal à 13. Il n'y a alors pas de solution au problème. Nous avons simplement attribué une probabilité d'inclusion de première phase égale à 1 aux personnes de ces ménages.

Le plan de sondage obtenu est exactement identique à un plan de sondage en deux phases où la première phase serait une sélection de ménages à l'aide d'un plan de Poisson, avec probabilités d'inclusion proportionnelles au nombre de personnes sélectionnables du ménage, et la deuxième phase la sélection avec probabilités égales d'une personne par ménage retenu. La raison pour laquelle nous n'avons pas directement fait cela est que la constitution des ménages dans le cadre de sondage, et en particulier le suivi de ces ménages dans le temps, n'étaient pas assurées entre septembre 2010, date de création du premier registre d'échantillonnage SRPH, et décembre 2012. En effet, l'obligation légale, pour les cantons et communes, de fournir un identifiant de ménage pour chaque personne n'entrait en vigueur qu'au 31 décembre 2012. Il paraissait ainsi plus sage d'organiser la coordination des échantillons directement au niveau des personnes.

La méthode utilisée pour n'obtenir qu'une sélection par ménage conduit à une coordination sous-optimale entre les enquêtes. En effet, la

coordination entre échantillons est réalisée entre les échantillons de première phase, dont les probabilités d'inclusion sont légèrement plus élevées que les probabilités d'inclusion finales. Pour la grande majorité des enquêtes, et des ménages cela ne fait pas une différence sensible. Pour les très grands ménages (plus de 13 personnes dans un canton qui aurait doublé son échantillon pour l'enquête structurelle), la coordination entre l'enquête structurelle et les autres enquêtes est inexistante.

Enquêtes thématiques

Le problème est plus complexe pour les enquêtes thématiques qui sont sélectionnés dans plusieurs cadres. Pour les ménages dont la composition reste identique lors des différents tirages, on peut assez simplement trouver une équation similaire à (2.1) pour relier les probabilités d'inclusion dans les différents blocs d'enquête et les probabilités d'inclusion finales. L'équation obtenue permet, lorsqu'une solution existe, de calculer explicitement les paramètres à utiliser lors de chaque tirage. Comme dans le cas des enquêtes structurelles et omnibus, il peut y avoir des problèmes pour certains très grands ménages liés à l'impossibilité de respecter les probabilités d'inclusion souhaitées tout en imposant au maximum une sélection par ménage sur l'ensemble des blocs d'enquête.

Dans les ménages dont la composition évolue, les probabilités d'inclusion des membres du ménage dans chaque bloc d'enquête ne sont pas nécessairement toutes égales. Par exemple, si le ménage contient lors du deuxième tirage une personne qui n'en faisait pas partie lors du premier tirage, il est possible que celle-ci ait eu une probabilité de sélection dans le premier bloc différente de celle des autres membres du ménage. Les calculs deviennent alors plus complexes. En effet, le polynôme (2.1) est alors remplacé, pour chaque personne, par un polynôme fonction des probabilités de première phase p_k de tous les membres du ménage à chaque bloc d'enquête passé. Le degré de ce polynôme est égal au nombre de personnes sélectionnables dans le ménage. Le problème est de calculer les racines de ces polynômes. On ne peut vraisemblablement pas obtenir une solution explicite dans les ménages où plus de quatre personnes sont sélectionnables. En effet, les polynômes de degré cinq ou plus ne sont en général pas résolubles par radicaux. Pour cette raison, nous avons développé une procédure de calcul des probabilités de première phase par approximations numériques successives. Cette procédure repose sur une version simplifiée de l'algorithme de Newton-Raphson dans lequel la différentielle de la fonction à inverser est remplacée par l'identité. En effet, cette différentielle est usuellement très proche de l'identité, et la perte d'efficacité de l'algorithme due à cette approximation est certainement largement compensée par l'économie de calculs réalisée à chaque étape.

Il faut noter que malgré tous nos efforts, la sélection d'échantillons qui respectent parfaitement ces exigences n'est en réalité pas possible car la composition des ménages est manquante pour 2 à 3% de la population dans le cadre d'échantillonnage. De plus, cette composition est susceptible de changer entre la date de constitution du fichier SRPH et la date

d'enquête. Toutefois, l'objectif peut être considéré comme essentiellement rempli si l'on ne sélectionne qu'une personne par ménage selon l'état de nos connaissances.

2.4 Estimation et Pondération

Les estimateurs utilisés par l'OFS sont en principe des estimateurs par dilatation, c'est à dire des sommes pondérées des valeurs observées, ou bien des fonctions d'estimateurs par dilatation, par exemple des ratios de deux estimateurs. Un rôle central est joué par l'estimateur de Horvitz-Thompson (Horvitz & Thompson, 1952) d'un total

$$\widehat{Y}_{HT} = \sum_{k \in s} d_k y_k, \tag{2.3}$$

où les y_k sont les valeurs de la variable dont on veut estimer le total et les $d_k = 1/\pi_k$ sont les inverses des probabilités d'inclusion des individus k dans l'échantillon s. Cet estimateur n'est cependant pratiquement jamais directement utilisé, mais il sert de base pour construire l'estimateur pondéré final

$$\widehat{Y}_w = \sum_{k \in r} w_k y_k,$$

où r est un sous-ensemble de répondants de s et $w_k \geq 0$.

Les poids d'extrapolation w_k sont calculés en trois, voire quatre grandes étapes. La première étape est le calcul des probabilités d'inclusion π_k, qui ne sont parfois pas connues avant l'enquête. La deuxième est la modélisation du mécanisme de non-réponse, et l'estimation de la probabilité de réponse de chaque unité observée. La troisième est le calage (Deville & Särndal, 1992). Suivant les enquêtes, une étape supplémentaire consiste à effectuer un partage des poids (Lavallée, 2002) lorsque les unités enquêtées ne sont pas directement les unités sélectionnées, ou bien une combinaison de poids lorsqu'un échantillon est obtenu par réunion de plusieurs autres échantillons. Chaque enquête nécessite le développement d'une pondération qui lui est propre dépendant de plusieurs facteurs : plan d'échantillonnage, procédé ou mode d'enquête, thématique, informations à disposition externes et internes à l'enquête.

2.4.1 Calcul des probabilités d'inclusion

Les tailles d'échantillons pour les enquêtes de l'OFS sont choisies en fonction de contraintes budgétaires et d'objectifs de précision. La répartition de ces échantillons dans les cantons est effectuée de manière équitable, c'est à dire proportionnellement à la population des cantons. Ces derniers ont la possibilité de financer des extensions de leur échantillon lorsque celui prévu par l'OFS est jugé insuffisant pour répondre à leurs besoins. Le plan de sondage naturel pour ces enquêtes est le sondage aléatoire simple stratifié (Tillé, 2006). Il n'est cependant en pratique quasiment jamais mis en œuvre : parfois le cadre de sondage ne contient pas l'information suffisante. Par exemple, pour sélectionner l'échantillon de l'Espa dans le

répertoire Castem, on ne connaît pas les tailles des ménages à l'avance. Parfois l'échantillon est obtenu en agrégeant des échantillons sélectionnés dans des cadres différents. C'est le cas pour SILC, EBM, et les enquêtes thématiques. Enfin, le plan de sondage utilisé pour les tirages coordonnés dans le SRPH est le plan de Poisson. Les probabilités d'inclusion ne sont en conséquence pas toujours égales au rapport entre la taille de l'échantillon et la taille de la population. Elles doivent être calculées pour chaque enquête.

Ce calcul, point de départ de la construction des estimateurs peut être relativement compliqué en raison de la procédure d'enquête, ou du cadre de sondage utilisé pour atteindre la population cible. Dans les cas les plus simples, les probabilités d'inclusion n'ont pas à être calculées car elles sont spécifiées dans le plan de sondage. C'est par exemple le cas pour une enquête ponctuelle sélectionnée dans le SRPH, telle l'enquête Omnibus du système de recensement. C'est également le cas pour l'enquête EBM, dont l'échantillon de ménages est obtenu en sélectionnant un nombre fixé de numéros de téléphones dans le registre de numéros de téléphone Castem. Dans d'autres cas, les probabilités d'inclusion sont calculables à partir des cadres de sondage. Par exemple, pour les enquêtes thématiques qui sont constituées de vagues d'enquêtes sélectionnées dans trois ou quatre cadres d'échantillonnage SRPH successifs, la probabilité d'inclusion d'une unité est simplement la somme de ses probabilités d'inclusion aux trois ou quatre vagues de l'enquête. Ceci est dû à la coordination négative entre les vagues qui assure qu'elles ne se recouvrent pas.

Une spécificité de l'enquête structurelle est que le questionnaire comprend une clause qui permet aux personnes enquêtées de décider laquelle doit répondre dans le cas où, malgré les procédures mises en place pour l'éviter, plusieurs personnes d'un même ménage reçoivent un questionnaire. Cette procédure, déterministe, repose sur les dates de naissance des personnes du ménage. Elle conduit, dans certains cas à faire diminuer la probabilité de sélection d'une personne lorsque son ménage effectif contient des personnes qui ne sont pas rencensées dans le cadre d'échantillonnage SRPH. Elles pourraient en effet avoir été sélectionnées en même temps, et, selon leur date de naissance, elles auraient alors été désignées par la procédure d'élimination pour répondre au questionnaire. La procédure d'élimination n'a pas toujours été comprise, mais les cas où elle a été appliquée et les cas où elle aurait dû l'être sont très peu nombreux (de l'ordre de quelques centaines). Les probabilités d'inclusion des répondants ont été recalculées en fonction de la composition de leur ménage déclarée à l'enquête et de sa comparaison avec la composition connue dans le cadre. Il faut toutefois noter que cela fait dépendre les probabilités de sélection, dans l'échantillon des personnes qui sont censées répondre au questionnaire, de la composition réelle des ménages déclarée à l'enquête et de leur compréhension ou bonne application des instructions qui accompagnent le questionnaire. En effet, ce sont les personnes contactées elles-même qui ont en charge d'effectuer l'ultime phase de filtrage de l'échantillon pour éliminer les cas de sélections multiples dans un ménage. La probabilité d'inclusion des ménages est elle aussi calcu-

lable en utilisant la composition déclarée des ménages observés, et en la rapprochant des probabilités de tirages connues pour toutes les unités du cadre de sondage.

Les différents panels qui constituent l'échantillon de l'enquête SILC sont des échantillons aléatoires simples de ménages, sélectionnés sans remise et stratifiés par régions géographiques. Les probabilités de tirage par strate sont proportionnelles à la taille relative de la population de chacune des régions. Tous les membres adultes des ménages contactés sont interrogés, soit directement, soit au travers d'une personne du ménage. L'unité d'enquête est donc tant le ménage que l'individu. SILC étant un panel rotatif, chaque année, le panel vieux de 4 ans est abandonné et un nouveau panel rejoint le relevé pour le remplacer. Plusieurs pondérations sont produites : une pondération pour extrapoler les informations au niveau des ménages une année donnée, une pondération pour extrapoler les informations au niveau des personnes une année donnée. Des poids dits "longitudinaux" sont aussi calculés pour les unités qui sont observées plusieurs années à la suite (Graf, 2008a). Les différentes probabilités d'inclusion : dans l'un des panels qui constituent l'échantillon annuel, dans l'un des panels qui sont enquêtés deux années à la suite, et ainsi de suite, sont donc évaluées (voir figure 2.6).

L'échantillon de l'enquête Espa est un panel rotatif de personnes. Il est constitué chaque trimestre de quatre blocs de rotation, dont un est enquêté pour la première fois, et les trois autres ont été enquêtés pour la première fois respectivement le trimestre précédent, un an avant, et quinze mois auparavant (Renfer, 2009; OFS, 2012). Ces blocs sont eux-mêmes sélectionnés selon une procédure mixte. Une partie de l'échantillon est obtenue en tirant des numéros de téléphone dans le répertoire Castem. La liste des personnes de chaque ménage contacté est alors établie interactivement, et l'une des personnes âgées de 15 ans ou plus est sélectionnée aléatoirement pour répondre à l'enquête (Feusi Widmer, 2004). Les probabilités d'inclusion correspondantes sont donc fonction de la taille des ménages. L'autre partie de l'échantillon est sélectionnée directement dans le registre du système d'information central sur la migration. Les probabilités d'inclusion dans chacun des blocs, ainsi que dans leur union un trimestre donné, où sur une période donnée sont évaluées. Il ne peut s'agir là que d'approximations car l'OFS ne dispose pas de toute l'information nécessaire : la composition des ménages n'est connue qu'au moment de l'enquête. On ne sait donc par exemple pas précisément avec quelle probabilité un individu du nouveau bloc de rotation aurait pu être sélectionné dans le bloc de rotation qui répond pour la quatrième fois.

Ces quelques exemples montrent qu'il existe une diversité de situations et de besoins qui conduisent à des calculs différents. La création du cadre d'échantillonnage SRPH et la généralisation de son utilisation pour toutes les enquêtes de l'OFS devrait permettre d'harmoniser ces calculs, et de les simplifier grâce à la connaissance précise de la population et de sa probabilité de sélection dans chaque vague d'enquête ou bloc de rotation pour les enquêtes concernées.

2.4.2 Calcul de probabilités de réponse

Toutes les enquêtes auprès de la population sont entachées de non-réponse, et seule une partie $r \subseteq s$ de l'échantillon sélectionné s répond à une enquête donnée. Cette non-réponse correspond effectivement à un sondage, dont on ne connaît pas le plan, dans l'échantillon sélectionné à l'origine. Pour utiliser l'estimateur de Horvitz-Thompson (2.3) avec les données de l'échantillon observé r, on doit connaître les probabilités d'inclusion dans r. En effet, les valeurs π_k ne sont pas égales aux probabilités d'appartenance à r, et l'estimateur

$$\widehat{Y} = \sum_{k \in r} \frac{y_k}{\pi_k} \tag{2.4}$$

est généralement biaisé. Pour que l'estimateur de Horvitz-Thompson soit sans biais, il faut en outre que les probabilités d'inclusion dans r soient toutes strictement positives. On est donc amené à modéliser le processus de réponse, et à estimer les probabilités de réponse des unités sélectionnées pour construire un estimateur qui ressemble à (2.3).

On désigne par la suite $q(\cdot|s)$ le mécanisme de réponse dans l'échantillon s. C'est la loi de probabilité, inconnue, qui a généré l'échantillon de répondants r parmi les s unités sélectionnées. Le fait qu'une unité sélectionnée ait effectivement répondu est indiqué par une variable aléatoire binaire r_k, indicatrice de réponse (0=pas de réponse, 1=réponse), pour $k \in s$. Il s'agit d'une variable de Bernoulli. Son espérance a priori inconnue est notée θ_k.

Si l'on connaissait les θ_k, on pourrait utiliser l'estimateur de Horvitz-Thompson

$$\widehat{Y}_{HT} = \sum_{k \in r} \frac{y_k}{\pi_k \theta_k}.$$

Dans la pratique, les θ_k sont toujours inconnus, et on les estime en supposant que $q(\cdot|s)$ appartient à une famille de lois spécifiée. Une fois les estimations $\hat{\theta}_k$, $k \in r$ réalisées, on peut estimer le total Y par

$$\widehat{Y} = \sum_{k \in r} \frac{y_k}{\pi_k \hat{\theta}_k}, \tag{2.5}$$

qui n'est au mieux qu'approximativement sans biais. Dans ce cas le facteur $(\pi_k \hat{\theta}_k)^{-1}$ joue le rôle de poids pour l'unité observée k. On l'appelle poids de non-réponse et on parle aussi de repondération pour compenser la non-réponse.

La modélisation de $q(\cdot|s)$ est une étape déterminante de la pondération. On suppose souvent que le plan $q(\cdot|s)$ est un plan de Poisson, c'est à dire que la décision de répondre ou pas d'une unité est indépendante de la décision prise par les autres unités. Cela est pertinent dans une enquête auprès des personnes où la probabilité de sélectionner deux personnes dans le même ménage est extrêmement faible, voire nulle par conception. Dans une enquête auprès des ménages, la non-réponse est dans un premier temps modélisée au niveau du ménage. Cette hypothèse n'est pas suffisante pour pouvoir estimer les θ_k puisque l'on ne dispose que d'une

observation de r_k par paramètre θ_k à estimer. On est donc amené à spécifier des contraintes sur les θ_k afin d'arriver à une situation où ceux-ci peuvent être estimés avec précision. Deux types de contraintes, de nature voisine, sont usuellement envisagées :

1. Les θ_k ont une certaine relation fonctionnelle avec des variables explicatives connues pour les répondants et les non-répondants à l'enquête. Ils sont alors généralement estimés en effectuant une régression logistique des valeurs observées des r_k sur les variables explicatives retenues. C'est la méthode utilisée pour les enquêtes EBM et Espa.

2. La population est partitionnée en groupes de personnes qui partagent la même valeur de θ_k : les Groupes Homogènes de Réponse (GHR). La répartition de l'échantillon dans ces groupes est estimée par une méthode de segmentation (Kass, 1980), en fonction de variables explicatives connues pour les répondants et les non-répondants à l'enquête. Les θ_k sont alors estimés par les taux de réponse observés dans les GHR. Cette méthode est utilisée à l'OFS pour l'enquête SILC.

Selon la constitution du processus d'enquête, plusieurs modélisations de la non-réponse peuvent être requises. C'est par exemple le cas quand, dans une collecte téléphonique, une description du ménage est demandée avant de sélectionner une personne pour l'enquête. Il peut y avoir alors non-réponse avant l'énumération du ménage, ou non-réponse de la personne sélectionnée. L'information disponible pour modéliser la première non-réponse est moins riche que celle disponible pour modéliser la seconde.

La première étape de la modélisation de la non-réponse consiste à déterminer quelles sont les variables explicatives qui doivent être considérées dans le modèle afin de bien prédire la variable dichotomique de réponse (voir LaRoche, 2007). Le choix des variables est parfois assez restreint car elles doivent être connues tant pour les unités répondantes que pour les unités non-répondantes. Pour les enquêtes sélectionnées dans le répertoire Castem, le cadre de sondage contient très peu d'information utilisable. Il est donc essentiel de collecter le maximum d'informations même sur les ménages ou personnes qui décident de ne pas répondre au questionnaire. Pour les enquêtes sélectionnées dans le SRPH, on dispose au contraire d'un nombre important de variables explicatives possibles, mais leur pouvoir explicatif n'est lui pas nécessairement très bon. Il s'agit ensuite de dresser une liste des variables disponibles candidates à décrire le processus de non-réponse. La nature de cette liste varie selon les thèmes abordés par l'enquête, le mode de collecte et le plan d'échantillonnage. Tout facteur susceptible d'expliquer la non-réponse doit être pris en compte dans la mesure du possible. Les variables peuvent êtres des grandeurs mesurées dans l'enquête elle-même ou des informations auxiliaires obtenues par d'autres biais (par exemple en appariant l'échantillon avec des registres) sur les unités répondantes et non-répondantes. Dans les enquêtes répétées on a en principe un choix

de variables beaucoup plus grand pour modéliser la non-réponse dès la deuxième vague car on peut alors se reposer sur des informations récoltées en première vague.

La régression logistique peut s'appuyer sur une procédure de sélection de variables pour obtenir la liste des variables explicatives les plus intéressantes afin de prédire la réponse des unités sélectionnées. Dans le cas de variables catégorielles, le croisement des modalités des variables retenues forme alors de facto des groupes homogènes de réponse.

Segmentation et régression logistique

Le choix entre une modélisation de la non-réponse par segmentation ou par régression logistique n'est pas simple. Dufour et al. (1998) ont montré sur deux études empiriques similaires à SILC que la méthode de segmentation était meilleure que la régression logistique pour créer des GHR. Effectivement, la méthode de segmentation est plus flexible que la régression logistique. La régression logistique force l'arbre de décision à être symétrique : une fois que les variables significatives sont déterminées, les GHR sont formés en prenant toutes les intersections possibles entre ces variables. Le modèle est donc nécessairement symétrique, et de petits GHR éventuellement non pertinents peuvent être formés. Cela augmente inutilement la variance des estimateurs. La modélisation par segmentation utilise quant à elle un processus itératif pour partitionner le fichier de données. Elle détermine à chaque embranchement de l'arbre de décision quelle variable est la plus significative, ce qui garantit que chaque GHR formé est pertinent. Ce procédé conduit habituellement à un modèle asymétrique. Si pour un nœud donné, aucune variable n'est significative pour expliquer la non-réponse, ou si l'un des critères d'arrêt de l'algorithme est rencontré, aucun nouvel embranchement n'est créé. La segmentation présente l'avantage de pouvoir utiliser un plus grand nombre de variables pour modéliser la non-réponse tout en arrivant à un nombre de GHR réduit. Plusieurs méthodes peuvent être employées pour déterminer quelle variable semble le plus influencer la réponse, Nakache & Confais (2003).

Selon LaRoche (2007), un inconvénient de la modélisation par segmentation est directement relié à l'arbre de décision lui-même. L'arbre de décision formé peut devenir très grand, comporter plusieurs embranchements, et se montrer difficile à interpréter. De plus, les arbres de décisions sont aussi assez instables : l'ajout, le retrait ou le changement d'une variable ou d'une unité peut avoir un effet important sur l'arbre lui-même. La stabilité de l'arbre est d'autant plus précaire que le nombre de variables explicatives est grand ou que certaines variables sont fortement corrélées entre elles. Cependant, conformément à Rakotomalala (2005), nous avons pu constater que, sur les enquêtes que nous avons traitées à l'OFS, ces variations dans la structure des arbres n'ont que peu d'influence sur les poids et les estimations finales.

Dans les deux méthodes, on a la possibilité d'imposer des contraintes dans la formation des GHR, typiquement sur le nombre d'observations

et le taux de réponse minimal dans chaque groupe. Ces choix ont une influence directe sur le nombre final de GHR et la taille de l'arbre de décision. Comme rapporté dans Graf (2009a), l'une et l'autre méthode visent à trouver, pour chaque unité répondante, un facteur d'ajustement pour la non-réponse. Si les GHR sont définis de telle sorte que la non-réponse soit complètement uniforme à l'intérieur d'un groupe, alors le biais de (2.5) est négligeable (voir Tambay et al., 1998; Kalton & Kasprzyk, 1986).

Correction de la non-réponse et calage

L'estimation à partir de données d'enquêtes sujettes à de la non-réponse peut aussi être abordée par des méthodes de post-stratification et plus généralement de calage (Deville & Särndal, 1992). Dans certains cas, les résultats sont même strictement identiques. Le calage généralisé introduit dans Deville (2002) et la méthode "lune et étoile" de Särndal & Lundström (2005) (voir section 2.4.3) permettent d'utiliser des informations disponibles seulement sur l'échantillon et d'obtenir la pondération finale en une seule étape. La méthode "lune et étoile" est une méthode élégante, applicable dans le cas d'une pondération simple d'enquête où le nombre d'étapes de correction de la non-réponse n'est pas élevé, et où il n'y a pas de partage de poids ou de combinaison d'échantillons s'intercalant entre les stades de correction de la non-réponse et le calage final. Cette méthode a été appliquée pour pondérer l'Enquête Suisse sur la Santé, Graf (2010).

Dans le cas de l'enquête structurelle, le calage exigé afin d'assurer la cohérence avec les totaux issus de l'exploitation exhaustive des registres est extrêmement complexe. Toutes les variables ou tous les croisements de modalités que l'on aurait envisagés pour tenter de modéliser la non-réponse ou constituer les GHR sont utilisés dans le calage. Il s'agit entre autres des croisements de lieu de résidence, sexe, de classe d'âge, d'état civil, de nationalité et permis de séjour. Dès lors on aurait pu considérer que la non-réponse est entièrement traitée par calage, suivant Särndal & Lundström (2005). Toutefois, on a décidé de ne pas complètement confondre la procédure de calage et la procédure de traitement de la non-réponse, mais simplement de simplifier cette dernière. Les coefficients $\widehat{\theta}_k$ sont estimés dans le cas du relevé structurel par les taux de réponse observés dans chaque canton. Cela correspond à un modèle de non-réponse homogène dans les cantons.

2.4.3 Calage

Les avantages des procédures de calage sont nombreux. Un calage peut permettre d'obtenir des estimateurs plus précis en profitant de la proximité entre une variable observée y_k est d'autres variables x_k dont on connaît les totaux sur la population. Il permet également d'obtenir des estimateurs qui reproduisent fidèlement des statistiques choisies, et ainsi de produire des séries de résultats cohérentes. On parle alors de calage cosmétique. Enfin, il peut être utilisé en complément ou à la place des méthodes présentées dans la section 2.4.2 pour calculer des poids qui prennent en

compte la non-réponse à l'enquête. Le calage peut donc être utilisé pour réduire la variance, sans but statistique précis, ou même pour réduire le biais de non-réponse de l'estimateur (2.4). Dans la pratique, on est toujours confronté à un mélange de ces trois raisons.

Les méthodes de calage utilisées à l'OFS ont été développées et présentées dans l'article fondateur de Deville & Särndal (1992), et dans Deville et al. (1993); Le Guennec & Sautory (2002); Sautory (2003); Deville (2002); Kott (2006). Le principe du calage est, partant d'un jeu de poids d_k, $k \in r$, de trouver un nouveau jeu de poids w_k, le plus proche possible des poids de départ d_k pour une distance spécifiée, et qui vérifie certaines contraintes. Dans la méthode originelle de Deville & Särndal (1992), il s'agit, partant des poids d_k de l'estimateur de Horvitz-Thompson, d'obtenir que

$$\widehat{X}_w = \sum_{k \in r} w_k x_k^*$$

soit égal au total connu sur la population, pour certaines variables x_k^*, $k \in U$ choisis. L'estimateur de Horvitz-Thompson étant sans biais, la faible distance entre les d_k et les w_k assure que l'estimateur calé obtenu est peu biaisé. Le choix de la distance, autorisé dans le logiciel Calmar2 (Le Guennec & Sautory, 2002; Sautory, 2003) utilisé à l'OFS, n'a théoriquement pas une grande importance. Ce choix est effectué de manière à éviter d'obtenir quelques poids négatifs, trop grands ou trop éloignés des poids de départ. On cherche aussi souvent à éviter qu'il y ait des poids inférieurs à 1, ce qui est conceptuellement difficile à envisager, ou même inférieurs aux poids de départ quand la procédure est utilisée pour traiter de la non-réponse. En effet, on ne pourrait alors plus interpréter le rapport d_k/w_k comme une probabilité de réponse estimée.

Calage "lune et étoile"

La méthodologie développée dans Särndal & Lundström (2005) a été appliquée à l'OFS pour pondérer l'Enquête Suisse sur la Santé de 2007 (Graf, 2010). Elle distingue deux types d'information auxiliaire :

1. les variables x_k^* connues sur l'échantillon observé r et dont on connaît le total sur la population. Dans la terminologie de Särndal & Lundström (2005), il s'agit du cas "InfoU",

2. les variables x_k° connues sur l'échantillon sélectionné s, et dont on ne connaît pas, ou pas nécessairement le total sur la population. Il s'agit du cas "InfoS".

Lorsque les deux types d'informations sont disponibles on est dans le cas "InfoUS".

Le calage sur les totaux des variables "étoile" x_k^* correspond au calage classique de Deville & Särndal (1992). L'information apportée par les variables "lune" x_k° peut être utilisée pour adapter les poids à la situation de non-réponse. Le calage "lune" vise à trouver un jeu de poids w_k qui vérifie les contraintes :

$$\sum_{k \in r} w_k x_k^\circ = \sum_{k \in s} d_k x_k^\circ,$$

où $d_k = 1/\pi_k$ est l'inverse de la probabilité de sélection de k dans s. Un exemple tout à fait classique est celui du plan aléatoire simple de taille fixe n dans une population de taille N, avec pour seule variable de calage $x_k^\circ = 1$, et un échantillon observé de m répondants. Partant de $d_k = N/n$, la méthode conduit naturellement à utiliser le poids $w_k = N/m$. Le calage "lune et étoile" a l'avantage de permettre de calculer une pondération qui respecte simultanément les totaux connus au niveau de la population et au niveau de l'échantillon sélectionné.

Le choix des variables "lune" x_k° peut être effectué selon les principes décrits dans la section 2.4.2. En particulier, il peut s'agir d'indicatrices d'appartenance à des groupes homogènes de réponse obtenus par segmentation en utilisant des variables connues sur l'échantillon s, ou encore de variables retenues par une procédure de sélection lors d'une régression logistique des indicateurs de réponse r_k sur les différentes variables disponibles. Les variables "étoile" x_k^* devraient être "proches" des variables d'intérêt y_k. Leur choix peut être assisté par des procédures de sélection de variables adaptées à la régression linéaire, sur l'échantillon des répondants, des y_k sur les différentes variables x_k^* dont on connaît le total sur la population. Dans ce dernier cas, il semble utile d'utiliser une pondération temporaire qui tienne déjà compte de la non-réponse. Dans la pratique, comme dans le cas du relevé structurel, la liste des variables imposées pour le calage cosmétique peut déjà être assez étendue. Les procédures de sélection de variables peuvent alors servir à confirmer qu'aucun facteur significatif n'a été oublié, ou de guider le choix des interactions ou croisements de catégories utilisé. Enfin, il permet de repérer les variables qui ont le moins de justification statistique d'entrer dans le calage et d'en motiver l'élimination lorsque les demandes formulées par les unités de diffusion semblent vraiment excessives.

Calage à deux niveaux

Dans les enquêtes auprès de la population, des résultats sont souvent produits au niveau des personnes et au niveau des ménages. Le calage simultané (Le Guennec & Sautory, 2002) permet de réaliser un calage à la fois sur des informations disponibles au niveau des personnes et sur des informations disponibles au niveau des ménages. En particulier, on peut exiger que les poids des personnes soient égaux au sein d'un même ménage. Des approches plus basiques ont été utilisées pour l'enquête SILC et pour le relevé structurel.

L'enquête SILC est une enquête ménage pour laquelle est produite à la fois une pondération ménages et une pondération personnes. Au moment de sa pondération, on ne disposait pas d'information précise sur la population des ménages, mais seulement sur la population des personnes.

Le relevé structurel comporte un questionnaire à remplir par tous les membres du ménage. Les répondants doivent en particulier fournir leur numéro de sécurité sociale (AVS), qui sert d'identifiant dans le cadre d'échantillonnage SRPH, indiquer leur langue principale et niveau de formation, et leurs liens familiaux avec les autres membres du ménage. Une

pondération a donc été calculée pour pouvoir extrapoler ces informations au niveau des ménages et non plus des personnes. Cela a été rendu difficile par le fait que la formation des ménages dans le fichier SRPH est imparfaite, et que l'on n'a donc pas de totaux de calage à ce niveau-là.

Pour ces deux enquêtes, un calage a été réalisé au niveau des ménages sur des totaux connus seulement au niveau des personnes. Les caractéristiques des personnes ont dans un premier temps été sommées dans chaque ménage m pour constituer les variables de calage x^*_m.

2.4.4 Poids partagés et poids combinés

Poids partagés

La méthode du partage des poids (Lavallée, 2002) permet de pondérer un échantillon obtenu de manière indirecte. Elle nécessite l'existence et le comptage de liens entre la population sélectionnée et la population observée. Dans l'enquête SILC, les individus enquêtés n'ont pas tous été directement sélectionnés. En effet, on fait la distinction entre les individus présent dans les ménages sélectionnés lors de leur première vague d'observation, ce sont les individus longitudinaux ou OSM (Original Sample Member), et les individus, aussi enquêtés, qui font partie de leurs ménages seulement aux vagues d'enquête suivantes. Ces individus, qui se joignent à l'étude après la vague 1, n'ont été sélectionnés que de manière indirecte puisqu'ils ne faisaient pas partie de l'échantillon effectivement tiré pour la vague 1. A la différence des OSM, qui possèdent un poids résultant de leur probabilité de sélection et de réponse, les nouveaux individus, appelés cohabitants, ne possèdent pas de poids a priori.

La méthode du partage des poids permet d'attribuer un poids aux nouveaux cohabitants en fonction des poids des OSM présents dans le ménage. Le lien entre la population sélectionnée et la population observée est la présence d'OSM dans les ménages observés. Plusieurs partages de poids ont été réalisés pour obtenir des pondérations individuelles et des pondérations ménages (Graf, 2008a, 2009a). La pondération ménage est caractérisée par le fait que tous les individus d'un même ménage ont un poids final identique. Pour la vague 2 ou plus, Lavallée (2002) montre que les poids calculés par la méthode du partage des poids permettent d'estimer sans biais des totaux et propose un estimateur de leur variance qui tient compte du partage. Massiani (2013a) relève que cet estimateur de variance n'est généralement pas sans biais, même lorsque les probabilités de réponse sont connues, et apporte une correction qui annule ce biais.

Poids combinés

Pour produire des estimations à partir de plusieurs enquêtes, on peut calculer des combinaisons convexes des estimations obtenues aux différentes enquêtes. Cela revient à assembler les échantillons en multipliant les poids individuels par les coefficients de la combinaison convexe.

Plusieurs enquêtes de l'OFS ont des échantillons constitués de blocs de rotation ou vagues d'enquête. Il s'agit des enquêtes thématiques, de

l'Espa, de SILC, de l'EBM. Pour les enquêtes thématiques, seule une estimation valable pour l'année d'enquête est produite. La pondération de ces enquêtes est réalisée sans modification spécifique liée à cette organisation de collecte en vagues : les probabilités d'inclusion à l'échantillon complet sont calculées puis les probabilités de réponse, etc.

L'Espa produit des résultats trimestriels, basés sur quatre blocs de rotation, et des résultats annuels, basés sur une partie des seize blocs d'enquête d'une année donnée. La pondération pour les résultats trimestriels est elle aussi calculée en évaluant la probabilité d'inclusion de chaque personne à l'union des quatre blocs correspondants, sa probabilité de réponse et en effectuant un calage.

Les informations utiles pour les estimations annuelles ne sont fournies que par une partie des blocs d'enquête : ceux qui sont en première et en troisième vague d'interrogation. En effet, selon le schéma de rotation décrit dans Renfer (2009), les individus enquêtés une année donnée en vague 2 et en vague 4 sont les mêmes que ceux enquêtés en vague 1 et en vague 3 cette même année, à l'exception du bloc en vague 4 au premier trimestre. Il y a ainsi une très forte corrélation entre les valeurs relevées en vagues 1 et 3 et celles relevées en vagues 2 et 4, ce qui peut justifier d'écarter ces dernières de l'estimation annuelle. Seuls huit blocs d'enquête constituent donc l'échantillon de l'Espa pour les estimations annuelles. Sa pondération est réalisée en deux temps :

1. une nouvelle pondération trimestrielle est réalisée en ne gardant que les blocs en vagues 1 et 3, sur le même principe que celle calculée pour produire les estimations trimestrielles de l'Espa,

2. les échantillons des quatre trimestres sont assemblés en divisant les poids des individus par 4. Cela revient à imposer que les estimations annuelles sont les moyennes simples d'estimations trimestrielles basées sur une partie des données.

L'enquête SILC fait l'objet de plusieurs pondérations : on calcule une pondération pour le panel qui a été enquêté quatre fois, une pondération pour les deux panels qui ont été enquêtés les trois dernières années, une pondération pour les trois panels qui ont été enquêtés les deux dernières années et une pondération pour les quatre panels enquêtés l'année courante (voir figure 2.6).

La méthode utilisée pour combiner les données des différents panels a été développée pour deux panels par (voir Merkouris, 2001; Lévesque & Franklin, 2000). Elle consiste à attribuer un facteur p_i à chacun des panels, avec $\sum_i p_i = 1$. Par exemple, l'estimateur transversal pour la dernière année est de la forme

$$\widehat{Y} = \sum_{i=1}^{4} p_i \widehat{Y}_i,$$

où $\widehat{Y}_i$ est l'estimation du total de la variable d'intérêt y_k obtenue en pondérant le panel i. La valeur optimale des coefficients p_i est calculée en minimisant la variance de $\widehat{Y}$. En pratique, un seul jeu de facteur est

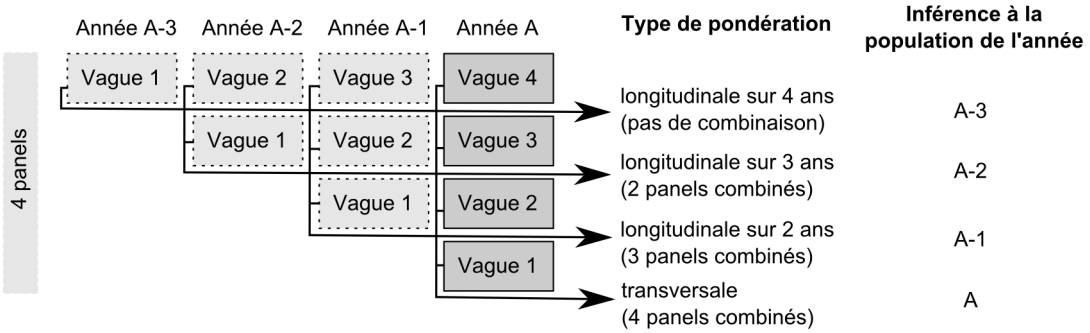


Figure 2.6 – *Plan rotatif de l'enquête SILC.*

utilisé pour plusieurs variables d'intérêt y, et l'optimisation n'est faite que de manière approchée. Le facteur p_i est en pratique défini par

$$p_i = \frac{n_i d_i}{\sum_{i=1}^{4} n_i d_i}, \; i = 1, ..., 4$$

où d_i désigne l'effet de plan d'échantillonnage pour le panel i et n_i représente le nombre de personnes, ou de ménages, du panel i répondantes à la vague d'enquête considérée pour l'année en cours. Il s'agit de la vague 1 pour le nouvel échantillon entrant, de la vague 2 pour l'échantillon interrogé pour la deuxième année consécutive, de la vague 3 pour celui interrogé pour la troisième année consécutive et de la vague 4 pour l'échantillon interrogé pour la quatrième et dernière année consécutive. Comme les plans de sondage des différents panels sont identiques, on fait l'hypothèse que les effets de plan sont sensiblement les mêmes, et les p_i sont in fine proportionnels aux tailles d'échantillons.

2.5 Traitement des données et imputation

Les données d'enquête présentent quasiment systématiquement des défauts : les valeurs sont manquantes pour certaines variables et certaines unités, ou bien elles sont incohérentes, atypiques, suspectes. Le problème peut venir d'une mauvaise compréhension du questionnaire, d'une divergence de concepts, d'une incapacité à répondre à certaines questions, ou encore d'un refus de répondre. Il peut aussi être lié aux traitements effectués lors de la production du fichier de données à partir du résultat de la collecte. On peut citer par exemple les erreurs lors de la numérisation de questionnaires papiers, et les erreurs de traitement informatique. Ces défauts doivent être traités, dans la mesure du possible, et systématiquement documentés pour que les utilisateurs finaux soient informés des possibles lacunes de l'enquête, et que des indicateurs de qualité puissent être calculés. Cela fait l'objet du processus de préparation statistique des données.

2.5.1 Processus de préparation statistique des données

L'OFS suit les recommandations formulées dans le cadre des projets Euredit (European Union, 2003) et Edimbus (European Union, 2007) sur le traitement des données d'enquête et les imputations. Ces recommandations font l'objet d'une formation interne dispensée régulièrement aux collaborateurs de l'OFS (Graf & Kilchmann, 2010). Dans un schéma d'enquête, l'édition des données, l'imputation et la pondération participent au processus de préparation statistique des données (PPSD, voir Figure 2.7). Le PPSD a pour objet la description des données récoltées,

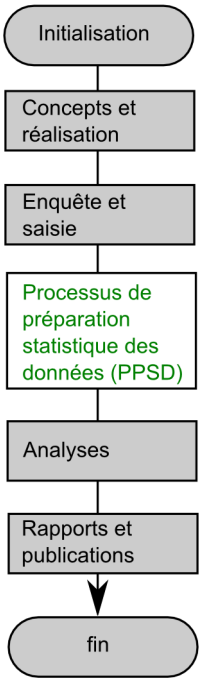


Figure 2.7 – *Schéma d'enquête.*

l'amélioration de leur qualité et de leur cohérence, et la préparation des données pour les analyses. À chacun de ces trois objets correspond une phase du PPSD, voir Figure 2.8. Chaque phase est constituée de plusieurs procédures, le tout pouvant être répété jusqu'à ce que les objectifs de qualité soient atteints. Les données initiales sont toujours conservées sans modifications, et des indicatrices sont créées pour repérer quelles valeurs des variables d'étude ont été générées par le PPSD.

La phase de préparation initiale des données permet de décrire la qualité des données, en créant des indicatrices de réponse pour chaque variable, des indicatrices de conformité à des règles devant être respectées par les valeurs observées. On peut aussi y trouver des procédures élémentaires d'imputation ou de modification automatique des données.

La phase de micro-préparation des données traite les observations au niveau de l'unité d'enquête, personne ou ménage. Elle s'attache à assurer la cohérence et la complétude des données pour chaque unité. Les traitements peuvent être automatisés ou bien interactifs, avec par exemple des comparaisons avec d'autres sources d'information, un retour vers l'unité

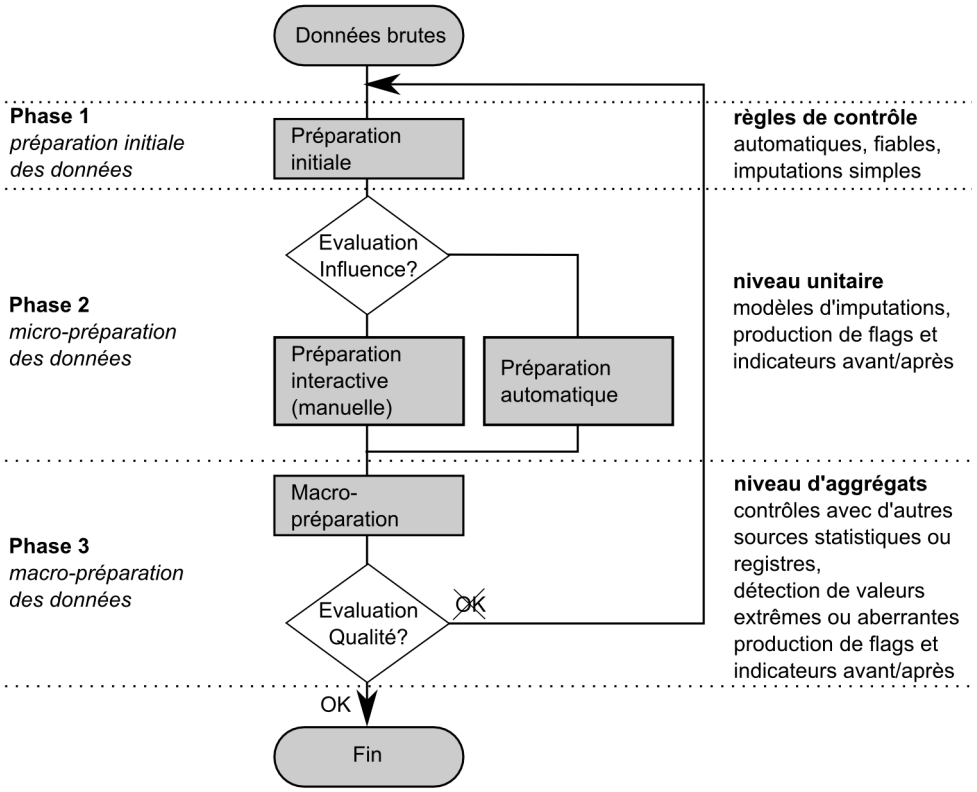


Figure 2.8 – *Processus de préparation statistique des données.*

enquêtée pour vérification des données, etc. Les traitements automatisés regroupent entre autre l'emploi de méthodes d'imputation. Les opérations réalisées dans cette phase ne doivent pas entrer en conflit avec les contrôles de qualité réalisés dans la phase de préparation initiale des données.

La phase de macro-préparation des données permet d'évaluer les données collectées en les comparant avec d'autres statistiques ou sources de données, et de repérer des données aberrantes en comparant chaque donnée individuelle avec le reste de l'échantillon. Elle peut déboucher sur la création de nouvelles règles intégrées aux deux phases précédentes. La qualité des données est évaluée à la fin de cette phase, et celles-ci sont resoumises au PPSD si la qualité n'est pas suffisante.

2.5.2 Non-réponse

Un défaut commun aux données d'enquêtes est l'absence de réponse à certaines questions par certaines unités enquêtées. On parle de non-réponse partielle lorsqu'un questionnaire est en partie rempli mais ne l'est pas complètement. Il est possible de laisser ces observations en l'état, mais cela conduira à produire des tableaux de résultats avec des catégories "réponse manquante". Typiquement le lecteur, déjà relativement averti, fera alors mentalement une règle de trois pour répartir cette catégorie manquante dans les catégories renseignées. On comprend donc que cela revient à

laisser au lecteur le choix d'un modèle d'imputation ou de repondération pour arriver à des résultats sans valeurs manquantes. On préfère en général, lorsqu'on en a les moyens, prendre en charge cette modélisation à l'OFS et traiter les observations avec non-réponse partielle. Deux sortes de traitement sont possibles :

1. imputer les valeurs manquantes, c'est à dire prédire les réponses qui n'ont pas été observées. Il s'agit d'un processus complexe, avec une part d'arbitraire, qui demande une grande expertise, et dont l'influence sur les résultats et leur précision est difficile à évaluer. Cette approche, discutée dans la section 2.5.4, est cependant souvent indispensable pour pouvoir exploiter les données d'enquête, car le traitement alternatif (voir ci-dessous) conduirait à perdre une trop grande part de l'échantillon.

2. éliminer les questionnaires présentant de la non-réponse partielle, et pondérer le reste de l'échantillon en conséquence. Cette approche est généralement utilisée pour les questionnaires qui présentent beaucoup de non-réponse et peu d'information utile, ou bien lorsqu'une imputation semble irréalisable. Cela a par exemple été effectué massivement dans le cadre de l'enquête structurelle lors de l'exploitation des questionnaires ménages, et du calcul d'une pondération d'extrapolation pour les ménages.

   En effet, les ménages observés à l'enquête structurelle sont tendanciellement plus petits que ceux attendus d'après le fichier SRPH. Il y a en particulier un déficit assez important de cohabitants âgés entre 20 et 40 ans. Cela peut être rapproché de la pratique courante pour les jeunes adultes et étudiants de rester enregistrés au domicile des parents plutôt que de s'inscrire dans leur commune de résidence effective. Or, si notre connaissance des ménages n'est pas parfaite, on estime qu'elle est tout de même de très bonne qualité. En particulier, le nombre de ménages dans le cadre d'échantillonnage SRPH est très proche de toutes les estimations qui avaient pu en être faites jusque-là. On pense donc que l'on est confronté à un problème de non-réponse. La non-réponse (totale) d'un individu membre du ménage est assimilable à une non-réponse partielle pour l'unité ménage concernée, dans la mesure où il manque toutes les informations de l'individu non-répondant sur le questionnaire ménage. Cette non-réponse peut être liée à un désir de ne pas répondre ou à une mauvaise compréhension de la notion de ménage décrite sur le questionnaire.

   Procéder à des imputations pour combler ces non-réponses (que l'on ne sait de plus pas identifier avec certitude puisque le fichier SRPH n'est pas absolument juste) semble hors de portée. Il faudrait en effet imputer toutes les relations familiales en respectant leur cohérence. On a donc choisi de traiter cette non-réponse par élimination et repondération. Tous les ménages dont la composition déclarée est différente de la composition relevée dans le fichier SRPH ont été écartés de l'échantillon de ménages. Ils représentent environ 20%

de l'échantillon. De fait tous les ménages dont la composition a réellement changé entre la constitution du cadre d'échantillonnage et l'enquête sont exclus. Toutefois, le nombre de ces ménages qui ont évolué est très faible devant le nombre total de ménages écartés.

Que ce soit pour procéder par imputation ou par élimination, il est nécessaire d'avoir une assez bonne idée d'un mécanisme qui pourrait expliquer la non-réponse survenue. Pour ce faire, on considère la variable indicatrice de réponse, résultat du contrôle de l'existence d'une réponse : $r_{kj} = 1$, si la variable j de l'observation k est renseignée, $r_{kj} = 0$ sinon. Le taux de réponse pour une variable j est alors la moyenne $\bar{r}_j = n^{-1} \sum_{k \in s} r_{kj}$, où n est la taille de l'échantillon s. Le cas échéant, il faut distinguer les valeurs qui sont structurellement manquantes, par exemple parce qu'une unité est dispensée de réponse à une partie du questionnaire, des valeurs qui auraient dû être observées mais ne sont pas renseignées. Le vecteur colonne $\mathbf{r}_j = (r_{1j}, r_{2j}, ..., r_{nj})'$ est l'indicatrice de réponse d'une variable j. La matrice des r_{kj} est générée par un processus de réponse inconnu. La modélisation de ce processus de réponse détermine les méthodes d'imputations et de repondérations utilisées.

2.5.3 Valeurs erronées et valeurs aberrantes

Un autre défaut fréquent des données d'enquête est la présence de valeurs manifestement fausses car en dehors des domaines des variables, possiblement fausses, incohérentes avec d'autres valeurs relevées, ou bien simplement atypiques car très éloignées du reste des valeurs observées. Ces observations peuvent provenir d'erreurs systémiques, par exemple d'erreurs de formulation dans le questionnaire, d'erreurs de numérisation, d'erreurs de programmation. Elles peuvent être liées à une incompréhension sur l'unité des valeurs demandées et plus généralement à une incompréhension du questionnaire. Elles peuvent ne répondre à aucun mécanisme évident. Enfin, elles peuvent être tout à fait correctes lorsqu'il s'agit de valeurs atypiques, mais leur présence implique une forte variabilité des estimateurs pour les variables concernées.

La détection de ces valeurs est effectuée au moyen de règles de contrôle. Fellegi & Holt (1976) ont montré que celles-ci peuvent être décomposées en règles simples implémentables dans un programme qui permet de tester les données. Ces règles peuvent être de diverses natures :

1. il y a les règles absolues qui permettent de repérer une valeur anormale sans considération pour les autres variables de la même unité enquêtée ou des autres unités. Par exemple, une année de naissance trop éloignée, un loyer négatif, peuvent-être détectés de cette manière.

2. Il y a aussi les règles de cohérence internes au questionnaire, par exemple les règles qui vérifient que la somme des composants déclarés d'un total déclaré est bien égale à ce total, ou bien les implications logiques entre variables déclarées. Lorsque des incohérences impliquent plusieurs variables sans que l'on sache déter-

miner lesquelles sont correctes et lesquelles sont erronées, le principe de parcimonie veut, en accord avec Fellegi & Holt (1976), que l'on essaye de traiter le plus petit nombre de variables pour rendre les observations cohérentes.

3. Enfin, il y a les règles qui permettent de détecter les valeurs influentes ou atypiques sans qu'il n'y ait nécessairement une incohérence ou une erreur. Celles-ci peuvent être une simple comparaison entre la valeur observée pour une unité et les valeurs observées sur le reste de l'échantillon. Un outil utile est alors le MAD (Median Absolute Deviation), défini comme la médiane des écarts absolus à la médiane de l'échantillon,

$$\mathrm{MAD}_y = \mathrm{med}_{k \in s}(|y_k - m_y|),$$

où $m_y = \mathrm{med}_{k \in s}(y_k)$ est la médiane des valeurs y_k observées. Cette mesure a l'avantage de ne pas être perturbée par un faible nombre de valeurs extrêmes. Les observations telles que l'écart $|y_k - m_y|$ excède une certaine fois MAD_y sont considérées comme extrêmes ou aberrantes. Hulliger (1999) propose de multiplier ces observations par un facteur $u_k < 1$ fonction de $\mathrm{MAD}_y / |y_k - m_y|$. D'autres règles reposent implicitement sur une modélisation des relations entre les variables observées. On calcule fréquemment certains quotients y_{kj}/y_{km} entre variables relevées, avec l'idée que ces variables sont quasiment proportionnelles et que le quotient doit rester dans une plage de valeurs restreintes. Dans la cas de panels ou d'enquêtes répétées, l'évolution relative y_t/y_{t-1}, où y_t et y_{t-1} sont les valeurs pour l'enquête courante et la précédente est aussi étudiée (Hidiroglou & Berthelot, 1986). Plus généralement, on peut régresser certaines variables observées sur d'autres variables de l'enquête, de préférence de manière robuste, puis comparer l'écart entre la valeur observée et la valeur prédite à une mesure de dispersion elle aussi, si possible, robuste.

Dans les enquêtes de l'OFS auprès de la population et des ménages, les deux premiers types de règles de contrôle sont toujours employés. Le troisième type de règle peut aussi être utilisé lorsque le questionnaire porte sur les salaires, loyers et dépenses. Le non-respect d'une ou plusieurs règles conduit à l'élimination des données de l'unité concernée dans les cas les plus sévères, et généralement à l'imputation d'une valeur admissible dans les autres cas.

2.5.4 Imputations

L'OFS utilise plusieurs stratégies pour imputer de nouvelles valeurs à la place des valeurs manquantes, et de certaines valeurs aberrantes. En général, plusieurs stratégies sont utilisées pour chaque enquête. Les méthodes les plus complexes ne sont, par manque de temps ou d'expertise, pas systématiquement utilisées. De plus, la plupart du temps les imputations ne peuvent pas tenir compte de toutes les règles de contrôle. Un contrôle après imputation doit donc être mené. La méthode de Fellegi & Holt

(1976) permet de tenir compte des règles de contrôle pour des variables qualitatives, mais cela nécessite qu'elles soient définies de manière stricte et complète, ce qui n'est pas toujours le cas.

Les stratégies d'imputation utilisées comprennent :

1. les imputations interactives, dont la reprise de résultats de rappels téléphoniques et les traitements manuels assistés par ordinateur. Ces opérations nécessitent une intervention humaine pour chaque observation. Leur volume est donc très limité. De plus, les imputations manuelles requièrent une grande expertise et ont le défaut de ne pas être reproductibles.

2. Les imputations automatiques basées sur des règles. Ce sont les résultats de déduction logique des valeurs d'imputations à l'aide de sources internes à l'enquête, comme les valeurs relevées d'autres variables, ou de sources externes, par exemple un registre qui serait apparié à l'échantillon. Ces imputations déterministes peuvent être implémentées de façon automatique. On leur accorde souvent un grand niveau de confiance, mais il faut prendre garde au fait qu'il y a parfois plusieurs valeurs imputées possibles, par exemple lorsque l'on a plusieurs sources d'information externes. Et dans les cas de reprise d'information externe, on prend le risque de dégrader les corrélations entre variables d'enquête.

3. Les imputations assistées par un modèle. On peut citer l'imputation par la moyenne, par la moyenne dans un groupe, par le quotient, par une prédiction de régression. Ces imputations reposent sur une modélisation des relations entre les variables observées et les variables inobservées. La modélisation est guidée par le type de non-réponse supposé. On peut distinguer trois classes de processus de réponse (Little & Rubin, 2002), les processus MCAR (Missing Completely at Random), MAR (Missing At Random) et NMAR (Not Missing At Random). La distinction entre ces trois classes réside dans la relation entre le processus de réponse et les valeurs observées et inobservées des variables y_j. Si la loi du processus ne dépend pas du tout de ces valeurs, on dit que le processus est MCAR. Si elle dépend de valeurs observées, mais pas des valeurs inobservées, on dit que le processus est MAR. Enfin, si elle dépend des valeurs observées et inobservées, on dit que le processus est NMAR. Ce dernier cas est typique des variables portant sur les revenus. En effet, la propension a répondre peut dépendre du montant du revenu, sans que cette dépendance puisse être contrôlée par l'utilisation des autres variables à disposition. Les paramètres du modèle sont estimés sur le sous-échantillon de répondants pour lesquels les questionnaires sont complets. La procédure d'imputation a un impact sur la distribution produite des variables d'enquête. Un problème classique de l'imputation par la moyenne ou par la prédiction est qu'elle tend à réduire la dispersion de la variable d'enquête. Ceci est parfois compensé par l'ajout aux valeurs prédites d'une erreur générée

aléatoirement. La cohérence entre toutes les variables d'enquête est difficile à maintenir.

4. Les imputations par donneur. Les valeurs déclarées par une ou plusieurs autres unités enquêtées sont utilisées pour l'imputation. Ces donneurs sont choisis à l'aide d'une distance définie entre les répondants. On peut citer par exemple la méthode des plus proches voisins (Sande, 1981). L'avantage de ce genre de méthode est que, grâce à un même donneur, il est possible de remplacer simultanément plusieurs valeurs manquantes ou erronées. On préserve ainsi la cohérence entre les valeurs imputées. La difficulté réside dans le choix de la fonction de distance entre les observations.

Elles sont en général utilisées à l'OFS dans cet ordre chronologique lors du traitement des données d'enquête.

L'expérience d'imputation la plus complexe à l'OFS a été menée dans le cadre de l'enquête SILC. On a essayé d'y utiliser une méthode d'imputation multiple assistée par des modèles de régression. Plusieurs variables ont été imputées à l'aide du programme IVEware (Raghunathan et al., 2001). IVEware permet d'effectuer des imputations individuelles ou multiples au moyen d'une méthode de régression séquentielle. Le type de régression utilisé est adapté à la nature des variables à imputer : quantitative, dichotomique, catégorielle, ou bien variable de comptage. Les variables sont traitées par proportion croissante de valeurs manquantes. Soit $\mathbf{X}$ la matrice regroupant le sous-ensemble des variables explicatives sans valeurs manquantes. Soient aussi $y_1, y_2, ..., y_p$ les p variable à imputer, où y_1 présente le moins de valeurs à imputer et y_p en présente le plus. Ces variables peuvent être de nature différentes. La fonction de densité commune de $y_1, y_2, ..., y_p$ sachant $\mathbf{X}$ peut être factorisée comme suit (Raghunathan et al., 2001) :

$$f(Y_1, Y_2, ..., Y_p|\mathbf{X}) = f_1(Y_1|\mathbf{X})f_2(Y_2|\mathbf{X}, Y_1)...f_p(Y_p|\mathbf{X}, Y_1, Y_2, ..., Y_{p-1})$$

où les f_j, $j = 1, 2, ..., p$ sont des fonctions de densité conditionnelles. On modélise chaque densité conditionnelle à l'aide d'un modèle de régression paramétrique approprié. Les densités conditionnelles sont alors estimées séquentiellement.

Le premier cycle d'une imputation est effectué comme suit : la densité conditionnelle f_1 est estimée à partir des observations pour lesquelles on dispose des valeurs de y_1. Des valeurs imputées sont alors générées selon cette loi conditionnelle pour compléter la variable y_1. La loi conditionnelle f_2 est estimée en utilisant à partir des observations pour lesquelles on dispose des valeurs de y_2, et en utilisant les valeurs imputées pour y_1 à l'étape précédente. Des valeurs imputées pour y_2 sont alors générées. Le processus est poursuivi jusqu'à disposer de valeurs imputées pour toutes les variables.

Le deuxième cycle d'une imputation est réalisé en estimant la densité de la variable y_1 conditionnellement à $\mathbf{X}$ et aux $p - 1$ autres variables. De nouvelles valeurs imputées sont générées selon cette densité estimée et

remplacent celles générées au premier cycle. Il est fait de même successivement pour les $p-1$ autres variables à imputer. Le nombre de tels cycles effectués peut être choisi, et on arrête le processus lorsque les résultats ne changent plus vraiment.

L'ensemble de la procédure, première imputation et cycles successifs compris, peut être répété m fois pour obtenir m jeux de valeurs imputées différents pour chacune des p variables. L'existence de plusieurs jeux de valeurs imputées permet alors d'avoir une idée de la distribution des valeurs imputées, et en particulier d'estimer simplement la variance due aux imputations. C'est là, avec la simplicité d'utilisation, l'un des avantages certains du logiciel IVEware. D'autres aspects peuvent être vus comme des désavantages : le tout fonctionne comme une boite noire, avec des cycles, des répétitions, on ne saurait calculer explicitement une densité même en connaissant les densités réelles des observations en entrée. Le code source du logiciel n'est pas disponible. On ne peut spécifier de pondérations. De plus, les régressions employées sont sensibles aux valeurs extrêmes, et mal adaptées aux variables monétaires. Enfin, ce qui est pour l'estimation de variance un avantage, la production de plusieurs valeurs imputées, est pour l'estimation directe source d'insatisfaction. En effet, il a été constaté que la variabilité des valeurs imputées était trop grande pour maintenir une cohérence acceptable avec d'autres informations du fichier de données si l'on choisit au hasard une des valeurs imputées. Actuellement, la valeur imputée correspond à la moyenne des valeurs obtenues sur 50 imputations. Un système d'imputation qui ne présente pas ces inconvénients est à l'étude (Graf & Tillé, 2012).

2.6 Estimation de variance

La variance des estimateurs utilisés dépend du plan de sondage, de la non-réponse et de son traitement, des éventuels partages ou combinaisons de poids, et des calages et imputations effectués. En toute rigueur, une estimation de cette variance devrait aussi prendre en compte tous ces éléments. En premier lieu, il faudrait pouvoir disposer d'une méthode de calcul spécifique à chaque plan de sondage. Or aucun outil n'est disponible à l'heure actuelle qui soit capable de répondre à ces besoins. Il est même douteux qu'il puisse en exister un tant les traitements peuvent être variés, et parfois arbitraires. La technique de Bootstrap (Efron, 1979; Davison & Hinkley, 1997) peut sembler être une manière de contourner ces difficultés. Mais elle demande également des développements spécifiques à chaque situation (Chauvet, 2007; Antal & Tillé, 2011).

2.6.1 Pour le relevé structurel et les autres enquêtes sélectionnées de manière coordonnée dans le fichier SRPH

Les méthodes d'estimation de variance s'appuient le plus possible sur les capacités offertes par les procédures spécifiques aux données d'enquêtes du logiciel SAS utilisé à l'OFS. Ces procédures n'offrent pas la souplesse nécessaire pour prendre en compte tous les aspects des plans

d'échantillonnage et des méthodes de pondération effectivement employées. En tout état de cause, ces plans d'échantillonnage excluent la possibilité de sélectionner simultanément deux personnes dont on sait qu'elles sont dans le même ménage. Il n'est donc pas possible d'estimer sans biais la variance des estimateurs de totaux (Särndal et al., 1992).

Toutefois, pour l'ensemble des enquêtes appartenant au nouveau système de recensement, les plans de sondages sont relativement proches de plans simples stratifiés par cantons, conditionnellement à la taille des échantillons sélectionnés dans les cantons. En modélisant le processus de non-réponse par un processus bernoullien de sélection des répondants dans les cantons, et en conditionnant par le nombre de répondants, on obtient encore un plan de sondage pour les répondants proche d'un plan simple stratifié. Cette approximation justifie l'utilisation d'estimateurs de variance classiques pour les plans stratifiés.

Les effets du calage réalisé pour pondérer les données sont pris en compte en utilisant une technique de linéarisation (Deville, 1999a). En pratique, une macro procédure SAS a été programmée pour permettre un calcul facile d'estimateurs de variance pour des estimateurs de totaux et de quotients. Cette procédure calcule dans un premier temps les résidus de régression des variables d'intérêts sur les variables de calages de l'enquête, puis elle fait appel aux procédures standard de SAS pour obtenir les estimations de variances. Les sorties sont ensuite mises en forme pour ressembler aux sorties standard des procédures SAS.

2.6.2 Pour l'enquête SILC

Lors du lancement du projet SILC en Suisse, trois options pour estimer les variances des estimateurs utilisés ont été envisagées : la technique de Bootstrap, le jackknife (Quenouille, 1949) et les techniques de linéarisation (Deville, 1999a). Des poids bootstrap ont été produits avec l'aide de Statistique Canada (Cauchon, 2006) utilisés avec la macro BOOTVAR (StatCan, 2010). L'OFS dispose de macros SAS utilisant la méthode jackknife pour estimer la variance de certaines statistiques (Canty & Davison, 1998).

Des comparaisons entre les variances obtenues sur la base de 1'000 jeux de poids bootstrap, par jackknife, ainsi que par linéarisation ont été menées pour les enquêtes pilotes de SILC ainsi que pour des vagues spécifiques du Panel Suisse de Ménages. Les estimations de variance obtenues pour les statistiques calculées, totaux, ratios, différences entre ratios, percentiles, régressions linéaires et logistique, tests du Chi-deux, sont assez proches les unes des autres pour les trois méthodes considérées. Ces essais nous ont conduit à retenir en production la méthode de linéarisation couplée à l'utilisation des procédures de SAS spécifiques pour le traitement des données d'enquête.

Récemment, une procédure qui tient compte de toutes les étapes de la construction de poids pour l'enquête SILC a été détaillée et programmée (Massiani, 2013a). Il y est tenu compte du plan d'échantillonnage, de la modélisation de la non-réponse, du partage des poids, de la combinaison des panels et du calage, par la technique de linéarisation. Des macros SAS

adaptées à SILC permettant des estimations de variances et tenant compte des spécificités de l'enquête ont été produites.

Des statistiques complexes, non linéaires, sont produites à partir de l'enquête SILC. Il s'agit des indices de pauvreté et d'exclusion sociale. Plusieurs grands projets européens se sont concentrés sur l'étude des indices de pauvretés estimés à partir d'enquêtes par échantillonnage. On peut citer Dacseis (European Union, 2004), KEI (European Union, 2008), Sample (European Union, 2011b) et Ameli (European Union, 2011a). Eurostat a également produit les rapports Eurostat (2005a, 2004c,a) et Eurostat (2013). Les programmes de linéarisation développés par Osier (2009) et basées sur Deville (1999a) et Demnati & Rao (2004a) sont utilisées dans Massiani (2013a) pour estimer les variances de ces estimateurs.

2.7 Anticipations

2.7.1 Sur le traitement de la non-réponse NMAR : le calage généralisé

Le calage généralisé de Deville (2002), voir aussi Le Guennec & Sautory (2002); Sautory (2003); Kott (2006), autorise l'utilisation de variables connues uniquement sur l'échantillon des répondants, en plus des variables auxiliaires disponibles pour le calage classique. Tout comme ce dernier, il permet de calculer un jeu de poids w_k, proche des poids initiaux, qui satisfait des contraintes de calage

$$\sum_{k \in r} w_k x_k = X,$$

où les totaux X sur la population des variables auxiliaires x_k sont connus. La différence entre le calage généralisé et le calage classique réside dans la manière d'obtenir ce nouveau jeu de poids. Les fonctions utilisées dans le calage généralisé dépendent des variables x_k et de variables z_k que l'on peut librement choisir, y compris parmi les variables qui ne sont connues que sur l'échantillon de répondants $r \subseteq s$. Ces variables z_k sont appelées instruments de calage, car cette méthode présente une analogie certaine avec la régression instrumentale (Theil, 1953). En particulier, la variance de l'estimateur pondéré du total obtenu peut être estimée par une technique de linéarisation. Elle demande de calculer les résidus de la régression instrumentale des variables d'intérêt y_k sur les variables de calage x_k avec les instruments z_k.

Cette méthode permet de traiter un problème de non-réponse NMAR lorsque certaines variables connues uniquement sur l'échantillon observé sont déterminantes de la propension à répondre (Deville, 2002; Graf & Tillé, 2012; Osier, 2013). Ces variables doivent alors être utilisées comme instruments de calage. L'utilité de la méthode du calage généralisé est conditionnée par un choix pertinent des variables z_k. De même que les techniques de sélection de modèles pour la régression linéaire peuvent être utilisées pour choisir les variables de calage dans le cas d'un calage classique, les résultats théoriques sur la régression instrumentale devraient pouvoir être transposés au calage généralisé. Nous étudions la

possibilité de mettre au point l'analogue de tests de Durbin-Wu-Hausman (voir Durbin, 1954; Wu, 1973; Hausman, 1978; Ruud, 1984) dans le cadre d'enquêtes par échantillonnage afin de disposer d'un outil pour identifier les bonnes variables instrumentales.

2.7.2 Sur l'estimation de la variance due à l'imputation

Comme mentionné dans la section 2.5.4, l'OFS utilise actuellement le logiciel IVEware pour procéder à des imputations multiples dans l'enquête SILC dans les cas de non-réponse partielle. Ce logiciel utilise des modèles de régression mal adaptés aux variables de revenu. Nous développons une méthode d'imputation paramétrique reposant sur l'ajustement d'une loi Bêta Généralisée de Seconde Espèce GB2 (McDonald, 1984; European Union, 2011a). En effet, plusieurs études empiriques (voir par exemple Kleiber & Kotz, 2003; Jenkins, 2007; Dastrup et al., 2007; Sepanski & Kong, 2008; Jones et al., 2011; McDonald et al., 2013) montrent qu'une distribution GB2 s'ajuste bien à de telles données et qu'elle est souvent plus adaptée que d'autres distributions à quatre paramètres. Cet ajustement est réalisé en s'appuyant sur des poids obtenus par calage généralisé qui corrigent au mieux la non-réponse NMAR dont est affectée la variable d'intérêt, voir Graf & Tillé (2012).

2.7.3 Sur la sélection des échantillons pour les enquêtes de l'OFS dans un avenir proche

Au cours de l'année 2014, les dernières enquêtes dont les échantillons sont tirés dans le cadre Castem, à savoir l'enquête SILC, l'Espa et l'EBM seront intégrées au système de coordination d'enquêtes et sélectionnées dans le fichier SRPH. Cela permettra de coordonner proprement ces enquêtes avec les enquêtes du système de recensement, et de ne plus avoir recours à la méthode "d'historisation". SILC et l'EBM sont des enquêtes qui touchent toutes les personnes des ménages sélectionnés. Or le système actuel ne permet pas de sélectionner directement et systématiquement toutes les personnes d'un même ménage dans les échantillons. Nous avons donc réfléchi à la meilleure manière d'organiser ces tirages.

Il est en outre apparu à l'usage que la coordination des sélections des personnes ne suffisait pas à répondre aux attentes des équipes de l'OFS qui commandent les échantillons pour leurs enquêtes. Celles-ci désirent en effet qu'il y ait une coordination au niveau des ménages, de sorte que deux personnes d'un même ménage ne puissent être sélectionnées de manière trop rapprochée.

Ces deux points, et l'entrée en vigueur de l'obligation légale de constitution des ménages dans les registres de population des cantons et communes nous à conduit à repenser toute notre procédure de tirage, et en même temps à proposer une méthode plus simple pour obtenir des échantillons avec une seule sélection par ménage. D'ici à fin 2013, l'OFS prévoit de constituer un cadre de sondage de ménages à partir du fichier SRPH et de sélectionner ses échantillons en deux phases. Une première phase sera un tirage coordonné de ménages dans le cadre de ménages, et, pour les

enquêtes auprès des personnes, la deuxième phase sera la sélection d'un individu par ménage sélectionné en première phase. On obtiendra de cette manière un système qui répondra bien aux besoins de l'OFS quant à la répartition de la charge d'enquête sur les ménages et qui permet d'éviter les complications évoquées dans la section 2.3.3. En outre, les plans transversaux des enquêtes ne seront pas modifiés par ce changement. La coordination au niveau des personnes sera par contre dégradée, puisque rien n'est prévu pour éviter que la même personne d'un ménage de plusieurs personnes ne soit re-sélectionnée lorsque, après un certain temps, le ménage lui-même est à nouveau re-sélectionné.

La transition vers ce nouveau processus de tirage est actuellement à l'étude afin de permettre une assez bonne coordination négative entre les enquêtes tirées selon la nouvelle procédure et celles tirées selon l'ancienne procédure. Plus précisément, les numéros aléatoires permanents des ménages seront générés en fonction des sélections et probabilités d'inclusion de leurs membres aux enquêtes passées.

2.8 Conclusions

Nous avons effectué un tour d'horizon des enquêtes par échantillonnage auprès des personnes et des ménages à l'OFS. La généralisation à toutes les enquêtes de la coordination des échantillons permettra, dans un avenir proche, de répartir au mieux la charge d'enquête sur les ménages et les personnes. Outre le confort supplémentaire que cela offre à la population, l'OFS espère que cette répartition équitable de la charge permettra d'améliorer les taux de réponse à ses enquêtes. Dans le domaine des pondérations, les quelques exemples traités montrent qu'il existe une diversité de situations et de besoins qui conduisent à des calculs et l'implémentation de méthodologies différents. La création du cadre d'échantillonnage SRPH et la généralisation de son utilisation pour toutes les enquêtes de l'OFS devraient permettre d'harmoniser davantage ces calculs. La simplification de ces calculs sera possible grâce à la connaissance précise de la population et de sa probabilité de sélection dans chaque vague d'enquête ou bloc de rotation pour les enquêtes de l'OFS. L'Office tente d'instaurer dans tous ses services l'utilisation de bonnes pratiques pour le traitement des données. L'imputation est un sujet complexe et le potentiel d'amélioration dans ce domaine est grand. Les progrès sont conditionnés par un investissement conséquent en temps et en moyens humains.

Par ailleurs, dans la mesure de ses capacités, l'OFS étudie et met en pratique les avancées récentes dans le domaine en constante évolution de la méthodologie d'enquêtes. Ainsi, les méthodes de calage "lune et étoile" et de calage généralisé ont déjà été mises à l'essai avec un certain succès. L'OFS participe également à des projets internationaux en collaboration avec d'autres instituts nationaux de statistique et universités. Il poursuit également sa collaboration étroite avec l'Institut de Statistique de l'Université de Neuchâtel. Ces collaborations couvrent des domaines variés comme l'estimation de variance, le traitement des données d'enquête,

les méthodes d'estimation sur des petits domaines, la mesure des inégalités. L'OFS s'attache à mettre en pratique les recommandations formulées par les grands projets européens sur la méthodologie d'enquête. La qualité des données et la probité des méthodes employées sont au cœur des préoccupations de l'Office.

Cet article est partiellement soutenu par une convention de collaboration entre l'Office fédéral de la statistique et l'Université de Neuchâtel. Son contenu n'engage que les auteurs et en aucun cas l'Office fédéral de la statistique. Pour obtenir des informations sur les enquêtes de l'OFS, le lecteur peut consulter le site internet de l'office (http://www.bfs.admin.ch/bfs/portal/fr/index.html).

3 Variance Estimation Using Linearization for Poverty and Social Exclusion Indicators

Abstract

We have used the generalized linearization technique based on the concept of influence function, as Osier has done (Osier, 2009), to estimate the variance of complex statistics such as Laeken indicators. Simulations conducted using the R language show that the use of Gaussian kernel estimation to estimate an income density function results in a strongly biased variance estimate. We are proposing two other density estimation methods that significantly reduce the observed bias. One of the methods has already been outlined by Deville (1999b). The results published in this article will help to significantly improve the quality of information on the precision of certain Laeken indicators that are disseminated and compared internationally.[1]

Keywords: influence function, EU-SILC survey, non-linear statistics, poverty and inequality indicators

3.1 Introduction

Deville (1999b) proposed that the precision of non-linear statistics in sampling designs be estimated using the generalized linearization method, which relies on the concept of influence function proposed by Hampel (1974) in the field of robust statistics. Osier (2009) applied these theories to estimate the variance of complex statistics such as the Laeken indicators (Eurostat, 2005a) in the European Statistics on Income and Living

[1]This chapter is an authorized reprint of: Graf, E. & Tillé, Y. (2014). Variance Estimation Using Linearization for Poverty and Social Exclusion Indicators. *Survey Methodology* **40**, 61–79.

Conditions (EU-SILC) survey. Goga et al. (2009) extend the theory of Deville (1999b) to two-sample surveys. Verma & Betti (2011) provide a comprehensive list of traditional poverty indicators and associated linearized variables, and they also compare the performance of the linearization technique with that of the jackknife repeated replication method. In this article, we will limit ourselves to poverty indicators published in the EU-SILC survey and focus on the way to estimate the income density function at various points in the distribution.

In Section 3.2, we review the required theoretical foundations, the expressions for the poverty and inequality indicators being studied, and the linearized variables of those indicators. Some linearized variables are dependent on the density function of the variable of interest, which is usually estimated using the Gaussian kernel estimation method. Two alternative methods are presented in Section 3.3. The R-language simulations are described and discussed in Section 4.7. We show that Gaussian kernel estimation can generate strong bias in the estimated variance of indicators when an estimate of the income density function is being used. We also show that the other two density estimation methods proposed in Section 3.3 reduce the observed bias, and this is discussed in the findings in the last part of this article.

3.2 Review of given poverty indicators and their linearized variables

Let U be a finite population consisting of N identifiable units $u_1, ..., u_k, ..., u_N$. To simplify the notation, let unit u_k be denoted simply by the index k. In practice the population U is a sampling frame with acceptable coverage of a given population for which we wish to make inferences. To each unit k is associated the value y_k for a given characteristic (in this case, income). Without loss of generality, to simplify the notation, assume that the values of y_k are distinct and sorted by order of magnitude, so that $y_k = y_{[k]}$. Data from sample surveys often contain duplicates, that is, a number of units with the same value y, as a result of rounding or range questions. In these cases and for this study, we can simply increase the values by a small (negligible), randomly selected, uniformly distributed amount so that the data may be sorted unambiguously.

Let S be a random sample of size n obtained using a sample design $p(s) = \mathrm{P}(S = s)$, for all $s \subset U$. In addition, let $\pi_k = \mathrm{P}(k \in s) > 0$ be the inclusion probability of unit k of U. As well, let $d_k = 1/\pi_k$ be the sampling weight, and let $w_k = w_k(s)$ be the estimation weight, which may be equal to d_k or may be more refined. For example, w_k may have been obtained after calibration (Deville & Särndal, 1992) and therefore also reflect a non-response adjustment.

The estimators of poverty and inequality indicators are non-linear statistics that cannot be expressed as regular functions of totals (that is, continuously differentiable up to the second order). In fact, they are rank statistics for the Gini coefficient and quantile statistics for the others. As

Osier (2009), points out, their variance therefore cannot be estimated using a Taylor linearization; the generalized linearization method is required instead (Deville, 1999b; Demnati & Rao, 2004a; Osier, 2009). An alternative for estimating variance would be to use bootstrap-type re-sampling techniques but, for the EU-SILC survey data, preference was given to the linearization technique, at least for a certain number of participating countries. Indeed, re-sampling methods often require more human and machine resources. As well, since Eurostat collaborates with some 30 countries that have different sampling designs and that may perform non-response adjustments and calibration to external sources, it seemed more appropriate to select an analytical solution for estimating variance. In addition, some countries might be using the existing SAS software POULPE (Ardilly & Osier, 2007) to generate the required estimates. That was the case for initial tests using Swiss EU-SILC data. Here we use a procedure that, as Antal et al. (2011), point out, reconciles the approach introduced by Deville (1999b) with that of Demnati & Rao (2004a). Both approaches use the concept of *influence function* initially developed in the field of robust statistics (Hampel, 1974). Antal et al. (2011) also state that the same linearized variables can be found by applying the method proposed by Graf (2013) that constructs a linearized variable on the basis of a Taylor expansion with respect to sample inclusion indicators. Note also the work by Kovacevic & Binder (1997) in which a linearization approach using estimating equations is developed.

Deville (1999b) states that the influence of a unit k on a population parameter of interest θ is determined by an infinitesimal variation in the importance assigned to the unit. The parameter is expressed as a functional $\theta = T(M)$, where M is a measure allocating a mass of 1, $M(k) = M_k = 1$, only at points on the continuum corresponding to units $k \in U$. The specialization of the general measure M into a discrete measure turns the functional T, predefined on a continuum, into a discrete functional, in the same way as the total Y is defined as the sum of all y_k over the given finite population. The *influence function* of T, or the *linearized variable*, is defined as

$$I[T(M)]_k = z_k = \lim_{t \to 0} \frac{T(M + t\delta_k) - T(M)}{t}, \text{ for all } k \in U,$$

where δ_k is the Dirac measure for unit k ($\delta_k(i) = 1$ if $i = k$ and 0 otherwise). In practice, we have known data only from a sample S and Deville (1999b) defines a linearized variable $\hat{z}_k$ or *empirical influence function*, by (1) determining the limit above using differential calculus and (2) replacing the unknowns in the evaluation with the corresponding estimated quantities using the sample. Deville justifies this procedure by showing that

$$T(\hat{M}) - T(M) \approx \left(\sum_{k \in S} w_k z_k - \sum_{k \in U} z_k \right).$$

The key result is that, under asymptotic conditions described by Deville (1999b), which are in theory satisfied when the sample is "sufficiently large", the variance of the estimated total of the variable $\hat{z}_k$ is an approxi-

mation of the variance of the (complex) statistic $\hat{\theta}$:

$$\text{var}\left[\sum_{k\in s}\hat{z}_k w_k\right] \approx \text{var}(\hat{\theta}).$$

The starting point of Deville's approach is therefore the population parameter and not the estimator that is proposed to be used for the evaluation using the sample. When the estimator used follows naturally from the population parameter expression (for example, the total Y approached by the Horvitz-Thompson estimator), the procedure is unambiguous. However, imprecision arises if we estimate the same total Y using the ratio estimator with an auxiliary variable x. In that case, Deville's approach, which does not specify the form of the total estimator to use, will yield a constant influence function equal to 1, instead of bringing the unknown ratio of interest into play.

An alternative that avoids these problems is the approach by Demnati-Rao, when used in Deville's framework, as done in Antal et al. (2011). They present the Demnati-Rao approach as resulting from Deville's framework when the measure M used is not the discrete measure defined for U described above, but rather the following measure defined for S, the sample:

$$\hat{M}(k) = w_k,\ k \in S$$

where w_k is a weight. By defining the measure for S, the starting point becomes the estimator and not the parameter; it is the parameter that is initially expressed as a functional, and not the population parameter to be estimated. That is, the functional corresponds to the estimator for which we are seeking a variance estimate using generalized linearization. We then obtain the linearized variable based on that functional as follows:

$$I[T(\hat{M})]_k = \hat{z}_k = \lim_{t\to 0}\frac{T(\hat{M}+t\delta_k) - T(\hat{M})}{t},\ \text{pour tout } k \in S.$$

Antal et al. (2011) note that, to the extent that the functional in this limit is expressed as an explicit function of the variables that are the weights assigned by the measure $\hat{M}$ to the observations, this linearized variable is in fact a function of the partial derivatives with respect to the weights:

$$I[T(\hat{M})]_k = \frac{\partial T(\hat{M})}{\partial w_k}.$$

Antal et al. (2011) point out that the linearized variables that we will discuss below can be obtained using either approach. In fact, computing the limit using the Demnati-Rao approach does not necessarily result in the variance estimate suggested by Deville (1999b). The practical approach used in this article might therefore be called the Deville-Demnati-Rao approach, in recognition of the theoretical framework provided by Deville (1999b) and the practical algorithm for Deville's framework provided by Demnati & Rao (2004a).

Using this method, the variance of $\hat{\theta}$ can be estimated for any sampling design, and a confidence interval can therefore be obtained by substituting

the linearized variable in the variance formula for a total for the selected sampling design. If the sampling design is simple random sampling without replacement, the estimator of the variance of an inequality indicator $\hat{\theta}$ is defined as

$$\widehat{\text{var}}_{lin}\left[\hat{\theta}\right] = \frac{N(N-n)}{n}\frac{1}{n-1}\sum_{k\in S}(\hat{z}_k - \overline{z})^2, \tag{3.1}$$

where

$$\overline{z} = n^{-1}\sum_{k\in S}\hat{z}_k.$$

Below, we review the empirical definitions of the inequality indicators considered with respect to population income measurement, as well as the expressions for the linearized variables of the indicators as we have implemented them.

3.2.1 Gini coefficient

The Gini coefficient, G, ranges from 0 (complete equality, that is, all individuals earn the same amount) to 1 (complete inequality, that is, one individual has all the income and the other individuals have no income). The coefficient G is expressed on the basis of the cumulative income of a given proportion of the poorest individuals. If $\mathcal{Y}$ is a random variable representing income, $f(y)$ its density function and $F(y)$ its distribution function, then the *Lorenz curve* Lorenz (1905) can be written as

$$L(\alpha) = \frac{\int_0^{F^{-1}(\alpha)} yf(y)dy}{\int_0^{\infty} yf(y)dy} = \frac{1}{\text{E}(\mathcal{Y})}\int_0^{\alpha} F^{-1}(u)du.$$

The Gini coefficient represents twice the area between the Lorenz curve and the line of complete equality (the diagonal line $f_{eg}(x) = x$) as shown in Figure 3.1. Therefore, the Gini coefficient can be defined as:

$$G = 2\int_0^1 [\alpha - L(\alpha)]\, d\alpha.$$

If a population U is finite, then the values of y_k will not be random and the Gini coefficient can be calculated as:

$$G = \frac{2\sum_{k\in U} ky_k}{N\sum_{k\in U} y_k} - \frac{N+1}{N},$$

where the values of y_k are sorted by rank. For a sample, the Gini coefficient can be estimated as

$$\begin{aligned}\widehat{G} &= \frac{2}{\widehat{N}\widehat{Y}}\sum_{k\in S} w_k\widehat{N}_k y_k - \left(1 + \frac{1}{\widehat{N}\widehat{Y}}\sum_{k\in S} w_k^2 y_k\right)\\ &= \frac{\sum_{k\in S}\sum_{\ell\in S} w_k w_\ell |y_k - y_\ell|}{2\widehat{N}\widehat{Y}},\end{aligned}$$

where $\widehat{N}_k = \sum_{\ell\in S} w_\ell \mathbf{1}_{[y_\ell \le y_k]}$ is the sum of the weights w_k, $\widehat{Y} = \sum_{k\in S} w_k y_k$ is the estimated total income of the population, and $\widehat{N} = \sum_{k\in S} w_k$ is the

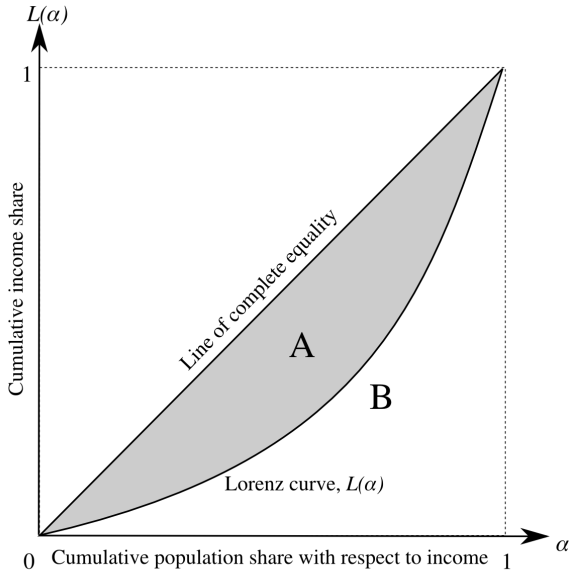


Figure 3.1 – *Gini coefficient, G, and Lorenz curve* $L(\alpha)$. $G = 2A$, $A + B = 1/2$.

estimated size of the population. The expression can be simplified as follows if all the weights are equal to N/n:

$$\widehat{G} = \frac{2\sum_{k\in S} k y_k}{n\sum_{k\in S} y_k} - \frac{n+1}{n}.$$

Note that the definition may vary by a factor of $n/(n-1)$ depending on the author (Osier, 2009; Eurostat, 2004c); however, this subtlety becomes negligible if the sample is large enough..

Langel & Tillé (2013) combine the various approaches to obtain the same estimated linearized variable of the Gini coefficient for the sample:

$$\hat{z}_k^{GINI} = \frac{1}{\widehat{N}\widehat{Y}}\left[2\widehat{N}_k(y_k - \widehat{\overline{Y}}_k) + \widehat{Y} - \widehat{N}y_k - \widehat{G}(\widehat{Y} + y_k\widehat{N})\right],$$

where $\widehat{\overline{Y}}_k = \sum_{\ell=1}^{k} w_\ell y_\ell / \widehat{N}_k$, and the values of y_ℓ are sorted and distinct.

3.2.2 Quintile Share Ratio (QSR or S_{80}/S_{20})

A good overview of this indicator is provided by Langel & Tillé (2011). Let q_{80} and q_{20} be the 80th and 20th percentiles of the distribution function $F(y)$. The QSR is the ratio of the total income of the 20% of the population with the highest income to the total income of the 20% of the population with the lowest income. In the continuous case, the QSR can be defined as

$$\text{QSR} = \frac{\text{E}(\mathcal{Y}|\mathcal{Y} > q_{80})}{\text{E}(\mathcal{Y}|\mathcal{Y} < q_{20})} = \frac{1 - L(0.8)}{L(0.2)},$$

where $\mathcal{Y}$ is a random variable representing income. For finite populations, the QSR can be expressed and estimated for a sample on the basis of partial sums,

$$\widehat{\text{QSR}} = \frac{\hat{Y} - \hat{Y}_{0.8}}{\hat{Y}_{0.2}},$$

where, given the results obtained by Langel & Tillé (2011), we will use the following definition of the partial sum, which differs slightly from the official definition of Eurostat (2004a),

$$\hat{Y}_\alpha = \sum_{k \in S} w_k y_k H\left(\frac{\alpha\hat{N} - \hat{N}_{k-1}}{w_k}\right), \tag{3.2}$$

with

$$H(x) = \begin{cases} 0 & \text{si } x < 0 \\ x & \text{si } 0 \leq x < 1 \\ 1 & \text{si } x \geq 1. \end{cases}$$

To obtain the linearized variable of the QSR, we must first calculate the linearized variable of the partial sum (3.2), which is

$$I(Y_\alpha)_k = y_k H(\alpha N - k + 1) + [\alpha - \mathbf{1}_{[y_k < Q_\alpha]}]Q_\alpha,$$

where $Q_\alpha = y_i$, with $\hat{N}_{i-1} < \alpha\hat{N} \leq \hat{N}_i$, corresponds to the first definition of the quantile of a finite population in the article by Hyndman & Fan (1996). Osier (2009) obtains a linearized variable that is dependent on the density of the variable $\mathcal{Y}$. However Langel & Tillé (2011) have shown that the problem of estimating this density for the QSR can be avoided through a simplification, so that it is not necessary to make a kernel approximation of income density as proposed by Osier (2009).

The influence function of the QSR is dependent on the influence functions of the partial sums:

$$I(\text{QSR})_k = z_k^{\text{QSR}} = \frac{y_k - I(Y_{0.8})}{Y_{0.2}} - \frac{(Y - Y_{0.8})I(Y_{0.2})}{Y_{0.2}^2}.$$

By making the necessary substitutions, we can see that the estimated linearized variable for a sample is

$$\begin{aligned} \hat{z}_k^{\text{QSR}} \quad = \quad & \frac{y_k - \left\{y_k H\left(\frac{0.8\hat{N} - \hat{N}_{k-1}}{w_k}\right) + \hat{Q}_{0.8}[0.8 - \mathbf{1}_{[y_k < \hat{Q}_{0.8}]}]\right\}}{\hat{Y}_{0.2}} \\ & - \frac{(\hat{Y} - \hat{Y}_{0.8})\left\{y_k H\left(\frac{0.2\hat{N} - \hat{N}_{k-1}}{w_k}\right) + \hat{Q}_{0.2}[0.2 - \mathbf{1}_{[y_k < \hat{Q}_{0.2}]}]\right\}}{\hat{Y}_{0.2}^2}. \end{aligned} \tag{3.3}$$

3.2.3 Linearized variable of a quantile

Before we discuss poverty indicators, we should give a few details on the linearized variable of an α-order quantile, which can be expressed as

$$\hat{z}_k^{Q_\alpha} = -\frac{1}{f(\hat{Q}_\alpha)}\frac{1}{\hat{N}}[\mathbf{1}_{[y_k \leq \hat{Q}_\alpha]} - \alpha],$$

where the weighted quantile can be defined in a manner similar to the partial sum (3.2) and $f(\cdot)$ is an income density function that will be discussed in details in Section 3.3. Note that Eurostat (2004a) recommends the second definition by Hyndman & Fan (1996). We could dispute

the Eurostat definition and use another quantile definition, for example $Q_\alpha = y_{k-1} + (y_k - y_{k-1})[\alpha N - (k-1)]$ where $\alpha N < k \leq \alpha N + 1$, which is the fourth definition according to Hyndman & Fan (1996). We then estimate the quantile for a sample as follows:

$$\hat{Q}_\alpha = y_{k-1} + (y_k - y_{k-1}) \left(\frac{\alpha \hat{N} - \hat{N}_{k-1}}{w_k} \right).$$

The linearized variable of a quantile is dependent on the value of the income density function in that quantile. However, the actual income density is unknown and therefore must also be estimated using the sample. Deville (1999b) and Osier (2009) suggest the use of Gaussian kernel estimation. We will discuss the problem of estimating f in more details in Section 3.3.

In addition to the problem of estimating the income density function, Croux (1998) shows that the empirical influence function of the median income is not a consistent estimator of the corresponding theoretical influence function. For a positive variable (such as income), the empirical influence function of the median income (the case discussed in Croux's article) converges toward an exponential distribution, the expectation of which is the influence function. It is not robust to large proportions of extreme values. It can be said to lack robustness in that the value of the estimator for a sample can differ greatly from the actual value for the population as a result of outliers (that is, values that are relatively very large) in the sample (see Hampel (1974) for a basic idea of robustness for infinite populations, and Beaumont et al. (2013) for recent thoughts on this topic for finite population sampling).

3.2.4 Median income and at-risk-of-poverty threshold

Let $\hat{m} = \hat{Q}_{0.5}$ be the estimated median income of the sample. The At Risk of Poverty Threshold (ARPT) is defined as 60% of the median income:

$$\begin{aligned} \text{ARPT} &= 0.6\, F^{-1}(0.5) \\ \widehat{\text{ARPT}} &= 0.6\, \hat{Q}_{0.5} = 0.6\, \hat{m}. \end{aligned}$$

This is an absolute measure that is scale-dependent. The linearized variable of the ARPT is proportional to that of the median income:

$$\hat{z}_k^{\text{ARPT}} = I(\text{ARPT})_k = 0.6\, I(\text{MED})_k = -\frac{0.6}{f(\hat{m})} \frac{1}{\hat{N}} [\mathbf{1}_{[y_k \leq \hat{m}]} - 0.5].$$

3.2.5 At Risk of Poverty Rate

The At Risk of Poverty Rate (ARPR), where ARPR $\in [0, 1]$, is the share of the population with an income below the ARPT: ARPR $= F(\text{ARPT})$. The ARPR is scale-independent, like the Gini coefficient, QSR and relative median poverty gap (see Section 3.2.7). The official Eurostat definition Eurostat (2004a) of the estimated ARPR for a sample is

$$\widehat{\text{ARPR}} = \frac{\sum_{y_k < \widehat{\text{ARPT}}} w_k}{\hat{N}}.$$

The linearized variable of the ARPR is defined by Osier (2009) as

$$\begin{aligned}\hat{z}_k^{\text{ARPR}} &= \frac{1}{\hat{N}}\left(\mathbf{1}_{[y_k\leq\widehat{\text{ARPT}}]} - \widehat{\text{ARPR}}\right) - \frac{\hat{f}(\widehat{\text{ARPT}})}{\hat{f}(\hat{m})}\frac{0.6}{\hat{N}}\left(\mathbf{1}_{[y_k\leq\hat{m}]} - 0.5\right)\\ &= \frac{1}{\hat{N}}\left(\mathbf{1}_{[y_k\leq\widehat{\text{ARPT}}]} - \widehat{\text{ARPR}}\right) + \hat{f}(\widehat{\text{ARPT}})\,\hat{z}_k^{\text{ARPT}}.\end{aligned}$$

Here, the income density function must be estimated at two points, namely the median income and the ARPT.

3.2.6 Median income of individuals below the ARPT

The median income of individuals below the ARPT is $m_p = F^{-1}(\frac{1}{2}F(\text{ARPT}))$. It is estimated in the same way as any other quantile, the exact definition of which may vary. The linearized variable of m_p (Osier, 2009) is dependent on that of the ARPR:

$$\hat{z}_k^{m_p} = \frac{1}{\hat{f}(\hat{m}_p)}\frac{\hat{z}_k^{\text{ARPR}}}{2} - \frac{1}{\hat{N}}\left(\mathbf{1}_{[y_k\leq\hat{m}_p]} - F(\hat{m}_p)\right).$$

The estimated income density therefore appears three times, namely in the median income and ARPT for $\hat{z}_k^{\text{ARPR}}$, and in the median income of individuals below the ARPT, m_p.

3.2.7 Relative Median Poverty Gap

The relative median poverty gap (RMPG) is the relative difference between the ARPT and the median income of individuals below the ARPT. RMPG $= 0$ if the income of all "poor" individuals is equal to the ARPT, and RMPG $= 1$ sif the income of all these individuals is zero. The RMPG is a measure of the extent to which the "poor" individuals are poor:

$$\text{RMPG} = \frac{\text{ARPT} - m_p}{\text{ARPT}}.$$

The estimated RMPG for a sample has already been described. The influence of each observation on the RMPG is defined by Osier (2009):

$$\hat{z}_k^{\text{RMPG}} = \frac{\hat{m}_p\,\hat{z}_k^{\text{ARPT}} - \widehat{\text{ARPT}}\,\hat{z}_k^{m_p}}{\widehat{\text{ARPT}}^2}.$$

The estimated income distribution density appears four times: once in the calculation of $\hat{z}_k^{\text{ARPT}}$ and three times in the calculation of $\hat{z}_k^{m_p}$.

3.3 Estimating the income density function

In a design-based approach with a finite population, inference is made in relation to the sampling design $\text{P}(S)$ used to select a sample S from a finite population U of size N. In this approach, only the sample inclusion indicators are random; all other quantities are fixed. The population income

distribution function is then a step function: $F_y(x) = \sum_{k \in U} \mathbf{1}_{y_k \leq x} / N$, and its derivative, the density function, does not exist due to discontinuities. If a model-based approach with a super-population model to justify the income density function term is not desired, then the distribution function must be artificially smoothed to make it differentiable. Therefore, our use of "density function" is not quite correct. For purposes of smoothing, Deville (1999b) and Osier (2009) suggest using Gaussian kernel estimation to estimate the income density function:

$$\begin{aligned} K(u) &= \frac{1}{h\sqrt{2\pi}} e^{-\frac{u^2}{2}}, \qquad u = \frac{x - y_k}{h} \\ \hat{f}_1(x) &= \frac{1}{\hat{N}} \sum_{k \in S} w_k \, K\left(\frac{x - y_k}{h}\right) \\ &= \frac{1}{h\sqrt{2\pi}} \frac{1}{\hat{N}} \sum_{k \in S} w_k \, \exp\left[-\frac{(x - y_k)^2}{2h^2}\right] \end{aligned} \tag{3.4}$$

where h is the bandwidth that Osier estimates using $\hat{h} = \hat{\sigma}\hat{N}^{-0.2}$ and $\hat{\sigma}$ is the estimated standard deviation of the empirical income distribution:

$$\hat{\sigma} = \sqrt{\frac{\sum_{k \in S} w_k y_k^2}{\hat{N}} - \left(\frac{\sum_{k \in S} w_k y_k}{\hat{N}}\right)^2} = \sqrt{\frac{\sum_{k \in S} w_k y_k^2}{\hat{N}} - \overline{y}_w^2}.$$

Note that this estimate of σ is not robust, since it is very sensitive to the extreme values of y. Income data often have a distribution tail extending to the right with values that may be extremely high; these are "representative outliers" as defined by Chambers (1986) et Hulliger (1999). As the simulations in Section 4.7 will show, this can generate a strong bias in the variance estimates. Verma & Betti (2011) also use kernel estimation, recalling that Silverman (1986), states that the choice of kernel is not critical to ensure that $\hat{f}(y)$ converges toward $f(y)$, but the choice of bandwidth is. They use a value recommended by Silverman for distributions with a positive skewness coefficient, $h = 0.79(\hat{Q}_{75} - \hat{Q}_{25})\hat{N}^{-0.2}$. In their findings, they point out that the linearization method may be problematic because of irregularities in the empirical density function. They also state that these problems are all the more cause for concern because survey data often contain groups of observations with the same value (due to rounding or range questions), which can make estimating the density more complicated. The rest of this article describes the solutions we are proposing to reduce bias in variance estimates.

3.3.1 Using the logarithm

One solution that produces very good results, as shown below, is simply to use the logarithm to estimate the density of x. If $v = \log(x + a)$, where x is the income and a is a positive real number equals to, say, $(|\min_k(y_k)| + 1)$ where there may be negative or zero incomes (ignoring that a would be estimated), then

$$F_v(v) = \mathrm{P}\left(\mathcal{V} \leq v\right) = \mathrm{P}\left(\log(\mathcal{Y} + a) \leq v\right) = \mathrm{P}\left(\mathcal{Y} \leq e^v - a\right) = F_y\left(e^v - a\right),$$

where $\mathcal{V}$ and $\mathcal{Y}$ would be random variables. Therefore,

$$f_v(v) = \frac{dF_v(v)}{dv} = \frac{dF_y(e^v - a)}{dv} = f_y(e^v - a)e^v.$$

That is, $f_v(v) = f_y(x)(x + a)$, which gives us the following estimator for the density of x:

$$\hat{f}_2(x) = \frac{\hat{f}_v(v)}{x + a} = \frac{\hat{f}_y\left(\log(x + a)\right)}{x + a}. \tag{3.5}$$

The estimated density at x for $\mathcal{Y}$ can therefore be determined by estimating the density of the logarithm of the variable divided by the value of the variable at a given point. This property is valid for finite populations. Using the logarithm has the advantage of reducing the leveraging effect of large income values in the kernel density approximation calculation. Simulations show that this simple method significantly reduces bias.

3.3.2 Nearest neighbour with minimum bandwidth

Deville (1999b) outlines another density estimation method that is a "nearest neighbour" method (see Silverman, 1986) using the kernel

$$K_D(u) = \begin{cases} \frac{1}{b-a} & \text{si } a \leq u < b \\ 0 & \text{sinon,} \end{cases},$$

where $u = y_k$ and the choice of a and b, with $x \in [a, b]$, is to be determined and could depend on x. The distance $(b - a)$ represents the bandwidth h. The density estimate would therefore be

$$\begin{aligned} \hat{f}_D(x, a, b) &= \frac{1}{\hat{N}} \sum_{k \in S} K_D(y_k) \\ &= \frac{1}{\hat{N}} \sum_{k \in S} w_k \frac{1}{b - a} \mathbf{1}_{y_k \in [a, b[} \\ &= \frac{\hat{F}_y(b) - \hat{F}_y(a)}{b - a}, x \in [a, b[\end{aligned} \tag{3.6}$$

where $\hat{F}_y(x) = \sum_{k \in S} w_k \mathbf{1}_{y_k \leq x} / \hat{N}$.

Note that the density estimate (3.6) is not a continuous function and would not be suitable for estimating density values at the end tails of the distribution. Since our work relies little on distribution tails, we shall consider this approach as an option.

Our second proposal for estimating the density of x is based on the idea above. It is a nearest neighbour method, but also imposes a minimum bandwidth. Specifically, our method requires the use of at least p observations nearest to point x with minimum bandwidth $h(p) \geq h_{opt}$ where

$$h_{opt} = \frac{0.9 \min(\hat{\sigma}, \hat{Q}_{75} - \hat{Q}_{25})}{1.34 \sqrt[5]{\hat{N}}}$$

is the rule of thumb Silverman (1986) for determining the bandwidth. This is also the default bandwidth value in the R function `density`. This solution is more robust than (3.4) and avoids the problems that arise when a number of values y_k are very close to each other, which is often the case because the individuals interviewed tend to round their income.

As the values y_k, $k = 1, ..., n$, are assumed to be ordered by rank, the width $h(p)$ of the window around x is initially determined by the p nearest observations, where $p \ll n$. In the simulations discussed in the next section, after various trials, p was initially set at 30. The density of x is imputed to be the estimated density at the nearest observed point y_j that is less than or equal to x, that is $j = \max(k|y_k \leq x)$, $k = 1, ..., n$. The bandwidth at x in fact depends on the p_j nearest observations around y_j, with $p_j \geq p$, which will be denoted $h(p_j)$ in the rest of this article. The density is therefore estimated only at observed points, with no smoothing or interpolation between the values $\hat{f}(y_j)$. The algorithm for estimating

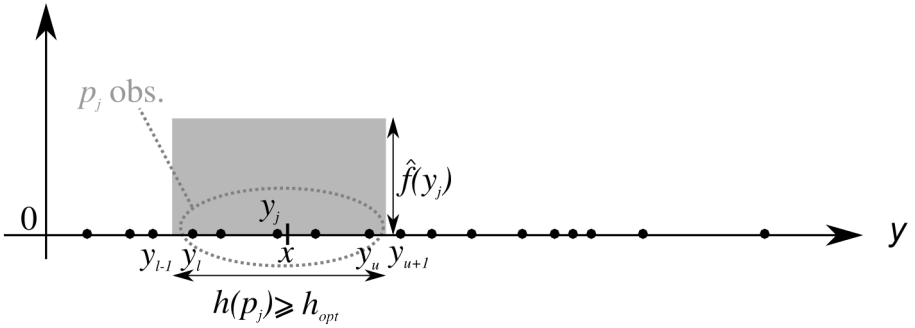


Figure 3.2 – *Window width* $h(p_j)$.

$\hat{f}(y_j)$ is as follows (see also Figure 3.2):

1. The initial width of the window around point y_j, with $p_j = p$, is defined as:

$$h(p_j) = \frac{y_u + y_{u+1}}{2} - \frac{y_\ell + y_{\ell-1}}{2}; \quad \begin{array}{rcl} u & = & \begin{cases} j + p_j/2 - 1 & \text{if } p_j \text{ is even} \\ j + \lfloor p_j/2 \rfloor & \text{if } p_j \text{ is odd} \end{cases} \\ \ell & = & j - \lfloor p_j/2 \rfloor. \end{array}$$

2. If the width of the resulting window $h(p_j)$ is less than h_{opt}, increment the two bounds:
upper bound: $u \to u + 1$, as long as $u < n$,
lower bound: $l \to l - 1$, as long as $l > 1$,
which implies $p_j \to p_j + 2$, unless $u = n$ or $l = 1$, in which case there is no longer the same number of points on each side of y_j.

3. Repeat step 2 until $h(p_j) \geq h_{opt}$.

4. The estimated density at x can be written as

$$\hat{f}(x) = \hat{f}(y_j) = \begin{cases} \dfrac{p_j}{n\, h(p_j)} & \text{without weighting,} \\ \dfrac{\sum\limits_{p_j \text{ closest to } y_j} w_j^{std}}{n\, h(p_j)} & \text{with weighting,} \end{cases}$$

with standardized weights $w_k^{std} = w_k/\overline{w}$, $k = 1, ..., n$.

The number of observations p_j used in the calculation may vary, and it depends on the local curvature of the empirical distribution function. The condition $h(p_j) \geq h_{opt}$ guarantees a minimum window width in places where numerous observations would be concentrated over a small interval. The procedure is made even more solid by combining this approach with the preceding approach, that is, by estimating the density of the logarithm of the variable divided by its (non-logarithmic) value:

$$\hat{f}_3(x) = \frac{\hat{f}(log(x+a))}{x+a}. \tag{3.7}$$

3.3.3 Robustness of the linearized variable

As stated above, for the median or for other quantiles, Croux (1998) points out that the empirical influence function or the linearized variable estimated using the sample is not as robust as it appears to be, even if the density function is known. We confirmed this for the EU-SILC data used in the model simulations with a Generalized Beta distribution of the second kind (GB2) by means of the R function `profml.gb2` (Graf & Nedyalkova, 2011). For small samples ($n \leq 100$), the potential bias of the linearized variable resulting from too many outliers may also bias the variance estimate calculated using the linearized variable. For larger samples ($n \geq 1000$), a maximum relative bias in the variance estimated using the empirical *versus* theoretical linearized variable may reach 5%. However, it is below the percentage in absolute terms three times out of four.

3.4 Results

Simulations were conducted on three sets of actual data to compare and assess the different density function estimation methods,. $\hat{f}_1(x)$, see (3.4), $\hat{f}_2(x)$, see (3.5) and $\hat{f}_3(x)$, see (3.7). These methods are required to estimate the variance of certain poverty and inequality indicators.

1. The first dataset contains equivalent household incomes from the EU-SILC survey conducted by the Swiss Federal Statistical Office in 2009. It includes 17,534 individuals with a non-zero income.

2. The second dataset also comes from the 2009 EU-SILC survey, but is limited to salaried individuals. It contains salaries from the register of the Central Compensation Office that has been linked with the survey respondents. We therefore have no non-response issues, and there are 7,922 individuals with a non-zero income.

3. The third test file, named *Ilocos*, comes with the R package `ineq` de R (Zeileis, 2012). It contains 632 observations, which are household incomes in Ilocos, one of the 16 regions of the Philippines. The data come from two surveys by the National Statistics Office of the Philippines, in 1997 and in 1998.

The three dataset have a positive skewness coefficient, which is typical of income distributions. Each data set is considered to be one population, and we initially selected 10,000 simple random samples without replacement of various sizes. The values of the various indicators were calculated for each sample, giving us a Monte Carlo estimate of their variance, $\text{var}_{sim}(\hat{\theta})$, for a poverty or inequality indicator θ. The variance estimator using linearization is denoted $\widehat{\text{var}}_{lin}(\hat{\theta})$ and is calculated using the linearization variable $\hat{z}^{\hat{\theta}}$ estimated for each sample:

$$\widehat{\text{var}}_{lin}(\hat{\theta}) = \frac{N(N-n)}{n}\text{var}(\hat{z}_S^{\hat{\theta}}),$$

where $n$is the size of the sample used for the simulations and $\text{var}(\hat{z}_S^{\hat{\theta}}) = \frac{1}{n-1}\sum_{k\in S}(\hat{z}_{S,k}^{\hat{\theta}} - \overline{z}_S^{\hat{\theta}})$ where $\overline{z}_S^{\hat{\theta}} = n^{-1}\sum_S \hat{z}_{S,k}^{\hat{\theta}}$, see (3.1).

The quality of the variance estimator using linearization is assessed by comparing the expected Monte Carlo value of the variance estimated using linearization, denoted $\text{E}_{sim}[\widehat{\text{var}}_{lin}(\hat{\theta})]$, with the "true" Monte Carlo variance $\text{var}_{sim}(\hat{\theta})$ in terms of relative bias:

$$\text{RB}[\widehat{\text{var}}_{lin}(\hat{\theta})] = \frac{\text{E}_{sim}[\widehat{\text{var}}_{lin}(\hat{\theta})] - \text{var}_{sim}(\hat{\theta})}{\text{var}_{sim}(\hat{\theta})}. \tag{3.8}$$

For the second data set (EU-SILC 2009, income of salaried individuals) we also, in a second step, selected 10,000 random samples without replacement under a stratified sampling design, and then calibrated the sampling weights to agree with the eight known sociodemographic marginal totals for the population of 7,922 individuals. The five strata used correspond to the age groups of the salaried individuals (see Table 3.1).

Table 3.1 – *Strata used in simulations with 2009 EU-SILC data and three sample sizes (income of salaried individuals, $N = 7,922$)*

Stratum h	Description	N_h	%	n_h		
1	individuals under 25	1187	15.0	75	112	150
2	26- to 35-year-olds	1359	17.2	86	129	171
3	36- to 45-year-olds	2137	27.0	135	202	270
4	46- to 55-year-olds	1864	23.5	117	177	235
5	individuals over 55	1375	17.4	87	130	174
	TOTAL	7922	100.0	500	750	1000

The eight calibration cells were obtained by crossing the three following dichotomous variables (auxiliary calibration variables):

1. MARIÉ which indicates whether or not the individual is married;
2. CHEF which indicates whether or not the individual's job is a management position; and
3. HOMME which indicates the individual's sex.

The totals for the population of 7,922 individuals for these calibration cells are shown in 3.2.

Table 3.2 – *Calibration margins in simulations with 2009 EU-SILC data (income of salaried individuals, $N = 7,922$)*

Margin	MARIÉ	CHEF	HOMME	Population total	%
1	0	0	0	1487	18.8
2	0	0	1	1208	15.2
3	0	1	0	323	4.1
4	0	1	1	457	5.8
5	1	0	0	1759	22.2
6	1	0	1	1278	16.1
7	1	1	0	328	4.1
8	1	1	1	1082	13.7
			TOTAL	7922	100.0

For each stratified sample, a calibration (linear method) was performed to make the sums of the weights agree with the eight margins shown above. Point estimates of the indicators and their linearized variable were computed for each sample using the calibrated weights.

Variance was estimated using the method developed by Deville (1999b). which consists of linearizing also with respect to the calibration by calculating the residuals $e^{\hat{\theta}}$ of the regression (weighted by the sampling weights) of the linearized variables of the indicators for the auxiliary calibration variables. The variance of the total of the residuals thus calculated, under a stratified random sampling plan without replacement is therefore an estimator of the variance of the estimated indicator; it is the quantity of interest:

$$\widehat{\text{var}}_{lin}(\hat{\theta}) = \sum_{h=1}^{H} \frac{N_h}{n_h}(N_h - n_h)s^2_{e^{\hat{\theta}}_h} \tag{3.9}$$

where

$$s^2_{e^{\hat{\theta}}_h} = \frac{1}{n_h - 1}\sum_{k \in S_h}(e^{\hat{\theta}}_k - \overline{e}^{\hat{\theta}})^2$$

The quality of the variance estimator using linearization is assessed analogously to the procedure for simple random sampling, see (3.8).

Table 3.3 – *Relative bias* (3.8) *of the variance obtained with 10,000 simple random samples without replacement from the 2009 EU-SILC data (equivalent household income, $N = 17,534$)*

	Sample size (sampling rate)								
Indicator	$n = 500(2.9\%)$			$n = 750(4.3\%)$			$n = 1000(5.7\%)$		
GINI	-0.02			-0.02			-0.02		
QSR	0.01			0.00			0.00		
	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_3$	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_3$	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_3$
ARPT	-0.08	-0.06	0.04	-0.09	-0.07	0.03	-0.09	-0.07	0.04
ARPR	-0.05	-0.01	-0.00	-0.09	-0.06	-0.05	-0.08	-0.05	-0.03
RMPG	-0.09	-0.07	**0.15**	**-0.10**	-0.07	**0.12**	-0.09	-0.06	**0.14**
MEDP	**-0.16**	**-0.12**	0.09	**-0.19**	**-0.13**	0.05	**-0.18**	**-0.11**	0.07
MED	-0.08	-0.06	0.05	-0.08	-0.06	0.04	-0.08	-0.06	0.04

Tables 3.3, 3.4 et 3.5 show the relative bias of the variance for the three data sets used and described above, using simple random sampling. Table 3.6 shows the relative bias of the variance using stratified random

Table 3.4 – *Relative bias* (3.8) *of the variance obtained with 10,000 simple random samples without replacement from the 2009 EU-SILC data (income of salaried individuals,* $N = 7,922$*)*

	Sample size (sampling rate)								
Indicator	$n = 500(6.3\%)$			$n = 750(9.5\%)$			$n = 1000(12.6\%)$		
GINI	-0.03			-0.03			-0.02		
QSR	-0.00			0.00			0.00		
	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_3$	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_3$	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_3$
ARPT	0.07	0.05	**0.13**	0.06	0.04	**0.10**	0.06	0.03	0.08
ARPR	-0.05	-0.04	-0.02	-0.05	-0.04	-0.01	-0.06	-0.05	-0.02
RMPG	**0.61**	**0.12**	**0.15**	**0.60**	**0.11**	0.08	**0.59**	0.09	0.05
MEDP	**0.73**	**0.17**	**0.18**	**0.72**	**0.16**	**0.10**	**0.72**	**0.15**	0.07
MED	0.07	0.04	**0.13**	0.06	0.04	**0.10**	0.05	0.03	0.07

Table 3.5 – *Relative bias* (3.8) *of the variance obtained with 10,000 simple random samples without replacement from Ilocos data (household income,* $N = 632$*)*

	Sample size (sampling rate)					
Indicator	$n = 50(7.9\%)$			$n = 63(10.0\%)$		
GINI	**-0.16**			**-0.13**		
QSR	0.00			0.00		
	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_3$	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_3$
ARPT	-0.05	-0.06	-0.01	-0.03	-0.03	-0.01
ARPR	**-0.31**	-0.01	**-0.12**	**-0.33**	-0.03	**-0.18**
RMPG	**1.55**	**0.83**	**0.26**	**1.54**	**0.16**	**0.39**
MEDP	**1.02**	**0.28**	**-0.26**	**1.05**	0.07	**-0.11**
MED	0.04	0.03	0.08	0.07	0.07	0.09

sampling with calibrated weights. The upper portions of the tables give the values for the Gini coefficient and QSR, which do not require estimating the income density function. The estimation of their variance works well. Note that there is a problem involving the underestimation of the variance of the Gini coefficient in the case of stratification with calibration (Table 3.6).

For the first data set, Table 3.3 does not reveal any major differences except that the estimation of income density using $\hat{f}_3(x)$ gives results that are more conservative. In fact, the relative bias remains of the same order of magnitude, but positive, while it is negative for the other two methods of estimating density. For the second data set, Table 3.4 shows that it is essential to use the logarithm or the nearest neighbour method with minimum bandwidth. The latter, all relative bias falls under 10% when the sample sizes are sufficiently large (see last column in the table). Simulations on the same data with a stratified sampling plan and calibration strengthen and confirm these results (see Table 3.6). For the third data set, Table 3.5 shows the same trends, although the results are less stable as a result of the small sample and population sizes. This is not surprising, since the minimum number of neighbours to consider is fixed at 30. In this case, for the Ilocos data set, simulations with a smaller value of p fixed at 10 makes no difference ultimately, because the condition $h(p_j) \geq h_{opt}$ automatically increases it above 30.

Table 3.6 – *Relative bias* (3.8) *of the variance obtained with 10,000 stratified random samples without replacement, with weights calibrated to eight sociodemographic margins, from the 2009 EU-SILC data (income of salaried individuals,* $N = 7,922$*).*

	Sample size (sampling rate)								
Indicator	$n = 500(6.3\%)$			$n = 750(9.5\%)$			$n = 1000(12.6\%)$		
GINI	**-0.21**			**-0.20**			**-0.20**		
QSR	-0.06			-0.06			-0.07		
	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_3$	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_3$	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_3$
ARPT	-0.07	-0.09	-0.01	-0.08	**-0.10**	-0.04	-0.09	**-0.11**	-0.06
ARPR	**-0.10**	**-0.10**	-0.08	-0.07	-0.06	-0.05	-0.06	-0.06	-0.05
RMPG	**0.63**	**0.13**	**0.13**	**0.61**	**0.11**	0.08	**0.59**	**0.10**	0.04
MEDP	**0.71**	**0.16**	**0.15**	**0.68**	**0.13**	0.09	**0.66**	**0.12**	0.04
MED	-0.07	-0.09	-0.01	-0.08	**-0.10**	-0.04	-0.08	**-0.11**	-0.06

Furthermore, generally speaking, we can see that the greater the use of Gaussian kernel density estimation -$\hat{f}_1(x)$- the greater the error. In fact, the relative bias of the variance for the median income of individuals below the ARPT and for the RMPG are almost systematically greater in absolute value that those for the other indicators. For the RMPG, the error may be offset (as in Table 3.3) if there are enough observations, since the density estimation appears in both the numerator and the denominator.

In short, we see that the variance can be overestimated ($\mathrm{RB}[\widehat{\mathrm{var}}_{lin}(\hat{\theta})] > 0$) or underestimated ($\mathrm{RB}[\widehat{\mathrm{var}}_{lin}(\hat{\theta})] < 0$) depending on the indicator and the data set. The use of the logarithm ($\hat{f}_2(x)$) provides significant improvement. The nearest neighbour method ($\hat{f}_3(x)$) eliminates all problems if there is enough data (as in Tables 3.3, 3.4 and 3.6). Slight problems arise with this method when the samples are small (as in Table 3.5). Illogical variations and bias that persist in the tables may also be the result of a lack of robustness in the linearized variables for certain samples, as stated in Section 3.3.3.

3.5 Conclusions

In a number of countries, national sample surveys publish extrapolated values for the Laeken indicators (Eurostat, 2005a) since they are key indicators that make it possible to direct decision makers with regard to political and social matters. It is therefore critical that we be able to quantify the precision of these measures, which raises the issue of the appropriateness of the precision estimates available. This article shows that a substantial improvement may be made in the precision estimates for poverty and inequality indicators by using a (local) estimate of the income density or given monetary variable.

The simulations conducted show that the Gaussian kernel density estimation method currently implemented in most cases is not recommended without at least using the logarithm as proposed in Section 3.3.1; otherwise, there may be significant bias in the estimated variance. The nearest neighbour method (Section 3.3.2) which also imposes a minimum bandwidth, may yield even better results, especially if there are agglomerations of observations with certain values in the given data. However, this

method requires setting a minimum number p of neighbours on the basis of the data used. If few observations are available, the use of the logarithm is preferable instead. In all cases, we hope that this work will help raise awareness of the importance of being meticulous during the implementation of calculations for the linearized variable of any indicator involving quantiles.

Acknowledgments

This work was carried out under a collaboration agreement between the Institute of Statistics at the University of Neuchâtel and the Swiss Federal Statistical Office (SFSO). We would also like to specifically thank the Income, Consumption and Living Conditions unit of the SFSO for providing us with data from the Swiss portion of the EU-SILC survey. Thanks also to Matti Langel and Anne Massiani for their support during our investigations.

4 Variance Estimation for Regression Imputed Quantiles

A first step towards Variance Estimation for Regression Imputed Inequality Indicators

Abstract

This work is done in the framework of sampling theory. It is based on a scenario in which a sample survey has been carried out and only a sub-part of the selected sample has accepted to answer (total non-response). Moreover, some respondents did not answer all questions (partial non-response). This is common scenario in practice. We are particularly interested in income type variables. Generally, total non-response is treated by re-weighting and partial non-response through imputation. One further supposes here that the imputation is carried out by a regression model adjusted on the respondents. We then resume the idea presented by Deville & Särndal (1994) and tested afterwards by Lee et al. (1994) which consists in constructing an unbiased estimator of the variance of a total based solely on the information at our disposal: the information known on the selected sample and the subset of respondents. While the two cited articles dealt with the exercise for a conventional total of an interest variable y, we reproduce here a similar development in the case where the considered total is one of the linearized variable of quantiles or of inequality indicators, and that, furthermore, it is computed from the imputed variable y. We show by means of simulations that regression imputation can have an important impact on the bias estimation as well as on the variance estimation of inequality indicators. This leads us to a method capable of taking into account the variance due to imputation in addition to the one due to the sampling design in the cases of quantiles. This method could be generalized to some of the inequality indicators.

Keywords: Influence function, SILC survey, linearization, bias, simulations, Laeken indicators

4.1 Introduction

This article considers firstly the estimation of variance for a total of a variable of interest. The data are collected through a sample survey and one supposes that, in addition to some re-weighting already done earlier in the process of statistical data preparation (see Graf & Qualité, 2014), to compensate for total nonresponse, the variable of interest y presents missing values. These are treated through parametric imputation consisting in a regression model adjusted on the respondents, which allows us to predict the values where they are missing. Other imputation methods are available, but it is not the goal of this article to debate between the different possible existing approaches allowing one to impute. The reason for our choice is fact that this work has been carried out with the goal of improving the methodology actually in place at the Swiss Federal Statistical Office for the treatment of the SILC (Survey on Income and Living Conditions, see European Union, 2014) data. In this context, after an evaluation process where different imputation methods were considered, it was decided to proceed through regression by the use of the Iveware SAS package (see Raghunathan et al., 2001). Once the variable is imputed, general linearization (Deville, 1999b; Demnati & Rao, 2004b; Osier, 2009) is used, following the operating mode described in Graf & Tillé (2014) in order to estimate the variance of different complex and non-linear statistics such as indicators of poverty and social exclusion also called inequality indicators (formerly called Laeken indicators, see Eurostat, 2003, 2005a). Presently, in the national statistical offices producing such estimations, imputed values are often considered as real values when it comes to applying the generalized linearization method. In other words, one neglects or forgets the fact that some of the values have been imputed. If the nonresponse rate is important, this leads to an underestimation of the variance, which can be considerable as we shall see. We concentrated ourselves on the case of quantiles while simultaneously underscoring and discussing the problem for inequality indicators published in the frame of the European SILC survey. The goal of the presented work is to provide a method allowing us to take into account the fact that some of the values have been imputed (through regression) in the variance estimation by linearization.

The notations are introduced in Section 4.2. In Section 4.3, we recall the method developed by Deville & Särndal (1994) in the case where one wants to estimate the variance of the total of the interest variable y. Lee et al. (1994) have tested this case in a convincing manner. Section 4.4 presents the additional issue we encounter when one no longer considers the variance of the total of y but the one of the total of a linearized variable based on y. In Section 4.5, we solve the problem for the case of quantiles leading us to formulae which have been implemented in the simulations of Section 4.6 for simple random sampling without replacement. The mentioned simulations, carried out on real data, show the effect of imputation on inequality indicators. They are described in the Section 4.6.1. The results are commented in Section 4.7 leading to the conclusions of Section 4.8.

4.2 Notations

Let U be a finite population comprised of N identifiable units $u_1, ..., u_k, ..., u_N$. To simplify the writing, in the following we shall use only k to designate unit u_k. For each unit k one associates a value y_k of the characteristic of interest (here some income). Without loss of generality and to reduce the notations, the y_k are supposed to be all distinct and ordered, that is $y_k = y_{[k]}$.

Let s be the set of individuals considered as respondents to the survey but not necessarily to the variable of interest y, r be the subset of respondents to y and $o = s \backslash r$ the nonrespondents to y. The vector of explanatory variables for the regression is noted $\mathbf{x}_{k\cdot} = (x_{k1}, x_{k2}, ..., x_{kJ})$. The information at our disposal for the individuals can be grouped in matrix form. One defines:

$$\mathbf{X} = (\mathbf{x}_{k\cdot})_{k\in s}, \qquad \mathbf{X}_r = (\mathbf{x}_{k\cdot})_{k\in r}, \qquad \mathbf{X}_o = (\mathbf{x}_{k\cdot})_{k\in o}.$$

The matrices $\mathbf{X}, \mathbf{X}_r, \mathbf{X}_o$ all have J columns, J being the number of explanatory variables considered in the regression model defined below. Let $y = \{y_k | k \in s\}$ be the complete variable without missing values (not at disposal in principle), $y_r = \{y_k | k \in r\}$ the values obtained on the respondents and $\hat{y}_{imp} = \mathbf{X}_o \hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\beta}} = (\mathbf{X}_r' \mathbf{X}_r)^{-1} \mathbf{X}_r' y_r$ being computed and based on the respondents. In other words, missing values are imputed with a regression model noted ξ which supposes for $k \in U$ that:

$$\xi : \quad y_k = \mathbf{x}_{k\cdot} \boldsymbol{\beta} + \varepsilon_k. \tag{4.1}$$

To simplify, we suppose that the error terms follow a normal distribution, are homoscedastic and not correlated: $\mathrm{E}_\xi(\epsilon_k) = 0$, $\mathrm{E}_\xi(\epsilon_k^2) = V_\xi(\epsilon_k) = \sigma^2$, $\mathrm{E}_\xi(\epsilon_k \epsilon_\ell) = 0 \, (k \neq \ell)$. It is possible to derive a similar, more complex, development than the one recalled here if one wishes to consider the fact that weights would have been used in the regression model ξ (heteroscedasticity). However, according to Deville & Särndal (1994), the variance estimation method shouldn't be too sensitive to the implication or not of weights in the computations. We satisfy ourselves here with a variance estimation after a non-weighted regression since this greatly simplifies the calculations. One defines the imputed variable there where it is necessary as follow:

$$y_{\bullet k} = \begin{cases} y_k & \text{si } k \in r \\ \hat{y}_{imp,k} = \mathbf{x}_{k\cdot} \hat{\boldsymbol{\beta}} & \text{si } k \in o. \end{cases}$$

Let us define also the following quantities:

- $Y = \sum_U y_k$, the population total of y,
- $\hat{Y} = \sum_s w_k y_k$, the estimated total of y from the sample s (unknown in principle because of the nonresponse affecting some values of y),
- $\hat{Y}_\bullet = \sum_s w_k y_{\bullet k} = \sum_r w_k y_k + \sum_o w_k \hat{y}_{imp,k}$, the estimated total of y on the sample s and after having imputed the missing values.

In Section 4.4 we will handle estimated linearized variables or empirical influence functions for a statistic A (population parameter of interest) which we will denote by z^A. Thus the influence of unit k on A is z_k^A and it depends on y_k. Similarly as above for y, let us define the following quantities:

- $Z = \sum_U z_k$, the population total of estimated linearized variable,
- $\hat{Z} = \sum_s w_k z_k$, the estimated total of z from the sample s (unknown in principle because of the nonresponse affecting some values of y),
- $\hat{Z}_\bullet = \sum_s w_k z_{\bullet k} = \sum_r w_k z_k + \sum_o w_k \hat{z}_{imp,k}$, the estimated total on the sample s of the estimated linearized variable z after having imputed the missing values for y.

4.3 Variance due to multiple regression imputation

This part summarizes the article of Deville & Särndal (1994) while developping the computations in the case of an imputation done by prediction through a multiple regression model with homoscedastic errors. This is a substantial introduction, but it is needed in order to articulate, argument and explain our work in the next section. Therefore Section 4.3 does not include any innovation. It allows us to introduce the essential concepts and notations needed for what follows.

4.3.1 Goal: obtain a confidence interval for the total Y

One wishes to estimate the **total variance** of $(\hat{Y}_\bullet - Y)$, $\hat{Y}_\bullet$ being the estimation of the total Y obtained by using the partially imputed data. Thus, if we assume that the estimator is normally distributed:

$$\hat{Y}_\bullet - Y \sim N(0, V_{tot})$$

$$q_{0.025} \leq \frac{\hat{Y}_\bullet - Y}{\sqrt{V_{tot}}} \leq q_{0.975},$$

if one can estimate V_{tot}, one can construct a 95% confidence interval for the true value of Y:

$$\hat{Y}_\bullet - q_{0.025}\sqrt{\hat{V}_{tot}} \lessapprox Y \lessapprox \hat{Y}_\bullet + q_{0.975}\sqrt{\hat{V}_{tot}}.$$

If we trust our model ξ, i.e. we think it does not introduce a bias, then we handle the $\hat{y}_{imp,k}$ as if they were real values and we admit therefore that $\hat{Y}_\bullet$, the *regression imputed Horvitz-Thompson estimator*, is an estimator of $\hat{Y} = \sum_s w_k y_k$ that is in itself the unbiased Horvitz-Thompson estimator of the population total $Y = \sum_U y_k$.

The variance due to the sampling design $p(s)$ of the total Y can be expressed by the quadratic form

$$V_p(\hat{Y}) = \sum_{k \in U} \sum_{\ell \in U} \Delta_{k\ell} y_k y_\ell \tag{4.2}$$

where $\Delta_{k\ell} = \pi_{k\ell}/(\pi_k\pi_\ell) - 1$, $\pi_{k\ell}$ being the probability that k and ℓ are included in the sample, and $\pi_{kk} = \pi_k$. For us $\pi_k = 1/w_k$. Without nonresponse and therefore without imputation, one can estimate the variance without bias through the standard formula

$$\hat{V}_p(\hat{Y}) = \sum_{k\in s}\sum_{\ell\in s} A_{k\ell} y_k y_\ell, \tag{4.3}$$

where $A_{k\ell} = \Delta_{k\ell}/\pi_{k\ell} = 1/(\pi_k\pi_\ell) - 1/\pi_{k\ell}$.

If one handles the imputed values as true values, we are inclined to estimate (4.3) by

$$\hat{V}_p(\hat{Y}_\bullet) = \sum_{k\in s}\sum_{\ell\in s} A_{k\ell} y_{\bullet k} y_{\bullet \ell}, \tag{4.4}$$

but this will tendentially underestimate the sought variance. Deville & Särndal (1994) show that, under some assumptions recalled hereafter, additional terms computable on the base of the observed and imputed data must be added to (4.4) in order to estimate (4.3) without bias.

One wishes that $\hat{Y}_\bullet$ is an unbiased estimator of Y, i.e. that the expectation with respect to the imputation model ξ, the sampling design $p(s)$ and the nonresponse mechanism $q(r|s)$ is zero: $\mathrm{E}_\xi\mathrm{E}_p\mathrm{E}_q(\hat{Y}_\bullet - Y|s,r) = 0$. In order to have this, Deville & Särndal (1994) must formulate some assumptions:

A1 The imputation model ξ is correct, i.e. that the expectation of the imputation error is zero: $\mathrm{E}_\xi(\hat{Y}_\bullet - \hat{Y}) = 0$.

A2 The variances of the model errors ξ can be written $\mathrm{E}_\xi(\varepsilon_k^2) = \sigma_k^2 = \sigma^2\mathbf{x}_{k\cdot}\boldsymbol{\lambda}$. In our case (homoscedasticity), this assumption is satisfied with choosing $\boldsymbol{\lambda} = (1, 0, ..., 0)'$ and setting $x_{k1} = 1$, for all k, in other words, one admits a constant term in the regression model.

A3 The nonresponse mechanism of y denoted by $q(r|s,\mathbf{X})$ is ignorable (MAR, *missing at random*, in the sense of Little & Rubin, 2002), which means that it may depend on the sample and on some of the x_{kj} but not on y_k.

Under these assumptions, Deville & Särndal (1994) can permute the expectations' order $\mathrm{E}_\xi\mathrm{E}_p\mathrm{E}_q$ en $\mathrm{E}_p\mathrm{E}_q\mathrm{E}_\xi$, one then finds that the bias under the model ξ, the sampling design p and the nonresponse q is null:

$$\mathrm{B}_{\xi pq}(\hat{Y}_\bullet) = 0. \tag{4.5}$$

This implies

$$\begin{aligned}\mathrm{E}_\xi\mathrm{E}_p\mathrm{E}_q(\hat{Y}_\bullet) &= \mathrm{E}_\xi\mathrm{E}_p\mathrm{E}_q(Y)\\ &= \mathrm{E}_\xi(Y) = \mathrm{E}_\xi\left(\sum_{k\in U}\mathbf{x}_{k\cdot}\boldsymbol{\beta} + \varepsilon_k\right) = \sum_{k\in U}\mathbf{x}_{k\cdot}\boldsymbol{\beta} = Y_{imp} \neq Y\end{aligned} \tag{4.6}$$

since Y is random under ξ. These authors can then decompose the total variance we seek to estimate of the difference with regards to the true total

of y in one component due to the sampling and one due to the imputation:

$$\begin{aligned} & V_{tot}(\hat{Y}_\bullet - Y) \\ &= \mathrm{E}_\xi \mathrm{E}_p \mathrm{E}_q (\hat{Y} - Y)^2 + \mathrm{E}_\xi \mathrm{E}_p \mathrm{E}_q (\hat{Y}_\bullet - \hat{Y})^2 + 2\mathrm{E}_\xi \mathrm{E}_p \mathrm{E}_q [(\hat{Y} - Y)(\hat{Y}_\bullet - \hat{Y})] \\ &= V_{sam} + V_{imp}. \end{aligned} \tag{4.7}$$

Deville & Särndal (1994) then construct estimators of V_{sam} and V_{imp} that we shall recall hereafter expressed in our notations and applied to our framework.

4.3.2 Sampling variance: $\hat{V}_{sam}$

Deville & Särndal (1994) demonstrate that one can estimate V_{sam} by the variance of the total of the variable ignoring impution plus a corrective term which is proportional to $\hat{\sigma}^2$, see (4.8). To see that, first note that:

$$V_{sam} = \mathrm{E}_\xi \mathrm{E}_p \mathrm{E}_q (\hat{Y} - Y)^2 = \mathrm{E}_\xi \mathrm{E}_p (\hat{Y} - Y)^2 = \mathrm{E}_\xi V_p (\hat{Y} - Y) = \mathrm{E}_\xi V_p (\hat{Y}),$$

since Y is fixed under p. Let us find a corrective term $C := \mathrm{E}_\xi[\hat{V}_p(\hat{Y}) - \hat{V}_p(\hat{Y}_\bullet)]$ compensating for the underestimation done when computing the sampling variance of the imputed variable and estimate it by $\hat{C}$ verifying $\mathrm{E}_\xi(\hat{C}) = C$. Thereby

$$\mathrm{E}_\xi[\hat{V}_p(\hat{Y}_\bullet) + \hat{C}] = \mathrm{E}_\xi[\hat{V}_p(\hat{Y}_\bullet) + \hat{V}_p(\hat{Y}) - \hat{V}_p(\hat{Y}_\bullet)] = E_\xi \hat{V}_p(\hat{Y}) = \hat{V}_{sam},$$

so $\hat{V}_p(\hat{Y}_\bullet) + \hat{C}$ will be an unbiased estimator of $\hat{V}_{sam}$. Furthermore, Deville & Särndal (1994) show that $\hat{C}$ can be written as $\hat{\sigma}^2(Q_s - Q_r)$. Q_s and Q_r being expressions entirely computable based on the information at our disposal on the sample and the weights defined below. Thus the sampling variance will be estimated by:

$$\hat{V}_{sam} = \hat{V}_p(\hat{Y}_\bullet) + \hat{C}, \tag{4.8}$$

the first term on the right side of the equality being the the estimated variance ignoring the imputation.

Development of the expression for C

Proposition 4.1 (from Deville & Särndal, 1994) *The corrective term for the sampling variance defined by*

$$C = \mathrm{E}_\xi[\hat{V}_p(\hat{Y}) - \hat{V}_p(\hat{Y}_\bullet)]$$

can be written

$$C = \sigma^2(Q_s - Q_r)$$

where $Q_s = \sum_{k\in s} A_{kk} - \sum_{k\in s}\sum_{\ell\in s} A_{k\ell}\mathbf{x}_{k\cdot}\boldsymbol{T}^{-1}\mathbf{x}'_{l\cdot}$ *and* $Q_r = \sum_{k\in r} A_{kk} + \sum_{k\in r}\sum_{\ell\in r} A_{k\ell}\mathbf{x}_{k\cdot}\boldsymbol{T}^{-1}\mathbf{x}'_{l\cdot}$, σ^2 *being the variance of the residuals of model* ξ *and* $\boldsymbol{T} = \sum_r \mathbf{x}'_{k\cdot}\mathbf{x}_{k\cdot} = \boldsymbol{X}'_r\boldsymbol{X}_r$, $J \times J$ *matrix.*

One obtains an estimation of C by estimating σ^2 (see next section):

$$\hat{C} = \hat{\sigma}^2(Q_s - Q_r). \tag{4.9}$$

Q_s and Q_r are entirely computable based on the auxiliary information available on the sample and the weights. If there are no nonrespondents, then $r = s$ and $C = 0$. According to the sampling design used, for implementation, it may be necessary to perform a spectral decomposition of the matrix $\mathbf{T}$ in order to be able to apply already programmed formulae for $V_p(\cdot)$ taking the sampling design into account.

Estimation of σ^2

We wish to re-use programmed formulae for $V_p(\cdot)$ taking the used sampling design p into account to estimate σ^2. Let $e_{\bullet k}$ be the pseudo-residuals defined as follow:

$$e_{\bullet k} = \begin{cases} e_k = y_k - \hat{y}_{imp,k} & \text{si } k \in r \\ 0 & \text{si } k \in o, \end{cases}$$

Proposition 4.2 (from Deville & Särndal, 1994) *An unbiased estimator of σ^2 under the model ξ is given by*

$$\hat{\sigma}^2 = \hat{V}_p(e_{\bullet s})/Q_r.$$

Sampling variance, final formula

By resuming (4.8), (4.9) and from the Proposition 4.2, Deville & Särndal (1994) obtain that:

$$\hat{V}_{sam} = \hat{V}_p(\hat{Y}_\bullet) + \left(\frac{Q_s}{Q_r} - 1\right)\hat{V}_p(e_{\bullet s}). \tag{4.10}$$

Note that without nonresponse (thus no imputation), $s = r$ and $o = \emptyset$, one falls back to $\hat{V}_{sam} = \hat{V}_p(\hat{Y})$ as it should be.

4.3.3 Imputation variance: $\hat{V}_{imp}$

The imputation variance is defined by, see (4.7):

$$V_{imp\xi} = \mathrm{E}_\xi\{(\hat{Y}_\bullet - \hat{Y})^2 + 2(\hat{Y}_\bullet - \hat{Y})(\hat{Y} - Y)|s, r\}. \tag{4.11}$$

Lemma 4.1 (from Deville & Särndal, 1994) *The first term of* (4.11) *is given by:*

$$\mathrm{E}_\xi\{(\hat{Y}_\bullet - \hat{Y})^2|s, r\} = \sigma^2(\boldsymbol{t}_{o,w}\boldsymbol{T}^{-1}\boldsymbol{t}'_{o,w} + \sum_{k\in o} w_k^2),$$

where $\boldsymbol{t}_{o,w} = \sum_{k\in o} w_k \mathbf{x}_{k\cdot}$ and w_k being the weights of the observations.

Lemma 4.2 (from Deville & Särndal, 1994) *The second term of* (4.11) *is given by:*

$$\mathrm{E}_\xi\left\{(\hat{Y}_\bullet - \hat{Y})(\hat{Y} - Y)|s, r\right\} = \boldsymbol{t}_{o,w}\boldsymbol{T}^{-1}\boldsymbol{t}'_{r,w} - \sum_{k\in o} w_k^2\sigma^2,$$

where $\boldsymbol{t}_{r,w} = \sum_{k\in r} w_k \mathbf{x}_{k\cdot}$.

Putting these two results together, the imputation variance can be expressed as follows

$$\begin{aligned} V_{imp\xi} &= \mathrm{E}_\xi\{(\hat{Y}_\bullet - \hat{Y})^2 + 2(\hat{Y}_\bullet - \hat{Y})(\hat{Y} - Y)|s,r\} \\ &= \sigma^2(\mathbf{t}_{o,w}\mathbf{T}^{-1}\mathbf{t}'_{o,w} + \sum_{k\in o} w_k^2) + 2\mathbf{t}_{o,w}\mathbf{T}^{-1}\mathbf{t}'_{r,w} - 2\sum_{k\in o} w_k^2\sigma^2 \\ &= \sigma^2\left[\mathbf{t}_{o,w}\mathbf{T}^{-1}(\mathbf{t}'_{s,w} + \mathbf{t}'_{r,w}) - \sum_{k\in o} w_k^2\right], \end{aligned}$$

where $\mathbf{t}_{s,w} = \sum_{k\in s} w_k\mathbf{x}_{k\cdot}$ (line-vector in our notations). Deville & Särndal (1994) show that the following expression is equivalent:

$$V_{imp\xi} = \sigma^2\left[\mathbf{t}_{o,w}\mathbf{T}^{-1}(\mathbf{t}'_{s,w} + \mathbf{t}'_{r,w}) - \mathbf{t}_{o,w^2}\mathbf{T}^{-1}\mathbf{t}'_{r,1}\right].$$

where $\mathbf{t}_{r,1} = \sum_{k\in r}\mathbf{x}_{k\cdot}$ and $\mathbf{t}_{o,w^2} = \sum_{k\in o} w_k^2\mathbf{x}_{k\cdot}$. Thus we shall estimate the imputation variance (which cancels itself out, as it should be, if $s = r$ and $o = \emptyset$) by:

$$\hat{V}_{imp} = \hat{\sigma}^2\left[\mathbf{t}_{o,w}\mathbf{T}^{-1}(\mathbf{t}'_{s,w} + \mathbf{t}'_{r,w}) - \sum_{k\in o} w_k^2\right] \tag{4.12}$$

with $\hat{\sigma}^2$ given by Proposition 4.2. Note that, except the estimated variance of the estimated total of the pseudo-residuals necessary in the estimation of $\hat{\sigma}^2$, see Proposition 4.2. The rest of the expression depends only on the explanatory variables of the regression and on the weights w_k and not on the values of y. Furthermore, if all the weights were equal, $w_k = w$, expression (4.12) could be simplified to:

$$\hat{V}_{imp} = \hat{\sigma}^2 w^2 \sum_{k\in o}\mathbf{x}_{k\cdot}\mathbf{T}^{-1}\sum_{k\in s}\mathbf{x}'_{k\cdot}, \tag{4.13}$$

which is obviously always ≥ 0. We have implemented this last expression in our simulations presented below, using a simple random without replacement sampling design.

4.3.4 Total variance: $\hat{V}_{tot}$

More complex expressions of formulae (4.10) and (4.12) for the two components of the total variance can be derived if one wishes to consider the fact that weights would have been used in the regression model ξ. However, according to Deville & Särndal (1994), the variance estimation method shouldn't be too sensitive to the implication or not of weights in the computations.

Bringing together (4.10) and (4.12), the total variance can thus be estimated by

$$\begin{aligned} \hat{V}_{tot} &= \hat{V}_{sam} + \hat{V}_{imp} \\ &= \hat{V}_p(\hat{Y}_\bullet) + \frac{\hat{V}_p(e_{\bullet s})}{Q_r}\left\{Q_s - Q_r + \left[\mathbf{t}'_{o,w}\mathbf{T}^{-1}(\mathbf{t}'_{s,w} + \mathbf{t}'_{r,w}) - \sum_{k\in o} w_k^2\right]\right\}. \end{aligned} \tag{4.14}$$

All the elements of (4.14) are computable with the same information than that necessary to perform the imputation. This is why this formula is very convenient and appealing.

Case of simple random sampling without replacement

For a sampling fraction of $\pi = n/N$,

$$Q_s = \sum_{k\in s} A_{kk} - \sum_{k\in s}\sum_{\ell\in s} A_{k\ell}\mathbf{x}_{k\cdot}\mathbf{T}^{-1}\mathbf{x}'_{l\cdot}$$

where, as a reminder, $\mathbf{T} = \mathbf{X}'_r\mathbf{X}_r$. $A_{k\ell}$ defined in (4.3) becomes

$$A^{sas}_{k\ell} = \begin{cases} -\frac{N(N-n)}{n^2(n-1)} & \text{si } k \neq \ell \\ \frac{N(N-n)}{n^2} & \text{si } k = \ell, \end{cases} \tag{4.15}$$

Thus the quantities Q_s et Q_r are computable without spectral decomposition of $\mathbf{T}$.

$$\begin{aligned} Q^{srs}_s &= \frac{N(N-n)}{n^2}(n - \sum_{k\in s}\mathbf{x}_{k\cdot}\mathbf{T}^{-1}\mathbf{x}'_{k\cdot}) + \frac{N(N-n)}{n^2(n-1)}\sum_{k\in s}\sum_{\ell\in s,\,\ell\neq k}\mathbf{x}_{k\cdot}\mathbf{T}^{-1}\mathbf{x}'_{l\cdot} \\ Q^{srs}_r &= \frac{N(N-n)}{n^2}(n_r - \sum_{k\in r}\mathbf{x}_{k\cdot}\mathbf{T}^{-1}\mathbf{x}'_{k\cdot}) + \frac{N(N-n)}{n^2(n-1)}\sum_{k\in r}\sum_{\ell\in r,\,\ell\neq k}\mathbf{x}_{k\cdot}\mathbf{T}^{-1}\mathbf{x}'_{l\cdot} \end{aligned}$$

By resuming (4.13) and (4.14), the total variance is approached by:

$$\hat{V}^{srs}_{tot} = \hat{V}_p(\hat{Y}_\bullet) + \frac{\hat{V}_p(e_{\bullet s})}{Q^{srs}_r}\left[Q^{srs}_s - Q^{srs}_r + \frac{N^2}{n^2}\sum_{k\in o}\mathbf{x}_{k\cdot}\mathbf{T}^{-1}\sum_{k\in s}\mathbf{x}'_{k\cdot}\right]. \tag{4.16}$$

4.4 Variance of inequality indicators

In our scenario, we are not interested in the variance of the total of $y_{\bullet k}$ on s estimated by (4.14) but rather in that of the total of an estimated linearized variable $z_{\bullet k}$ as an approximation of the variance of the corresponding inequality indicator, with $z_{\bullet k}$ obtained from $y_{\bullet k}$. The goal is now to estimate $\hat{V}_{tot}(\hat{Z}_\bullet - Z)$ and not no longer $\hat{V}_{tot}(\hat{Y}_\bullet - Y)$. In particular, the linearized variables we are dealing with are those of quantiles or indicators of poverty, inequality and social exclusion. We compute these linearized variables as described in Graf & Tillé (2014) where their expressions are also recalled.

One might be tempted to apply a plug-in system (named "D2"-variant in the simulations of Section 4.6) consisting in mechanically replacing $\hat{V}_p(\hat{Y}_\bullet)$ by $\hat{V}_p(\hat{Z}_\bullet) = \hat{V}_p(\sum_s w_k z_{\bullet k})$ in (4.14), and also $\hat{V}_p(e_{\bullet s})$ by $\hat{V}_p(\sum_s w_k\zeta_{\bullet k})$ with $\zeta_{\bullet k} = z_k - \hat{z}_{imp,k}$ for $k \in r$ and $\zeta_{\bullet k} = 0$ if $k \in o$:

$$\hat{V}_{tot} = \hat{V}_p(\hat{Z}_\bullet) + \frac{\hat{V}_p(\zeta_{\bullet s})}{Q_r}\left\{Q_s - Q_r + \left[\mathbf{t}_{o,w}\mathbf{T}^{-1}(\mathbf{t}'_{s,w} + \mathbf{t}'_{r,w}) - \sum_{k\in o} w_k^2\right]\right\}. \tag{4.17}$$

However, by doing so, one would implicitly assume that the stage of imputation and the one of calculating the linearized variable are interchangeable. In other words, to first impute $\hat{y}_{imp,k}$, $k \in o$, and than linearize to obtain $z_{\bullet k}$, $k \in s$, would lead to the same result as first calculating the linearized variable for the respondents, and then imputing values of the linearized variable for the nonrespondents. This assumption does not seem to be realistic since the linearized variables for the considered indicators are non linear functions of the y_k or $y_{\bullet k}$. The simulations presented in Section 4.6 verify this and show that the plug-in system leads to over- or under-estimation of the variance of the considered indicators. However, the method seems to be well adapted to the case of variance estimation under imputation of the total Y (see Table 4.2), which was the goal of Deville & Särndal (1994)'s article. The same simulations also show that ignoring the imputation in the calculation of variance through linearization almost always leads to under-estimation of the variance, sometimes severe.

Thus we need to conduct more advanced thinking to apply Deville & Särndal (1992)'s method to complex non-linear statistics. This article proposes a solution.

As we shall see hereafter, the bias generated by regression imputation may become serious and problematic. Indeed, even if the (already strong!) assumption **A1** is satisfied for the total Y, that is that $E_\xi(\hat{Y}_\bullet - Y) = 0$, the property is no longer satisfied for the statistics we consider, nor for their linearized variables as we show in Section 4.5.2. We shall return to the bias problem in the results' analysis of the simulations in Section 4.7.3.

4.4.1 Degree of a functional

In this section, we shall recall a few definitions formulated by Deville (1999b) leading to the important property that the sum of the linearized variables we are considering is always zero.

Definition 4.1 *A functional $T(M)$ is said to be **homogeneous**, if there is a positive real number α depending on T such that*

$$T(tM) = t^\alpha T(M) \ \forall t \in \mathbb{R}^+.$$

*α is the **degree** of the functional.*

A total Y is a homogeneous functional of degree 1, but a mean, a ratio, a quantile, as well as the functionals of all the indicators which we are interested in, are of degree 0.

Definition 4.2 *One says that a statistic S is of degree α if $N^{-\alpha}S$ tends to a finite limit as $N \to \infty$.*

Definition 4.3 *A statistic S of degree α can be **linearized** if there is a synthetic variable z_k (called the linearized variable of S) such that the variance of $\hat{Z}$ is equivalent to the one of S in the sense that $\sqrt{n}(N^{-\alpha}S - N^{-1}\hat{Z}) = O_p[f(n)]$, with $f(n) \to 0$ ($N \to \infty$, $n \to \infty$).*

Result: If T is homogeneous of degree α we have:

$$\int IT(M,x)d(M)x = \sum_{k\in U} IT(M,x) = \alpha\, T(M).$$

The particular case of $\alpha = 0$ means that every homogeneous functional of degree o has an influence functuon (linearized variable) summing up to o.

4.5 Total variance for quantiles

We will build an estimator of the total variance for the regression imputed quantile. Next we will detail how we obtain the final expression:

$$\hat{V}_{tot}^{Q_\alpha} = \underbrace{\hat{V}_p(\hat{Z}_\bullet) + \hat{C}^{Q_\alpha}}_{V_{sam}} + \underbrace{V_{imp\xi}}_{V_{imp}} - \underbrace{\hat{B}^2_{\xi pq}}_{bias^2}$$

4.5.1 Linearized variable of a quantile

The estimated linearized variable of a quantile Q_α is given by (Deville, 1999b; Osier, 2009; Graf & Tillé, 2014):

$$z_k^{Q_\alpha} = -\frac{1}{f(\hat{Q}_\alpha)}\frac{1}{\hat{N}}[\mathbf{1}_{[y_k\le\hat{Q}_\alpha]} - \alpha],$$

where the α-order quantile Q_α is defined by (see Hyndman & Fan, 1996; Graf & Tillé, 2014):

$$\widehat{Q}_\alpha = y_{\bullet k-1} + (y_{\bullet k} - y_{\bullet k-1})\left(\frac{\alpha\hat{N} - \hat{N}_{k-1}}{w_k}\right),$$

with $\widehat{N}_k = \sum_{\ell\in s} w_\ell \mathbf{1}_{[y_{\bullet\ell}\le y_{\bullet k}]}$ and $f(\cdot)$ is the income density function which has to be estimated as well. Note that, in (4.18), we have left out the hat on z_k although it is an estimated linearized variable based on the sample, we shall keep the hat-notation to designate the value

$$\hat{z}_{imp,k}^{Q_\alpha} = -\frac{1}{f(\hat{Q}_\alpha)}\frac{1}{\hat{N}}[\mathbf{1}_{[\hat{y}_{imp,k}\le\hat{Q}_\alpha]} - \alpha],$$

which we will abusively call *imputed value*. It corresponds to the value of the linearized variable computed out of an imputed value of y.

4.5.2 Sampling variance: $\hat{V}_{sam}$

We wish to adapt the calculations done for the total of y in (4.8) to the case where y is replaced by the linearized variable of a quantile:

$$\hat{V}_{sam}(\hat{Z}) = \hat{V}_p(\hat{Z}_\bullet) + \hat{C}^{Q_\alpha} \tag{4.18}$$

This requires us to develop an expression to estimate the corrective term C^{Q_α}. Similarily to Proposition 4.1, the latter is defined by:

$$\begin{aligned} C^{Q_\alpha} &= \mathrm{E}_\xi[V_p(\hat{Z}) - \hat{V}_p(\hat{Z}_\bullet)] \\ &= \mathrm{E}_\xi\left[\sum_{k\in o}\sum_{\ell\in o} A_{k\ell}(z_k z_\ell - \hat{z}_k\hat{z}_\ell) + 2\sum_{k\in r}\sum_{\ell\in o} A_{k\ell} z_k(z_\ell - \hat{z}_\ell)\right] \end{aligned}$$

To achieve this, we see that we need to develop expressions for $\mathrm{E}_{\tilde{\xi}}(z_k z_\ell)$, $\mathrm{E}_{\tilde{\xi}}(z_k \hat{z}_\ell)$ and $\mathrm{E}_{\tilde{\xi}}(\hat{z}_k \hat{z}_\ell)$.

Proposition 4.3 *Neglecting the fact that $\hat{Q}_\alpha$ is also random and assuming the ϵ_k iid $\mathcal{N}(0, \sigma^2)$,*

$$\mathrm{E}_{\tilde{\xi}}(z_k^{Q_\alpha}) = -\frac{1}{f(\hat{Q}_\alpha)} \frac{1}{\hat{N}} [\Phi_k - \alpha], \tag{4.19}$$

$$\mathrm{E}_{\tilde{\xi}}(z_k^{Q_\alpha} z_l^{Q_\alpha}) = \frac{1}{f^2(\hat{Q}_\alpha)} \frac{1}{\hat{N}^2} [\Phi_{kl} - \alpha(\Phi_k + \Phi_\ell) + \alpha^2], \tag{4.20}$$

$$\mathrm{E}_{\tilde{\xi}}(\hat{z}_k^{Q_\alpha}) = -\frac{1}{f(\hat{Q}_\alpha)} \frac{1}{\hat{N}} \left[M_k - \alpha\right], \tag{4.21}$$

$$\mathrm{E}_{\tilde{\xi}}(\hat{z}_k^{Q_\alpha} \hat{z}_l^{Q_\alpha}) = \frac{1}{f^2(\hat{Q}_\alpha)} \frac{1}{\hat{N}^2} \left[M_{kl} - \alpha(M_k + M_l) + \alpha^2\right], \tag{4.22}$$

$$\mathrm{E}_{\tilde{\xi}}(z_k^{Q_\alpha} \hat{z}_\ell^{Q_\alpha}) = \frac{1}{f^2(\hat{Q}_\alpha)} \frac{1}{\hat{N}^2} \left[\Phi M_{k\ell} - \alpha(\Phi_k + M_\ell) + \alpha^2\right]. \tag{4.23}$$

With

$$\Phi_k = \Phi\left(\frac{\hat{Q}_\alpha - \mathbf{x}_{k\cdot}\boldsymbol{\beta}}{\sigma}\right),$$

Φ is the distribution function of the standard normal distribution, $\Phi_{k\ell} = \Phi_k \Phi_\ell$ if $k \neq \ell$, $\Phi_{kk} = \Phi_k$,

$$M_k = \Phi\left(\frac{\hat{Q}_\alpha - \mathbf{x}_{k\cdot}\boldsymbol{\beta}}{\sigma\sqrt{\mathbf{x}_{k\cdot}\boldsymbol{T}^{-1}\mathbf{x}'_{k\cdot}}}\right),\ M_{kl} = \Phi^{2d,A}\left(\frac{\hat{Q}_\alpha - \mathbf{x}_{k\cdot}\boldsymbol{\beta}}{\sigma\sqrt{\mathbf{x}_{k\cdot}\boldsymbol{T}^{-1}\mathbf{x}'_{k\cdot}}}, \frac{\hat{Q}_\alpha - \mathbf{x}_{l\cdot}\boldsymbol{\beta}}{\sigma\sqrt{\mathbf{x}_{l\cdot}\boldsymbol{T}^{-1}\mathbf{x}'_{l\cdot}}}\right),$$

$M_{kk} = M_k$, $\Phi^{2d,A}$ is the distribution function of the centred two dimensional normal distribution with variance $\left((1, \hat{\rho}_{kl}^M)', (\hat{\rho}_{kl}^M, 1)'\right)$ where

$$\hat{\rho}_{kl}^M = \frac{\mathbf{x}_{k\cdot}\boldsymbol{T}^{-1}\mathbf{x}'_{l\cdot}}{\sqrt{(\mathbf{x}_{k\cdot}\boldsymbol{T}^{-1}\mathbf{x}'_{k\cdot})(\mathbf{x}_{l\cdot}\boldsymbol{T}^{-1}\mathbf{x}'_{l\cdot})}},\ \Phi M_{kl} = \Phi^{2d,B}\left(\frac{\hat{Q}_\alpha - \mathbf{x}_{k\cdot}\boldsymbol{\beta}}{\sigma}, \frac{\hat{Q}_\alpha - \mathbf{x}_{l\cdot}\boldsymbol{\beta}}{\sigma\sqrt{\mathbf{x}_{l\cdot}\boldsymbol{T}^{-1}\mathbf{x}'_{l\cdot}}}\right),$$

$\Phi^{2d,B}$ is the distribution function of the centred two dimensional normal distribution with variance $\left((1, \hat{\rho}_{kl}^{\Phi M})', (\hat{\rho}_{kl}^{\Phi M}, 1)'\right)$ where

$$\hat{\rho}_{kl}^{\Phi M} = \frac{\mathbf{x}_{k\cdot}\boldsymbol{T}^{-1}\mathbf{x}'_{l\cdot}}{\sqrt{\mathbf{x}_{l\cdot}\boldsymbol{T}^{-1}\mathbf{x}'_{l\cdot}}}.$$

To illustrate the reasoning here is the proof of the first statement, (4.19), of Proposition 4.3:

Proof. If we neglect the fact that $\hat{Q}_\alpha$ is also random, it can be extracted from the expectation. Doing so it is assumed that the estimated quantile based on the sample is close to the real quantile of the population. This is acceptable only if the sample is of sufficient size to ensure the stability

of the estimated quantile and if assumption **A3** (ignorable nonresponse) is valid.

$$\begin{aligned}
\mathrm{E}_{\xi}(z_k^{Q_\alpha}) &= -\frac{1}{f(\hat{Q}_\alpha)}\frac{1}{\hat{N}}\left[E_{\xi}(\mathbf{1}_{[y_k\leq\hat{Q}_\alpha]})-\alpha\right] \\
&= -\frac{1}{f(\hat{Q}_\alpha)}\frac{1}{\hat{N}}\left[P_{\xi}(y_k\leq\hat{Q}_\alpha)-\alpha\right] \\
&= -\frac{1}{f(\hat{Q}_\alpha)}\frac{1}{\hat{N}}\left[P_{\xi}(\mathbf{x}_{k\cdot}\boldsymbol{\beta}+\varepsilon_k\leq\hat{Q}_\alpha)-\alpha\right] \\
&= -\frac{1}{f(\hat{Q}_\alpha)}\frac{1}{\hat{N}}\left[P_{\xi}\left(\frac{\varepsilon_k}{\sigma}\leq\frac{\hat{Q}_\alpha-\mathbf{x}_{k\cdot}\boldsymbol{\beta}}{\sigma}\right)-\alpha\right] \\
&= -\frac{1}{f(\hat{Q}_\alpha)}\frac{1}{\hat{N}}\left[\Phi\left(\frac{\hat{Q}_\alpha-\mathbf{x}_{k\cdot}\boldsymbol{\beta}}{\sigma}\right)-\alpha\right] \\
&= -\frac{1}{f(\hat{Q}_\alpha)}\frac{1}{\hat{N}}[\Phi_k-\alpha],
\end{aligned}$$

with

$$\Phi_k := \Phi\left(\frac{\hat{Q}_\alpha-\mathbf{x}_{k\cdot}\boldsymbol{\beta}}{\sigma}\right).$$

□

The proofs of the other statements of Proposition 4.3 are detailed in the Appendix A. Using the results of Proposition 4.3, we obtain the following expression for the corrective term of the variance due to sampling in the case of a quantile:

$$\begin{aligned}
C^{Q_\alpha} &= \mathrm{E}_{\xi}[V_p(\hat{Z})-\hat{V}_p(\hat{Z}_\bullet)] \\
&= \mathrm{E}_{\xi}\left[\sum_{k\in o}\sum_{\ell\in o}A_{k\ell}(z_kz_\ell-\hat{z}_k\hat{z}_\ell)+2\sum_{k\in r}\sum_{\ell\in o}A_{k\ell}z_k(z_\ell-\hat{z}_\ell)\right]
\end{aligned}$$

From (4.20),(4.22), (4.23), we have:

$$\begin{aligned}
C^{Q_\alpha} &= \frac{1}{f^2(\hat{Q}_\alpha)}\frac{1}{\hat{N}^2}\left\{\sum_{k\in o}\sum_{\ell\in o}A_{k\ell}[\Phi_{kl}-\alpha(\Phi_k+\Phi_\ell)+\alpha^2-M_{kl}+\alpha(M_k+M_l)-\alpha^2]\right\} \\
&+ 2\frac{1}{f^2(\hat{Q}_\alpha)}\frac{1}{\hat{N}^2}\left\{\sum_{k\in r}\sum_{\ell\in o}A_{k\ell}[\Phi_{kl}-\alpha(\Phi_k+\Phi_\ell)+\alpha^2-\Phi M_{kl}+\alpha(\Phi_k+M_l)-\alpha^2]\right\} \\
&= \frac{1}{f^2(\hat{Q}_\alpha)}\frac{1}{\hat{N}^2}\left\{\sum_{k\in o}\sum_{\ell\in o}A_{k\ell}[\Phi_{kl}-M_{kl}-\alpha(\Phi_k+\Phi_\ell-M_k-M_l)]\right. \\
&+ \left.2\sum_{k\in r}\sum_{\ell\in o}A_{k\ell}[\Phi_{kl}-\Phi M_{kl}-\alpha(\Phi_\ell-M_l)]\right\} \qquad (4.24)
\end{aligned}$$

The sampling variance of the total of the linearized variable of the quantile α is thus given by the sampling design variance of the linearized variable obtained from partially imputed data plus the corrective term C^{Q_α} given by (4.24).

$\xi pq-$**bias of** $\hat{Z}_\bullet^{Q_\alpha}$

Although the sum on the population of the linearized variable is zero (see paragraph 4.4.1), contrary to what was observed in (4.5) for the imputed total $\hat{Y}_\bullet$, its bias under the model ξ, the design p and the nonresponse q does not vanish. Indeed, as for the bias of $\hat{Y}_\bullet$, also with reversing the expectations order (the population size and thereby the sums on the expectations are all finite):

$$\begin{aligned} \mathrm{B}_{\xi pq}(\hat{Z}_\bullet^{Q_\alpha}) &= \mathrm{E}_\xi \mathrm{E}_p \mathrm{E}_q (\hat{Z}_\bullet - Z) \\ &= \mathrm{E}_p \mathrm{E}_q \mathrm{E}_\xi (\hat{Z}_\bullet - Z) \\ &= \mathrm{E}_p \mathrm{E}_q \mathrm{E}_\xi \left(\sum_{k \in o} w_k \hat{z}_k + \sum_{k \in r} w_k z_k - \sum_{k \in U} z_k \right) \end{aligned}$$

From (4.19) and (4.21), one has:

$$\begin{aligned} \mathrm{B}_{\xi pq}(\hat{Z}_\bullet^{Q_\alpha}) &= \frac{-1}{f(\hat{Q}_\alpha)\hat{N}} \left[\sum_{k \in o} w_k (M_k - \alpha) + \sum_{k \in r} w_k (\Phi_k - \alpha) - \sum_{k \in U} (\Phi_k - \alpha) \right] \\ &= \frac{1}{f(\hat{Q}_\alpha)\hat{N}} \left[\sum_{k \in U} \Phi_k - \sum_{k \in o} w_k M_k - \sum_{k \in r} w_k \Phi_k \right]. \end{aligned}$$

Therefore the $\xi pq-$bias of $\hat{Z}_\bullet^{Q_\alpha}$ is nonzero and in principle not computable, but it can be estimated by replacing $\boldsymbol{\beta}$ with $\hat{\boldsymbol{\beta}}$ and $\sum_{k \in U} \Phi_k$ with $\sum_{k \in s} w_k \Phi_k$; one then obtains:

$$\hat{\mathrm{B}}_{\xi pq}(\hat{Z}_\bullet^{Q_\alpha}) \approx \frac{1}{f(\hat{Q}_\alpha)\hat{N}} \left[\sum_{k \in o} w_k (\Phi_k - M_k) \right]. \tag{4.25}$$

For the median, $\alpha = 0.5$, it is generally small compared to the sampling variance and the imputation variance in the case of simple random sampling, even squared. However, the more the considered quantile moves away from the median, the more it increases. Intuitively, one understands it, remembering the definitions of Φ_k et M_k:

$$\Phi_k := \Phi\left(\frac{\hat{Q}_\alpha - \mathbf{x}_{k\cdot}\boldsymbol{\beta}}{\sigma} \right), \quad M_k := \Phi\left(\frac{\hat{Q}_\alpha - \mathbf{x}_{k\cdot}\boldsymbol{\beta}}{\sigma\sqrt{\mathbf{x}_{k\cdot}\mathbf{T}^{-1}\mathbf{x}'_{k\cdot}}} \right)$$

The quantities $\tilde{h}_k = \mathbf{x}_{k\cdot}\mathbf{T}^{-1}\mathbf{x}'_{k\cdot}$ do not vary much for $k \in o$, for the simulations conducted $1/\sqrt{\tilde{h}_k} \approx 10$, this implies roughly, that the M_k take the values 0 or 1. The quantities $\tilde{h}_k$ would correspond to the diagonal elements of the *hat matrix* if $k \in r$ (because $\mathbf{T} = \mathbf{X}'_r\mathbf{X}_r$). But since $k \in o$ we cannot interpret them as leverages even if they are of the same magnitude. It is thus seen that, in the case of the median, if the imputed values' median $\hat{y}_{imp,k} = \mathbf{x}_{k\cdot}\hat{\boldsymbol{\beta}}$ is close to the one of the respondents $\hat{Q}_{0.5}$ - this will not happen if the nonresponse is not ignorable, that is to say, the assumption **A3** is not valid, or if the regression model does not fit the data well - then, if all the weights w_k are also equal, the two sums of (4.25) will both

approach $N_o/2$ because $\Phi(x) + \Phi(-x) = 1$, $N_o = \sum_o w_k$ and thus the bias is nearly zero. Under the same conditions, for a higher quantile than the median, the sign of the argument for Φ_k or M_k will more often be positive and therefore the bias moves away from zero. For a lower quantile than the median, the sign of the argument for Φ_k or M_k will more often be negative and therefore the bias again moves away from zero. Table 4.3 confirms this for the quantiles $\alpha = (0.2, 0.5, 0.8)$ obtained through the realized simulations and are commented below. Further observe that the sum of (4.25) is on the set of nonrespondents. A high nonresponse rate or extreme weights naturally increase the risk of bias.

We must therefore take account of this bias in the estimation of total variance:

$$\begin{aligned} V_{tot}(\hat{Z}_\bullet^{Q_\alpha} - Z^{Q_\alpha}) &= \mathrm{E}_\xi \mathrm{E}_p \mathrm{E}_q \left[(\hat{Z}_\bullet^{Q_\alpha} - Z^{Q_\alpha})^2 \right] - \left[\mathrm{E}_\xi \mathrm{E}_p \mathrm{E}_q (\hat{Z}_\bullet^{Q_\alpha} - Z^{Q_\alpha}) \right]^2 \\ &= V_{sam} + V_{imp} - \mathrm{B}^2_{\xi pq}(\hat{Z}_\bullet^{Q_\alpha}) \end{aligned} \tag{4.26}$$

Assuming a normal distribution approaches well enough the quantity of interest and using the estimated bias to correct only the variance component of the total of the linearized variable based on partially imputed data:

$$\hat{Q}_\alpha - Q_\alpha \sim N(0, V_{tot})$$

$$\frac{\hat{Q}_\alpha - Q_\alpha}{\sqrt{V_{tot}}} \sim N(0,1)$$

If we can estimate V_{tot}, a 95% confidence interval for the true value of Q_α is given by:

$$\hat{Q}_\alpha - q_{0.025}\sqrt{\hat{V}_{tot}} \lessapprox Q_\alpha \lessapprox \hat{Q}_\alpha + q_{0.975}\sqrt{\hat{V}_{tot}}$$

4.5.3 Imputation variance: $\hat{V}_{imp}$

Taking over (4.11), the imputation variance of the total of the linearized variable based on partially imputed data is defined by:

$$\begin{aligned} V_{imp\xi} &= \mathrm{E}_\xi \{ (\hat{Z}_\bullet - \hat{Z})^2 + 2(\hat{Z}_\bullet - \hat{Z})(\hat{Z} - Z) | s, r \} \\ &= \mathrm{E}_\xi \{ (\hat{Z}_\bullet - \hat{Z})^2 | s, r \} + 2\mathrm{E}_\xi [(\hat{Z}_\bullet - \hat{Z})(\hat{Z} - Z) | s, r] \end{aligned}$$

Proposition 4.4 *The imputation variance of the total of the linearized variable based on partially imputed data can be approximated by:*

$$V_{imp\xi} = \frac{1}{f^2(\hat{Q}_\alpha)} \frac{1}{\hat{N}^2} \left\{ \sum_{k \in o} \sum_{\ell \in o} w_k w_\ell (M_{k\ell} - 2\Phi M_{k\ell} + \Phi_{k\ell}) \right\},$$

The quantities $M_{k\ell}$, $\Phi M_{k\ell}$ and $\Phi_{k\ell}$ being defined in Proposition 4.3.

The proof is found in Appendix B.

4.5.4 Total variance for the linearized variable of a quantile

The estimated total variance will be given by:

$$\begin{aligned} V_{tot}^{Q_\alpha} &\overset{(4.26)}{=} V_{sam} + V_{imp} - \mathrm{B}^2_{\xi pq} \\ \hat{V}_{tot}^{Q_\alpha} &\overset{(4.18)}{=} \hat{V}_p(\hat{Z}_\bullet) + \hat{C}^{Q_\alpha} + V_{imp\xi} - \hat{\mathrm{B}}^2_{\xi pq} \end{aligned} \tag{4.27}$$

Case of simple random sampling without replacement

In the case of simple random sampling without replacement (srs), $\pi_k = \pi = n/N$ for all k, and $A_{k\ell}$ is given in (4.15). By taking over formula (4.24), $C^{Q_\alpha,srs}$ is written as (see the development in Appendix C):

$$\begin{aligned} C^{Q_\alpha,srs} &= \frac{1}{f^2(\hat{Q}_\alpha)} \frac{N-n}{Nn^2} \{\Phi_o - M_o \\ &\quad - \frac{1}{n-1} [\Phi_{\overline{oo}} - M_{\overline{oo}} + 2(\Phi_{ro} - \Phi M_{ro})]\}, \end{aligned} \tag{4.28}$$

where

- $\Phi_o = \sum_{k\in o} \Phi_k$, $M_o = \sum_{k\in o} M_k$,
- $\Phi_{\overline{oo}} = \sum_{k\in o} \sum_{\ell\in o, \ell\neq k} \Phi_{k\ell}$,
- $M_{\overline{oo}} = \sum_{k\in o} \sum_{\ell\in o, \ell\neq k} M_{k\ell}$,
- $\Phi_{ro} = \sum_{k\in r} \sum_{\ell\in o} \Phi_{k\ell}$,
- $M_{ro} = \sum_{k\in r} \sum_{\ell\in o} M_{k\ell}$.

Furthermore, $V_{imp\xi}$, given by (4.4), is written as:

$$V_{imp\xi}^{srs} = \frac{1}{f^2(\hat{Q}_\alpha)} \frac{1}{n^2} (M_{oo} - 2\Phi M_{oo} + \Phi_{oo}),$$

and the bias given by (4.25) squared is estimated by:

$$\hat{\mathrm{B}}^{2,srs}_{\xi pq}(\hat{Z}^{Q_\alpha}_\bullet) \approx \frac{1}{f^2(\hat{Q}_\alpha)} \frac{1}{n^2} (\Phi_o - M_o).$$

The general formula for the quantile α, always in the case of a simple random sampling without replacement is obtained by plugging in these quantities in equation (4.27).

4.6 Simulations from real data

We conducted simulations on three real data sets considered in the following as populations:

1. The equivalent household income of Swiss SILC data 2009 (see SFSO, 2014),

2. the employees' income, also from the Swiss SILC data 2009,

3. the determining income from the tax register of the Canton of Neuchâtel.

Only the second simulation on the employees' income is presented in this article to remain concise. The other simulations are available upon request. We have artificially eliminated from all these populations all incomes greater than 200,000 Swiss francs to avoid problems due to extreme values in the regressions. In a real production setting, we could obviously not afford to do so, one could for example take the logarithm of income and impute the transformed variable, then this change should also be taken into account in the calculations presented in the previous sections. In addition, we would create imputation classes with one model per class. The purpose of this exercise is not to achieve the best possible imputation but rather to study the behaviour of estimators built. We therefore preferred to avoid destabilizing the regression models and the calculation of linearized variables with extreme values for these effects, which should not be neglected in practice, might disturb the results of assessments that we want to conduct. Table 4.1 and Figure 4.1 give some indications of the distribution of the data used. In the following three paragraphs we give details on how the simulations were made before moving on to the description of the results.

Table 4.1 – *Descriptive statistics of the population used for the simulations presented (employees' income, SILC09), the distribution has been cut to 200,000 Swiss francs to avoid problems due to extreme values.*

Min.	16
1st Qu.	29,040
Median	61,530
Mean	64,240
3rd Qu.	90,610
Max.	199,900
N	6,561

The simulations are constructed according to the following scheme repeated 2000 times:

1. Select a sample s of size $n = 750$ according to a simple random sampling without replacement in the considered population,

2. generate a non-ignorable nonresponse depending on some auxiliary variables contained in a matrix $\mathbf{X}$,

3. on one hand, imputation by predicting the missing values using a regression model adjusted on the subset $r \subset s$ of respondents, on the other hand, to compare, calculation of weights compensating for the nonresponse according to a model assuming the nonresponse as completely uniform, and other weights obtained according to a

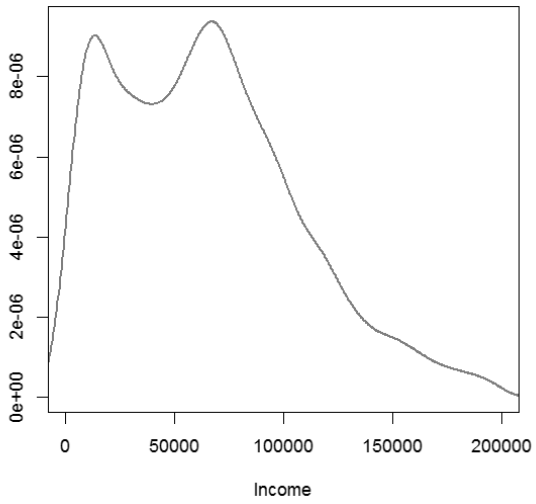


Figure 4.1 – *Descriptive statistics of the population used for the simulations (employees' income smaller than 200'000.- Swiss Francs, SILC09): kernel estimated density function with the R function* `drvkde` *from package* `feature`.

logistic model using the same auxiliary variables as those used for the regression imputation.

Indicators of interest for us are $\theta = med$, the median; Q_{20}, the 20$^{\text{th}}$ quantile; Q_{80}, the 20$^{\text{th}}$ quantile; $arpt$, the at risk of poverty threshold; $arpr$, the at risk of poverty rate; $gini$, the Gini index; qsr, the quintile share ratio; $rmpg$, the relative median poverty gap; $medp$, the median of the poor. Their definitions can be found in Eurostat (2003, 2005a); Graf & Tillé (2014). For comparison purposes, we have calculated different linearized variables for these indicators:

A. The linearized variables of the indicators on the complete population in order to have a variance estimation through linearization on all the data at disposal;

B. the linearized variables on the selected sample s;

C. the linearized variables solely on the respondents $r \subset s$:

   C1. assuming a completely uniform nonresponse, that is to say by simply expanding the weights of respondents by a single factor,

   C2. using weights that model the nonresponse mechanism;

D. the linearized variables on the selected sample s based on the respondents and the imputed values for the nonrespondents.

The variance of the indicators has been estimated on the basis of these linearized variables:

for A. $\text{var}(\theta) \approx \text{var}(\sum_{k \in U} \hat{z}_k)$,

for B. one obtains 2000 values $\text{var}(\hat{\theta}) \approx \text{var}(\sum_{k \in s} N/n\, \hat{z}_k)$,

for C1. one obtains 2000 values $\text{var}(\hat{\theta}) \approx \text{var}(\sum_{k \in r} N/n\, n/n_r \hat{z}_k)$,

for C2. one obtains 2000 values $\text{var}(\hat{\theta}) \approx \text{var}(\sum_{k\in r} N/n\, 1/f_k^{logit}\, \hat{z}_k)]$, where f_k^{logit} is the adjustment factor obtained by the logistic model,

for D. two variants were calculated for all indicators:

D1. ignoring that a part of the data was imputed, one obtains 2000 values through: $\text{var}(\hat{\theta}) \approx \text{var}(\sum_{k\in s} N/n\, \hat{z}_k)$

D2. the plug-in system is applied, one obtains 2000 values of the variance by (4.14).

For the quantiles ($\theta = med$, Q_{20}, Q_{80}) we have also calculated the variant we propose:

D3. one obtains 2000 values of the variance by (4.27).

All these variants are each compared with the Monte-Carlo estimation of the variance based on the 2000 point estimates obtained.

We have not tested other designs than the simple random sampling, nor made any calibrations. These aspects are of course important. However, on the one hand they are not the sought goal. On the other hand, we want to focus on the effect of regression imputation only without mixing this effect to that of a calibration on margins, for example.

4.6.1 Simulations: employees' income SILC09

The variable of interest here is the employees' income of the respondent household at the Swiss SILC survey in 2009. This is the variable of the Central Compensation Office (CCO) register, which was matched with data from the SILC survey. As explanatory variables we find the cost of housing *hous*\$, the occupation rate *occ_r*, two indicator variables of age classes 20 to 29 years old (*ac_3* = 1, 0 otherwise) and 40 to 49 years (*ac_5* = 1, 0 otherwise), the variable *boss* is equal to one if the person occupies a position of chief, zero otherwise, *man* = 1 for men, 0 otherwise and *house* = 1 for individuals living in a house, 0 otherwise. On the entire population ($N = 6,561$) all the model variables $y_{cco} \sim hous\$ + occ\_r + ac\_3 + ac\_5 + boss + man + house$ are significant and the model has a coefficient of determination R^2 of 44.8%. The probability p of each individual not responding has been fixed once and for all through

$$\mathbf{p} = \frac{1}{1+\exp(\mathbf{Xb})} \in [0,1],$$

which is specific to the characteristics of the individual identified in the matrix $\mathbf{X}$ comprising in order the variables from the regression model with the a column of one in first position. The vector $b = (0.5, 0.8, 0.5, 0.5, 0.4, .8, 0, 0, 0)$ was determined such that the probability of not responding is approximately 20%: min = 5%, Q_{25} = 14%, med = 20%, mean = 21%, Q_{75} = 27%, max= 38%. The simulated nonresponse is ignorable, it depends on variables *ac_3*, *ac_5*, *boss*, *man*, and *house* from

X, not from **y**. The expected response rate throughout the population by randomly deciding who are the respondents is 80%.

Models adjusted for respondents from each sample regression had on average a R^2 of 43.7%. The variables *ac*_3, *ac*_5, *boss*, *man*, *house* and *occ_r* were used as explanatory variables.

4.7 Results

4.7.1 Total variance

Table 4.2 summarizes the results of the simulations presented in Section 4.6 concerning the variance. The following points are observed:

1. Generally, variant C2, where weights supposed to compensate for the nonresponse mechanism were produced, leads to an overestimation of the variance through linearization. This overestimation can be massive in the case of the total Y. Conversely, as expected, variant C1, where nonresponse is considered wrongly as uniform, almost always leads to a substantial underestimation of the variance. Alternative B systematically checks, what we assume implicitly, that the variance estimation through linearization works everywhere when there is no nonresponse.

2. To not consider the imputation in the variance estimation (variant D1), i.e. to act as if the imputed values were true values, leads to a systematic underestimation. The latter is particularly important in the case of the total Y, ARPT, ARPR as well as for two of the considered quantiles. It is interesting to note that, for the dataset used, the variance of Q_{20}, Gini, QSR, RMPG and MEDP is only slightly underestimated when ignoring the fact that 20 to 30% of the values were imputed. However simulations on other datasets showed that the latter observation is not valid in general.

3. The method of Deville & Särndal (1994), variant D2, actually seems well suited to the estimation of variance under imputation of the total Y as these authors argue. The relative bias of the variance of -19% with D1 is reduced to 8% with D2

4. For all other indicators, in addition to not being defendable theoretically, the plug-in method D2 is not reliable:

   - for the median, the Q_{80} and the ARPT it seems to offset some of the underestimation but not completely,
   - for the ARPR, MEDP and RMPG it leads to an important overestimation. Other simulations on other datasets have shown that it can inflate more than ten times the true value in the case of MEDP and RMPG.
   - for Gini and QSR the plug-in method seems to be particularly unstable, underestimation can be maintained as observed in

other simulations or an important overestimation can be introduced, as is the case here.

5. On the method proposed in this paper (variant D3)

   - The method works very well for the median and Q_{80}. It is better than all other alternatives, does not provide unreasonable values; the theory is respected; and it can be applied without too much risk when the samples are large enough to ensure that the point estimates of the quantile in question are fairly stable. It is indeed important to respect this restriction because, to recall, in the developments of Section 4.5, we neglected the fact that the quantile is random.
   - For Q_{20}, one overestimates the variance by 30%, when, by chance, it was only overestimated by 5% using the method D1 (= ignore the imputation). But we have observed underestimates up to -26% on other datasets with the D1 method Method D2 (plug-in) has an intermediate behaviour. This is mainly due to the shape of the left tail of the income distribution (see Figure 4.1). In reality, if the population's income had such a distribution, we should refine our model to better fit lower incomes. On the other hand, if the survey sample presented such an income distribution in practice, one would not be surprised and more at ease obtaining a high variance estimate for low quantiles like the Q_{20}.
   - For the median (Q_{50}) and the Q_{80}, our method compensates for the underestimation we commit by ignoring imputation.

Table 4.2 – *Relative biases of the variance compared to Monte Carlo variance for different variants of variance estimation in simulations carried out for the employee's income from SILC 2009. A: values for the entire population. B: samples without NR. C1: weighted samples but without modeling the NR. C2: weighted samples with weights modeling the NR. D1: imputed samples but imputation ignored. D2: imputed samples, plug-in method Deville & Särndal. D3: imputed samples, our method.*

	TY	med	Q20	Q80	gini	qsr	arpt	arpr	rmpg	medp
A (Pop)	0.03	-0.01	**0.24**	0.07	-0.02	0.02	-0.00	-0.01	0.01	0.02
B	0.03	0.05	0.04	**0.14**	-0.02	-0.00	0.05	0.01	0.09	0.10
C1 (unif)	0.13	-0.07	**-0.17**	-0.07	**-0.20**	**-0.22**	-0.07	**-0.20**	**-0.15**	**-0.12**
C2	**2.75**	**0.20**	0.01	**0.24**	0.00	-0.01	**0.21**	0.01	0.09	**0.13**
D1	**-0.19**	**-0.26**	0.05	**-0.20**	-0.04	-0.00	**-0.26**	**-0.12**	-0.05	-0.01
D2	0.08	**-0.10**	**0.22**	**0.10**	**0.11**	**0.71**	**-0.10**	**0.71**	**0.60**	**0.73**
D3	-	-0.05	**0.30**	-0.02	-	-	-	-	-	-

4.7.2 Components of total variance in the case of quantiles

Table 4.3 shows the weights of the different estimated contributions to total variance $V_{tot}^{Q_\alpha}$, given by (4.27), for the conducted simulations. The four components are the design variance ignoring imputation, see (4.4),

$V_p(\hat{Z}_\bullet^{Q_\alpha})$, the corrective term $\hat{C}^{Q_\alpha}$, see (4.24), the imputation variance $V_{imp\xi}$, see Proposition 4.4, and minus the squared bias $\hat{B}^2_{\xi pq}(\hat{Z}_\bullet^{Q_\alpha})$, see (4.25).

- For the median, we find that the corrective term $\hat{C}^{Q_\alpha}$ can be ignored and it is the smallest of all contributions to the total variance in the case of other quantiles ($< 7\%$).
- The design variance ignoring the imputation component, $V_p(\hat{Z}_\bullet^{Q_\alpha})$, oscillates stably for a contribution between 75% and 78% on average to the total variance. Observing other measures of its distribution, we find that the estimate is stable and varies little.
- The imputation variance component $V_{imp\xi}$ as well as the one for squared bias $\hat{B}_{\xi pq}(\hat{Z}_\bullet^{Q_\alpha})$ fluctuate much more. For some samples they may even take fairly large values balancing out one another. For the median, theses components fluctuate less than for the other two considered quantiles.

Table 4.3 – *Estimated proportions of the various components of the total variance, see* (4.27), $V_{tot}^{Q_\alpha}$ *for the simulation on the employee's income from SILC 2009. The four components are: the design variance ignoring imputation,* $V_p(\hat{Z}_\bullet^{Q_\alpha})$*, the corrective term* $\hat{C}^{Q_\alpha}$*, see* (4.24)*, the imputation variance* $V_{imp\xi}$*, see Proposition 4.4, and minus the squared bias* $\hat{B}_{\xi pq}(\hat{Z}_\bullet^{Q_\alpha})$*, see* (4.25).

stat	$V_p(\hat{Z}_\bullet^{Q_\alpha})$	$\hat{C}^{Q_\alpha}$	$V_{imp\xi}$	$-\hat{B}^2_{\xi pq}(\hat{Z}_\bullet^{Q_\alpha})$
		Q_{50}		
Min.	0.46	0	0.12	-1.13
1stQu.	0.71	0	0.24	-0.11
Median	0.76	0	0.31	-0.04
3rdQu.	0.81	0	0.39	-0.01
Max.	0.88	0	1.31	0
mean	**0.75**	**0**	**0.33**	**-0.08**
		Q_{20}		
Min.	0.63	0.03	0.48	-5.39
1stQu.	0.76	0.06	1.51	-2.22
Median	0.78	0.06	1.94	-1.77
3rdQu.	0.79	0.07	2.39	-1.36
Max.	0.86	0.1	5.59	-0.31
mean	**0.77**	**0.06**	**1.99**	**-1.83**
		Q_{80}		
Min.	0.5	0.02	0.31	-5.48
1stQu.	0.76	0.04	1.26	-2.03
Median	0.78	0.05	1.7	-1.53
3rdQu.	0.8	0.06	2.2	-1.08
Max.	0.85	0.09	5.68	-0.16
mean	**0.78**	**0.05**	**1.8**	**-1.62**

4.7.3 Biais of estimators

The variance and bias of the variance having been discussed, what about the bias of the point estimator? Table 4.4 shows the relative biases obtained. For total Y, it can be seen that even for the variant C1, which

wrongly uses a weighting not correcting for nonresponse, the bias is negligible, also under imputation. For other estimators considered (quantile and inequality indicators), even the variant C1 shows a low bias. Alternative C2 (= compute weights offsetting the nonresponse) provides estimates with negligible bias for all measures.

Under imputation, bias remains acceptable for most measures, but it becomes troublesome for Q_{20} (16%), *qsr* (-17%) and the median of the poor (11%). Observing the density plot of incomes on the population used for the simulations (see Figure 4.1), we see that low incomes have a very different left tail distribution than a normal distribution (which is what the regression model used to predict missing values implicitly assumes). It is not surprising that the estimate of Q_{20} under imputation suffers from this. This has implications on the *qsr* which is the ratio of the sum of the incomes of the 20% richest on the 20% poorest. Similarly, the median of the poor will be sensitive to it and therefore the *rmpg* ($= (arpt - medp)/arpt$). The purpose of this article was to explore the possibilities of a methodology for estimating variance through linearization taking imputation into account. So we did not take all precautions to minimize the risk of bias in the estimators. If one wants to transpose the system into production, it is obvious that we should at least take the logarithm of income variable before imputing by regression and should form two to three imputation classes to try to model separately low income, middle income and high income[1]. This is what is currently done in the imputation models in production at the Swiss Federal Statistical Office in the SILC survey. But then, one must also take all these transformations into account in the estimation of variance through linearization as we have done it for quantiles. That is why, in the interest of keeping things simple and understandable, we have not done all these potential improvements to the imputation models.

In the case of variant D3 (our estimate of the total variance) for quantiles, the bias $B_{\xi pq}(\hat{Z}_{\bullet}^{Q_\alpha})$ was estimated to correct the total variance, see (4.27). One might be tempted to also use this bias estimate to correct the point estimator. Table 4.5 shows that indeed, on the 2,000 simulations, the average estimated bias through linearization generally takes comparable proportions to the true Monte Carlo bias (= value estimated on the whole population less the sample imputed value). The estimated biases through linearization on each sample have a correlation of 38%, -1% and 37% with true Monte Carlo Biases in the cases of, respectively, the median, Q_{20} and Q_{80}. In simulations on other datasets, this correlation went up to 34%, 28% and 52% respectively. The option of correcting the estimator with the estimated bias still does not seems to be a good option because it would further increase the total variance and we would then have to consider it in the correction. It is better to improve the imputation models as mentioned above, in particular by creating imputation classes.

[1]See Appendix F for more details on the creation of imputation classes. Note that the appendix is not part of the original publication.

Table 4.4 – *Relative biases of the point estimates for the different variants with the simulations conducted on employee's income in SILC 2009. B: samples without NR. C1: weighted samples but without modeling NR. C2: weighted samples with weights modeling the NR. D: imputed samples.*

	TY	med	Q20	Q80	gini	qsr	arpt	arpr	rmpg	medp
B. (full resp.)	-0.00	-0.00	-0.01	-0.00	-0.00	0.00	-0.00	-0.00	0.00	-0.00
C1 (unif)	0.02	0.01	0.01	0.01	-0.01	-0.00	0.02	-0.01	0.01	0.00
C2	0.00	-0.00	-0.02	-0.00	-0.00	0.00	-0.00	-0.00	0.00	-0.00
D. (imp.)	-0.00	0.01	**0.16**	-0.03	-0.06	**-0.17**	0.02	-0.07	-0.07	**0.11**

Table 4.5 – *Mean Monte-Carlo biases and estimated average bias* $\overline{\hat{B}_{\xi pq}(\hat{Z}_{\bullet}^{Q_\alpha})}$ *(bias through linearization) on the 2,000 simulations for the three quantiles.*

	Q_{50}		Q_{20}		Q_{80}	
	Bias	Rel. bias	Bias	Rel. bias	Bias	Rel. bias
Bias MC	894	0.01	**3,540**	**0.16**	-2,652	-0.03
Bias lin.	275	0.00	**2,898**	**0.13**	-3,303	-0.03

4.8 Conclusions

Almost all countries seeking to estimate inequality indicators (Eurostat, 2005a) and their accuracy proceed through sample surveys. They must then deal with nonresponse of various natures and causes. Generally speaking, in the end, the nonresponse at the respondent level (household or individual in the cases discussed here) is most often handled by reweighting (see for example Graf, 2008b, 2009b) the nonresponse on the question level, i.e. to a specific variable, by means of imputation. Parametric imputation using a regression model is a method among others which is often implemented. When it comes to estimating the variances of extrapolated values for complex and non-linear statistics as indicators of poverty and social exclusion, this study confirms a phenomenon that is already known, that ignoring the stage of imputation in the variance estimation leads to underestimation of variance. Underestimation is dangerous because it suggests that the survey as a measurement tool is more accurate than it really is. It may, for example, lead one to over-interpret the results which, instead of helping decision-makers at the political and social levels, complicates their task without them realizing it. It is therefore essential to quantify the accuracy of these measures, and one prefers to slightly over-estimate the variance than to underestimate it. This raises the question of the relevance of the available precision estimates.

This work also shows that the method given by Deville & Särndal (1994) and tested by Lee et al. (1994) can be adapted to quantiles and a substantial improvement can be made in estimating the accuracy of the quantiles under imputation. We thus solved the problem for the at risk poverty threshold (ARPT) since it simply corresponds to 60% of the median. For other indicators of poverty and inequality we may try a similar application to the one presented here with a little bit of calculation.

Although we are more concerned with the estimation of variance, it should be noted that the bias of the estimators can also become problematic. The risk increases with a decreasing response rate and/or with poorly fitted regression models used for the prediction of missing values. Models we tested have nonresponse rates varying between 20% and 33% and coefficients of determination from 40% to 47%. Simulations for one of these models have been presented in this article. In this context, some estimators showed a significant bias, but as mentioned, more efforts in modelling the distribution of income would significantly improve this aspect. But that was not the primary goal of this article.

In all cases, we hope that this work will contribute to an awareness of the fact that we must act with caution when implementing the estimation of variance through linearization for all indicators involving quantiles. Indeed, the linearization method proposed by Deville (1999b); Demnati & Rao (2004b); Osier (2009) works well, but as shown in Graf & Tillé (2014), we must not only pay attention to the way the income density is estimated for the calculation of linearized variables, and be wary of the effect of outliers on the linearized variables as noted in Croux (1998), we must also not neglect the component due to imputation as was demonstrated in this

paper. One must also take care that the imputation system in place limits at best the risk of introducing a bias into the estimates produced with partially imputed data.

4.9 Acknowledgements

This work was carried out within the framework of a cooperation agreement between the Institute of Statistics of the University of Neuchâtel and the Swiss Federal Statistical Office (SFSO). We also wish to thank in particular Pr. Yves Tillé for his help and comments and the section *Income, Consumption and Living Conditions* of the SFSO for having made available the data of the Swiss part from the European survey on income and living conditions. Our gratitude also goes to the Statistical Service of the Canton of Neuchâtel for having provided a anonymized extract from the tax-register.

5 Imputation of income data with generalized calibration procedure and GB2 distribution: illustration with SILC data

Abstract

In sample surveys of households and persons, questions about income are often sensitive and thus subject to a higher nonresponse rate. Nevertheless, the household or personal incomes are among the important variables in surveys of this type. The distribution of such collected incomes is neither normal, nor log-normal. Hypotheses of classical regression models to explain the income (or their log) are not satisfied. Imputations using such models modify the original and true distribution of the data. This is not suitable and may lead the user to wrong interpretations of results computed from data imputed in this way. The generalized beta distribution of the second kind (GB2) is a four-parameter distribution. Empirical studies have shown that it fits income data very well. The advantage of a parametric income distribution is that there are explicit formulae for the inequality and poverty indicators as functions of the parameters. We present a parametric method of imputation relying on weights obtained by generalized calibration. A GB2 distribution is fitted on the income distribution in order to assess that these weights can compensate for nonignorable nonresponse that affects the variable of interest. The success of the operation greatly depends on the choice of auxiliary and instrumental variables used for calibration, which we discuss. We validate our imputation system on data from the Swiss Survey on Income and Living Conditions (SILC) and compare it to imputations performed through the use of IVEware software (Raghunathan et al., 2001) running on SAS®. We have made great efforts to estimate variances through linearization, taking all the steps of our procedure into account (sampling design, weighting, imputations).[1]

Keywords: GB2, generalized calibration, inequality measures, Laeken indicators, nonignorable nonresponse, SILC, IVEware

[1]This chapter is a reprint of: Graf, E. (2014). Imputation of income data with generalized calibration procedure and GB2 distribution: illustration with SILC data. *Submitted article*.

5.1 Introduction

Undoubtedly, knowledge about income distribution of the population is of vital interest to all studies of economic markets in order to guide economic and social decision-making. In economics and social statistics, the study of income distribution is at the heart of inequality measures and more generally of evaluations of social welfare. In official national sample surveys of households and individuals, nonresponse in income variables is often a difficult and serious problem in the sense that the phenomenon affects the image of the reality that the survey reflects. Without proper treatment, the outcomes of political or social decision-making based on the results of the survey may be wrong. We show that generalized calibration can be used to obtain weights allowing us to compensate even for nonignorable nonresponse (see for instance Osier, 2013; Lesage & Haziza, 2013a,b)). Moreover these weights can be used to perform an imputation and fit a Generalized Beta distribution of the second kind (GB2) on the collected income data. The procedures developed are applied to the Swiss Survey on Income and Living Conditions (SILC) data from year 2009 to assess their validity. We compare the outcomes of our imputation system with those obtained through the use of the IVEware (Raghunathan et al., 2001) SAS® software and show that we perform much better in terms of satisfying the original income distribution as well as in the bias of several inequality indicators computed on the partially imputed data. The present work is of direct relevance to National Statistical Institutes and is endorsed by a very productive collaboration with the Swiss Federal Statistical Office (SFSO). Our work uses well known statistical tools but combine them in an innovative way.

We aim to develop a method of imputation for income variables allowing direct analysis of the distribution of such data, but also the estimation of complex statistics such as inequality indicators for poverty and social exclusion. We focus on five of them, formerly called *Laeken indicators* (Eurostat, 2003, 2005a): the At Risk of Poverty Threshold (ARPT), At Risk of Poverty Rate (ARPR), Relative Median at-risk-of Poverty Gap (RMPG), Quintile Share Ratio (QSR) and the Gini index (GINI).

It is very important that the imputation method is transparent and allows us to produce reliable variance estimates for any statistic based on the imputed data including the complex statistics mentioned above. Today the SFSO uses IVEware to conduct the necessary imputations for several income variables. IVEware can perform single or multiple imputations of missing values using the Sequential Regression Imputation Method. Although IVEware is a very convenient tool, in our context it has several drawbacks:

- It produces a poor output documenting the models that were fitted (lack of transparency).
- If the data contain extreme values, as is often the case with income, these can destabilize the models.

- Output information is insufficient for producing variance estimates for (complex) statistics.
- Somehow a (multi-)normal distribution is assumed and used to impute, therefore, imputed data cannot fit heavy tailed distributions.
- It does not allow survey weights to be used, which prevents the use of nonresponse correction factors.
- It is not easy to decide how many imputations to produce and then which measure to take as the final imputed value.

Convincing results of empirical studies on income (see for instance Kleiber & Kotz, 2003; Jenkins, 2007; Dastrup et al., 2007; Sepanski & Kong, 2008; Jones et al., 2011; McDonald et al., 2013) have shown that the Generalized Beta distribution of the second kind (GB2) fits well with such data and it is often more suitable than other four-parameter distributions. Moreover, the European project European Union (2011a) confirmed this fact for EU-SILC data. Encouraged by these works, we decided to work out another parametric imputation method partially based on a weights stemming from a generalized calibration and assessed through GB2 adjustment.

The Swiss SILC data could be matched with the register of the Central Compensation Office (CCO). So some of the variables collected through the survey showing item nonresponse could be compared, on the unit level, with their official corresponding records found in the register. The register presents almost no missing values. The nonresponse mechanism observed in the survey was then applied to the register-variable, without the need to be modelled, providing a unique occasion to test imputation models and compare their outputs with the true values. It is worth noting that the nonresponse mechanism does not need to be fully describable with measured or known variables. This allows us to illustrate and evaluate our imputation strategy. We deal with about 30% of nonresponse rate.

In Section 5.2 we introduce the notations needed in our work. Section 5.3 presents the Generalized Beta distribution of the second kind. Section 5.4 is dedicated to generalized calibration and its link with instrumental regression is recalled. Section 5.5 deals with weighted ranks of observations and their transformation into normal scores. The core of our imputation strategy is summarized in Section 5.6. Finally, Section 5.7 emphases the problem of variance estimation for our data. We illustrate and explain in a more detailed fashion our methodology in Section 5.8 providing also some results and conclude in Section 5.9.

5.2 Notations

Through an individual or household survey (example: SILC), we collect, among other measures, an income variable, say y. In our scenario y is affected by nonignorable nonresponse (worst case) (NMAR, *Non Missing At Random*, in the sense of Little & Rubin, 2002). That means that

the missingness pattern depends not only on auxiliary variables, but also on the (unknown) values of y itself. We place ourselves in an end user situation where we are dealing with a partially missing y-variable (item nonresponse) accompanied by a set of weights which may reflect the sampling design, different sorts of nonresponse adjustments and calibration on known population total. In other words, the questionnaire non response has already been treated by reweighting. We intend to deliver an imputed variable: $y_{\bullet k}$ draw inferences to the target population by publishing estimated values with confidence intervals for several indicators of poverty and social exclusion based on this imputed variable.

Let U denote the finite population of size N, s the surveyed sample of size n, w_k the survey weight for unit $k \in s$. The sample s may already be a subset of the selected sample according to a sampling design and the weights w_k may consequently be defined as a product of the sampling weight with questionnaire nonresponse and calibration correction factors. Let $r \subseteq s$ be the set of respondents of size n_r to the interest variable y resulting from the unknown NMAR mechanism. Further, let $o = s \backslash r$ be the set of non-respondents to y of size n_o ($n_o + n_r = n$). Let $\mathbf{X}_U$ be the $N \times J$ matrix grouping the J *auxiliary variables* defined on U. Our notations are thought as follow

$$\mathbf{X}_U = \begin{pmatrix} \mathbf{x}_{1\cdot} \\ \vdots \\ \mathbf{x}_{N\cdot} \end{pmatrix} = (\mathbf{x}_{\cdot 1} \cdots \mathbf{x}_{\cdot J}).$$

Let also $\mathbf{X}_s$ and $\mathbf{X}_r$ be similar matrices but of size $n \times J$, respectively $n_r \times J$, grouping the auxiliary variables defined on s, respectively on r. The row vector $\mathbf{x}_{k\cdot} = (x_{k1}, ..., x_{kj}, ..., x_{kJ})$ of length J is the k^{th} row of $\mathbf{X}_U$, $\mathbf{X}_s$ or $\mathbf{X}_r$, and we will denote by $\mathbf{x}_{\cdot j}$ a column vector of length N, n or n_r corresponding to the j^{th} column of $\mathbf{X}_U$, $\mathbf{X}_s$, respectively $\mathbf{X}_r$. Similarly, $\mathbf{Z}_r$ will be the $n_r \times J$ matrix grouping the J *instrumental variables* defined on r. The j^{th} column of $\mathbf{Z}_r$ is the column vector $\mathbf{z}_{\cdot j}$ of length n_r. The k^{th} row of $\mathbf{Z}_r$ is the row vector $\mathbf{z}_{k\cdot} = (z_{k1}, ..., z_{kj}, ..., z_{kJ})$. We assume that all these matrices have the same number of columns J.

The vector $\mathbf{x}_{k\cdot}$ is used for calibration and prediction through a regression model (see Section 5.4). The $\mathbf{x}_{k\cdot}$ are assumed to be known on s, particularly their totals on s are available: $\hat{\mathbf{t}}_x = \sum_{k \in s} w_k \mathbf{x}'_{k\cdot}$ is known. The vector $\mathbf{z}_{k\cdot} = (z_{k1}, ..., z_{kj}, ..., z_{kJ})$ of *instrumental variables* is assumed to be measured at least on r. Note that some of the $\mathbf{z}_{\cdot j}$ can be identical to some of the $\mathbf{x}_{\cdot j}$. In short, one observes $(y_k, \mathbf{x}_{k\cdot}, \mathbf{z}_{k\cdot})$ and disposes from $\hat{\mathbf{t}}_x$. We shall denote by $y_{\bullet k}$, $k \in s$, the partially imputed variable:

$$y_{\bullet k} = \begin{cases} y_k & \text{if } k \in r \\ \hat{y}_{imp,k} & \text{if } k \in o, \end{cases}$$

where $\hat{y}_{imp,k}$, $k \in o$, are imputed values obtained by an imputation procedure. Before describing how we impute the missing values of y in Section 5.6, the next few Sections present the necessary concepts and methods we shall apply afterwards.

5.3 GB2 distribution and fit

The Generalized beta distribution of the second kind (GB2) is a four-parameter distribution: $GB2(a, b, p, q)$, it was developed by McDonald (1984). Its density is given by

$$f_{GB2}(y; a, b, p, q) = \frac{a}{b \cdot B(p, q)} \frac{(y/b)^{ap-1}}{(1 + (y/b)^a)^{p+q}},$$

where $B(p, q) = \int_0^1 t^{p-1}(1-t)^{q-1}dt$ is the beta function. Many other probability distributions can be seen as special cases of the GB2 by re-parametrizing or setting some of the parameters equal to suitably fixed values (Generalized Gamma, Dagum, Beta2, Singh-Maddala, Lognormal, Gamma, Fisk, Weibul). Parameter a represents the sharpness of the distribution, p governs the left tail, q the right tail and b is a scale parameter. For large values of a, b tends to the expectation.

The advantage of a parametric estimation for a distribution of income is that there are explicit formulae for the inequality measures as functions of the four parameters $\theta = (a, b, p, q)^T$ of the GB2 adjusted to the data. The detailed expressions can be found in McDonald (1984); European Union (2011a); Graf & Nedyalkova (2011, 2013). The main inequality measures of concern in SILC are programmed and available in the statistical software **R**, `package("GB2")`. This package contains in particular the function `profml.gb2(y,w)` allowing one to fit a GB2 distribution on a dataset, where y is the variable of interest (in our case the collected income through the survey or matched register data) and w the associated weights. Estimating equations obtained by applying the method of maximum likelihood for the GB2 distribution are used (see European Union, 2011a; Graf & Nedyalkova, 2013).

Note, and this is an important limitation, that we need a large enough sample (a few thousand) in order to estimate the GB2-parameters a, b, p, q precisely enough (European Union, 2011a). But this is not a worry in the SILC survey except if one would work on small sub-populations, which is not our case here.

5.4 Generalized calibration

Calibration methods have been introduced by the reference articles of Deville & Särndal (1992) and Deville et al. (1993). Almost ten years later, the generalized calibration method has been presented and applied by Le Guennec & Sautory (2002); Sautory (2003); Deville (2002). Kott (2006); Osier (2013); Lesage & Haziza (2013b,a); Park & Kim (2014) studied generalized calibration and its ability to correct nonresponse. Classically, the objective is to estimate a population total $t_y = \sum_U y_k$, ideally by $\hat{t}_y = \sum_s w_k y_k$ (Horvitz-Thompson estimator of the total) if all the units of s are available, more realistically based on the respondents' subset:

$$\hat{t}_y^{cal} = \sum_{k \in r} w_k^{cal} y_k. \tag{5.1}$$

The calibrated weights w_k^{cal}, $k \in r$, are, in the case of classical calibration, "near" to the previous weights w_k according to some (pseudo-)distance and must satisfy the constraints

$$\sum_{k \in r} w_k^{cal} \mathbf{x}'_{k.} = \hat{\mathbf{t}}_x, \tag{5.2}$$

where $\hat{\mathbf{t}}_x$ is the $J-$vector of known totals estimated on the complete sample: $\hat{\mathbf{t}}_x = \sum_s w_k \mathbf{x}'_{k.}$. Note that, classically, one calibrates the sample (s-level) on known population totals (U-level). In this paper, we calibrate the subsample of respondents to y (r-level) on some known totals at the s-level. In the case of calibration, the calibrated weights are of the form

$$w_k^{cal} = w_k F(q_k \mathbf{x}_{k.} \boldsymbol{\lambda}),$$

where $F(u) : \mathbf{R} \to \mathbf{R}$ is the so-called calibration function for which the usual choices are available: linear, raking ratio, logit, truncated, see for example Deville et al. (1993). It is a regular function and satisfies $F(0) = \partial F(0)/\partial u = 1$. The $J-$vector $\boldsymbol{\lambda}$ is a Lagrange multiplier. One can give more or less importance to certain units through the q_k. Note that the q_k are often simply set equal to one for all k in practice, which is what we did. For the sake of simplicity, in the following, the q_k are considered to be equal to one for all k.

For generalized calibration the weights obtained are of the form $w_k^{cal} = w_k F(\mathbf{z}_{k.} \boldsymbol{\lambda})$ and the function F could even be observation-dependent and thus be from $\mathbf{R}^J \to \mathbf{R}$. However, the notion of minimized distance to the weights before calibration is less straightforward.

The simplest case is the linear one where $F(u) = 1 + u$, and one has $F(\mathbf{x}_{k.} \boldsymbol{\lambda}) = 1 + \mathbf{x}_{k.} \boldsymbol{\lambda}$ in the case of calibration, $F(\mathbf{z}_{k.} \boldsymbol{\lambda}) = 1 + \mathbf{z}_{k.} \boldsymbol{\lambda}$ for generalized calibration.

By solving (5.2) in $\boldsymbol{\lambda}$ one obtains, in the case of calibration,

$$\hat{t}_{ylin} = \hat{\mathbf{t}}'_x \hat{\mathbf{B}}_r + \sum_{k \in r} w_k e_k^r, \tag{5.3}$$

where

$$e_k^r = y_k - \mathbf{x}_{k.} \hat{\mathbf{B}}_r \tag{5.4}$$

are the residuals of the regression of y on the J auxiliary variables $\mathbf{x}_{k.}$.

$$\hat{\mathbf{B}}_r = \mathbf{T}_r^{-1} \sum_{k \in r} w_k \mathbf{x}'_{k.} y_k \tag{5.5}$$

is the J-parameters vector of the regression with

$$\mathbf{T}_r^{-1} = \left(\sum_{k \in r} w_k \mathbf{x}'_{k.} \mathbf{x}_{k.} \right)^{-1}.$$

Obviously, this means that the better the auxiliary variables explain y, the smaller the residuals e_k will be, thus also the variance of $\hat{t}_{ylin}$.

For generalized calibration, similarly, solving for $\boldsymbol{\lambda}$ so that w_k^{cal} satisfy the constraints (5.2) leads to

$$\hat{t}_{ylinG} = \hat{\mathbf{t}}_x' \hat{\mathbf{B}}_{rzx} + \sum_{k \in r} w_k e_k^g, \tag{5.6}$$

where

$$e_k^g = y_k - \mathbf{x}_{k\cdot} \hat{\mathbf{B}}_{rzx} \tag{5.7}$$

are the residuals of the instrumental regression of y on the J auxiliary variables $\mathbf{x}_{k\cdot}$ on sample r, with the J instrumental variables $\mathbf{z}_{k\cdot}$ (Deville, 2000, 2002).

$$\hat{\mathbf{B}}_{rzx} = \mathbf{T}_{rzx}^{-1} \sum_{k \in r} w_k \mathbf{z}_{k\cdot}' y_k \tag{5.8}$$

is the J-parameters vector of the instrumental regression with

$$\mathbf{T}_{rzx}^{-1} = \left(\sum_{k \in r} w_k \mathbf{x}_{k\cdot}' \mathbf{z}_{k\cdot} \right)^{-1}.$$

Note that in the case of generalized calibration, the estimated vector of parameters also depends on the $\mathbf{z}_{k\cdot}$. A crucial point is that the $\mathbf{z}_{k\cdot}$ only needs to be known on the sub-sample r of respondents to y. Asymptotically, in the case of full response, all the choices of calibration functions are equivalent to the linear one presented here. But, in the presence of nonresponse, as Lesage & Haziza (2013a,b) show, the choice of calibration function reflects implicit assumptions on the nonresponse model. If the choice of calibration function is not aligned with the true nonresponse mechanism, the calibrated estimator may show an important bias.

5.4.1 Instrumental variable regression

As stated in Stock & Trebbi (2003), the theory of instrumental variables was first derived by Wright (1928). The topic is widely covered and discussed in econometrics, we have in particular referred to Fuller (1987); Davidson & MacKinnon (1989, 1993); McFadden (1999); Cameron & Pravin (2005). We shall here recall some fundamental aspects of the method without going into details and by considering its utility in the framework of sampling in an innovative way. Comparing generalized calibration to instrumental variable regression allows us to apply known properties to control the choice of auxiliary and instrumental variables used.

The variable of interest y, which exists for the whole population, has been observed through the survey only on the respondents' subset $r \subset s$. Anyway, if one could do it, by regressing y on the auxiliary variables on U one would try to explain the greatest possible part of variation in y with the help of the $\mathbf{x}_{\cdot j}$'s:

$$y_k = \mathbf{x}_{k\cdot} \hat{\mathbf{B}}_U + e_k^U,\ k \in U, \tag{5.9}$$

where $\hat{\mathbf{B}}_U = (\mathbf{X}_U' \mathbf{X}_U)^{-1} \mathbf{X}_U' \mathbf{y}$. This regression on the whole population is of course hypothetical and cannot really be performed. Still, the residuals e_k^U are completely determined by y and $\mathbf{X}_U$. Geometrically, by construction

$\mathbf{X}'_U\mathbf{e}^U = \mathbf{0}$. The estimated value $\hat{y}_k = \mathbf{x}_{k\cdot}\hat{\mathbf{B}}_U$ is the projection of y on the hyperplane spanned by $\mathbf{X}_U$. Note that we do not necessarily assume a model of the type $y_k = \mathbf{x}_{k\cdot}\boldsymbol{\beta} + \epsilon_k$, $k \in U$, with ϵ_k following a distribution. We simply try to construct the best possible guess of y with the help of the auxiliary variables at our disposal.

Now consider sample $s \subset U$ drawn according to a sampling design $p(s)$. The adjusted regression at the level of the sample s is

$$y_k = \mathbf{x}_{k\cdot}\hat{\mathbf{B}}_s + e_k^s,\ k \in s, \tag{5.10}$$

where $\hat{\mathbf{B}}_s = (\mathbf{X}'_s\mathbf{D}_s\mathbf{X}_s)^{-1}\mathbf{X}'_s\mathbf{D}_s\mathbf{y}_s$, $\mathbf{D}_s = \text{diag}(1/\pi_k)$ and $\pi_k = \Pr(k \in s)$ denotes the probability that unit k belongs to s. In the case of full response, we know that

$$\sum_{k\in s}\frac{1}{\pi_k}\mathbf{x}'_{k\cdot}e_k^U$$

is an unbiased estimator of $\mathbf{X}'_U\mathbf{e}^U$, which is equal to zero. That means

$$E_p(\sum_{k\in s}\frac{1}{\pi_k}\mathbf{x}'_{k\cdot}e_k^U) = \mathbf{0}.$$

Deng & Chhikara (1991) show by using an important result from David & Sukhatme (1974) that, under simple random sampling without replacement, $\hat{\mathbf{B}}_s$ is an asymptotically unbiased estimation of $\hat{\mathbf{B}}_U$. Their set-up for asymptotic theory in a finite population is as follow. Let a sequence of finite populations of sizes N_n be drawn randomly from a superpopulation. The N_n are strictly increasing so that $n = f_nN_n$, $\lim f_n = f$, as $n \to \infty$ for $0 < f < 1$ and $0 < f_n < 1$ for all n. It is also assumed that $\mathbf{X}'_s\mathbf{D}_s\mathbf{X}_s$ is nonsingular for all possible samples. In practice, as recalled in David & Sukhatme (1974) we only have one finite population of size N from which we draw a simple random sample of size n. Under these settings and conditions, Deng & Chhikara (1991) show the following Lemma which does not need a superpopulation model, except for studying the asymptotic behaviour. Note that, since we calibrate from the r-level on the s-level, as exposed in Section 5.4, one could consider the population U as a superpopulation and the sample s as the "population".

Lemma 5.1 *(Deng & Chhikara, 1991). Let $\hat{\mathbf{B}}_s$ and $\hat{\mathbf{B}}_U$ be defined as in* (5.10) *and* (5.9)*, respectively.*

$$\hat{\boldsymbol{B}}_s - \hat{\boldsymbol{B}}_U = O_p(n^{-\frac{1}{2}}) = \left(\frac{1}{N}\boldsymbol{X}'_U\boldsymbol{X}_U\right)^{-1}\left(\frac{1}{n}\boldsymbol{X}'_s\boldsymbol{D}_s\boldsymbol{e}_s^U\right) + O_p(n^{-1}),$$

with $e_{s,k}^U = e_k^U$ for $k \in s$ and provided that $\boldsymbol{X}'_U\boldsymbol{e}^U = 0$.

It follows from the Lemma that

$$E_p(\hat{\mathbf{B}}_s) = \hat{\mathbf{B}}_U + O(n^{-\frac{1}{2}}),$$

and

$$E_p\left(\frac{1}{n}\mathbf{X}'_s\mathbf{D}_s\mathbf{e}_s^U\right) = E_p\left(\frac{1}{n}\sum_{k\in s}\frac{1}{\pi_k}\mathbf{x}'_{k\cdot}e_k^U\right) = O(n^{-\frac{1}{2}}).$$

The validity of these results is maintained for a stratified random sample in each stratum. Now, suppose that, because of a MAR nonresponse mechanism Ψ, we observe only $r \subset s \subset U$. That nonresponse mechanism can be seen as a second stage of stratified random sample in the sampling design. Let ψ_k be the probability that unit k responds. Assume that ψ_k can be entirely modelled with the help of some known variates by the construction of response homogeneity groups: ψ_k is known for all k in r. Then, the adjusted regression at the level of the sample r

$$y_k = \mathbf{x}_{k\cdot}\hat{\mathbf{B}}_r + e_k^r,\ k \in r, \tag{5.11}$$

where $\hat{\mathbf{B}}_r = (\mathbf{X}_r'\mathbf{D}_r\mathbf{X}_r)^{-1}\mathbf{X}_r'\mathbf{D}_r\mathbf{y}_r$, $\mathbf{D}_r = \text{diag}(1/\pi_k \cdot 1/\psi_k)$, would still yields an asymptotically unbiased estimation $\hat{\mathbf{B}}_r$ of $\hat{\mathbf{B}}_U$ for the same reasons as those at the s-level. In particular, we will still have

$$E_p E_\psi\left(\frac{1}{n_r}\mathbf{X}_r'\mathbf{D}_r\mathbf{e}_r^U\right) = E_p E_\psi\left(\frac{1}{n_r}\sum_{k\in r}\frac{1}{\pi_k}\frac{1}{\psi_k}\mathbf{x}_{k\cdot}'e_k^U\right) = O(n_r^{-\frac{1}{2}}).$$

Note that equation (5.11) means the same as (5.4) and (5.5).

Source of the problem: suppose that the response probability ψ_k has been unperfectly modelled or that it is not missing at random (NMAR). This is often the case in practice because the survey practitioner does not have all the variables explaining the nonresponse at disposal or the nonresponse depends partly on the variable of interest itself. Nevertheless, with what is at his disposal, he estimates ψ_k with $\widehat{\psi}_k$ by modelling the nonresponse through a logit model, a segmentation algorithm or by conducting some calibration. In addition, some of the auxiliary variables (columns of $\mathbf{X}_r$) may be subject to measurement errors (respondents lying, not recalling themselves precisely some amounts or roughly rounding up values for example). All together, the result is that, for some $\mathbf{x}_{\cdot j}$'s, if we could verify it,

$$\sqrt{n_r}\, E_p E_\Psi\left(\frac{1}{n_r}\sum_{k\in r}\frac{1}{\pi_k}\frac{1}{\widehat{\psi}_k}x_{kj}e_k^U\right) \text{ can be very large,}$$

for one or several j's. Then $\hat{\mathbf{B}}_r$ will be a biased estimator of $\hat{\mathbf{B}}_U$.

However, if we dispose of some other variables grouped in matrix $\mathbf{Z}_r$, such that, on r:

(i) $\frac{1}{n_r}\sum_{k\in r}\frac{1}{\pi_k}\frac{1}{\widehat{\psi}_k}z_{kj}e_k^U = O_p(n_r^{-\frac{1}{2}})$ for all $j = 1, ..., J$, which means that the instrument $\mathbf{z}_{\cdot j}$ is valid or *exogenous*. The instruments are centered with respect to the weighted population residuals. If this condition is not satisfied, $\mathbf{z}_{\cdot j}$ is said to be *endogenous*.

(ii) $\left(\frac{1}{n_r}\mathbf{Z}_r'\hat{\mathbf{D}}_r\mathbf{X}_r\right)^{-1}$ exists, with $\hat{\mathbf{D}}_r = \text{diag}(1/\pi_k \cdot 1/\widehat{\psi}_k)$.

Then $\hat{\mathbf{B}}_U$ can still be estimated via the so-called *weighted instrumental variable regression estimator*:

$$\begin{aligned}\hat{\mathbf{B}}_{rzx} &= (\mathbf{X}_r'\mathbf{P}_z\hat{\mathbf{D}}_r\mathbf{X}_r)^{-1}\mathbf{X}_r'\mathbf{P}_z\hat{\mathbf{D}}_r\mathbf{y}_r \\ &= (\mathbf{Z}_r'\hat{\mathbf{D}}_r\mathbf{X}_r)^{-1}\mathbf{Z}_r'\hat{\mathbf{D}}_r\mathbf{y}_r,\end{aligned}$$

where $\mathbf{P}_z = \mathbf{Z}_r(\mathbf{Z}_r'\mathbf{Z}_r)^{-1}\mathbf{Z}_r'$ is the hat-matrix (projector on L_z, the hyperplane generated by the instrumental variables). Indeed, by reusing (5.9):

$$\begin{aligned}\hat{\mathbf{B}}_{rzx} &= (\mathbf{Z}_r'\hat{\mathbf{D}}_r\mathbf{X}_r)^{-1}\mathbf{Z}_r'\hat{\mathbf{D}}_r(\mathbf{X}_r\hat{\mathbf{B}}_U + \mathbf{e}_r^U)\\ &= \hat{\mathbf{B}}_U + (\mathbf{Z}_r'\hat{\mathbf{D}}_r\mathbf{X}_r)^{-1}\mathbf{Z}_r'\hat{\mathbf{D}}_r\mathbf{e}_r^U,\\ &= \hat{\mathbf{B}}_U + \left(\frac{1}{n_r}\mathbf{Z}_r'\hat{\mathbf{D}}_r\mathbf{X}_r\right)^{-1}\left(\frac{1}{n_r}\mathbf{Z}_r'\hat{\mathbf{D}}_r\mathbf{e}_r^U\right),\end{aligned}$$

with $e_{r,k}^U = e_k^U$ for $k \in r$. Thus, under conditions (i) and (ii)

$$\hat{\mathbf{B}}_{rzx} = \hat{\mathbf{B}}_U + O(n_r^{-\frac{1}{2}}).$$

Which implies that $\mathbf{e}^g = \mathbf{e}^r$ is an estimate of $\mathbf{e}_r^U$, where $\mathbf{e}^g$ are the residuals from the generalized calibration in the linear case defined in (5.7) with $w_k = 1/(\pi_k \cdot \widehat{\psi}_k)$ and $\mathbf{e}^r$ is defined in (5.11).

Condition (i) is fulfilled if $1/n_r(\mathbf{z}_{.j}'\hat{\mathbf{D}}_r\mathbf{e}_r^U) = O(n_r^{-\frac{1}{2}})$ or if every $\mathbf{z}_{.j}$ can counterbalance the $1/\widehat{\psi}_k$ (for instance if $z_{kj} = \widehat{\psi}_k/\psi_k$). The first means that the weighted population residuals are orthogonal to $\mathbf{z}_{.j}$, it cannot be checked. The second means that the $\mathbf{z}_{.j}$'s completely explain the nonresponse. Empirically, one would check if the correlation between $\mathbf{z}_{.j}$ and $\mathbf{e}_r^U$ is small enough (see Section 5.8). The problem is that $\mathbf{e}_r^U$ is unknown. In some situations, the statistician has other ways to assess that some instruments are valid. For example through his knowledge of the research field or by using established physical laws. We will use a graphical test, based on an appreciation of GB2-fits, to identify the good instruments, as explained in Section 5.4.2.

In a survey sampling context, the idea is that, for some auxiliary variables, the true $\mathbf{x}_{kj}$, $k \in U$, at the population level are not endogenous, but their observed portion x_{kj}, $k \in r$, may be. In other words, the nonresponse mechanism may have distorted some of the auxiliary variables or the not perfectly modelled weights make their observed part appear as endogenous. Good instruments will be those being able to describe the nonresponse mechanism as we saw and as stated by Lesage & Haziza (2013a,b).

As stated in Davidson & MacKinnon (1993), intuitively, the procedure through instrumental variables, goes as follows: Weighted Least Squares (WLS) minimize the distance between y and L_x. But if some of the $\mathbf{x}_{.j}$ are endogenous, the WLS-estimate is inconsistent. The space of n_r-dimensions where y is a point can be divided into two subspaces orthogonal to one-another: L_z and $L_z^{\perp}$. Weighted Instrumental Variable (WIV) regression minimizes only the part of distance between y and L_x which is in L_z. If $\mathbf{e}_r^U$ is orthogonal to L_z, any correlation between $\mathbf{e}_r^U$ and L_x must be in $L_z^{\perp}$. Thus, restricting the minimization within L_z avoids the consequences of correlations between $\mathbf{e}_r^U$ and $\mathbf{X}_r$.

The estimated total $\hat{t}_{ylinG}$, see (5.6), therefore appears hence as being the projection of t_y on L_x parallel to an orthogonal direction to L_z. Thus the closer the subspaces spanned by $\mathbf{X}_r$ and $\mathbf{Z}_r$ (condition (ii) implies that

$\mathbf{Z}_r'\hat{\mathbf{D}}_r\mathbf{X}_r$ is of maximal rank), the smaller the gap between t_y and $\hat{t}_{ylinG}$ (Sautory & Le Guennec, 2005; Deville, 2002).

By viewing generalized calibration as instrumental regression, we thus realize that generalized calibration will provide a stable and unbiased estimator $\hat{t}_{ylinG}$ if, on the level of the subsample r:

A) we are able to distinguish which $\mathbf{x}_{.j}$ are endogenous and which are exogenous,

B) we have as many instruments $\mathbf{z}_{.j}$ satisfying the mentioned condition as there are endogenous $\mathbf{x}_{.j}$.

The problem is that A) or condition (i) cannot be verified. As explained in the next section, we use the GB2-fit to find the best combination of $\mathbf{x}_{.j}$'s and $\mathbf{z}_{.j}$'s. The GB2-fit directly indicates that $\hat{\mathbf{B}}_{rzx}$ is unbiased and the conditions (i) and (ii) can be post-verified. Which means that endogeneity has been compensated. Which also means, in our framework, that the nonresponse has been compensated as well.

Note that, ultimately, we are not interested in $\hat{\mathbf{B}}_U$, see (5.9), neither are we interested in estimating $\hat{t}_y^{cal}$, see (5.1), but only in finding weights stemming from generalized calibration, capable of compensating for the nonresponse. The instrumental variable regression is only introduced here to help us develop an intuition of how to tune the generalized calibration and to verify that the $(\mathbf{X}_r;\mathbf{Z}_r)$-combination is suitable.

5.4.2 Choosing the auxiliary and instrumental variables

In practice, we would like to partition the matrix of auxiliary variables in an endogenous and an exogenous part, $\mathbf{X} = [\mathbf{X}_{end}', \mathbf{X}_{exo}']'$. The exogenous $\mathbf{x}_{.j}$ may be instrumented by themselves since they satisfy the desired conditions for instruments. The other endogenous $\mathbf{x}_{.j}$ corrupt the WLS-estimator of $\hat{\mathbf{B}}_U$ and make it non consistent. These variables cannot be used as instruments. Thus, valid instruments are $\mathbf{Z} = [\mathbf{Z}_\star', \mathbf{X}_{exo}']'$. One needs to find as many instruments in $\mathbf{Z}_\star$ as there are endogenous variables in $\mathbf{X}_{end}$. In summary, to provide an unbiased estimator for generalized calibration, it is necessary that, for the realized sample:

- the observed auxiliary variables in $\mathbf{X}_r$ and instrumental variables in $\mathbf{Z}_r$ must be as correlated as possible, with $\mathbf{Z}_r'\hat{\mathbf{D}}_r\mathbf{X}_r$ of maximal rank invertible (condition (ii)), but at the same time, $\mathbf{Z}_r$ must not be correlated to the unknown population residuals $e_k^U = y_k - \mathbf{x}_{k.}\hat{\mathbf{B}}_U$, $k \in r$ (condition (i)).

- the $\mathbf{Z}$ should describe the nonresponse mechanism (condition (i)) and $\mathbf{X}$ should explain y as we supposed in (5.9), (5.10) and (5.11). In other words, once the $\mathbf{z}_{.j}$ are taken into account, the remaining unexplained part of nonresponse is completely ignorable.

A balance must be found between these two requirements on the $\mathbf{Z}_r$. Otherwise we risk a strong bias in estimating $\hat{\mathbf{B}}_{rzx}$, see (5.8), and thus $\hat{t}_y^{cal}$, see (5.1), as well as an increase of the estimated variance of t_{ylinG}, see (5.6),

which corresponds to the findings of Lesage & Haziza (2013a,b) who interpret the calibration function $F(\mathbf{z}_k.\boldsymbol{\lambda})$ as a model of the inverse of the response probability, observing the formulae for the bias and the variance.

The novel idea of this paper is to use GB2-adjustments relying on weights obtained through generalized calibration in order to identify the best combination of auxiliary and instrumental variables, given two lists of potentially eligible variables of both kinds. The first list contains potential auxiliary variables. These must explain y and their total must be known on s. The second list contains potential instrumental variables, they explain the nonresponse and are observed at least on r. Yet, a priori, we do not know which of the auxiliary variables are endogenous, if any, neither do we know if all our instruments are valid. The best combination of auxiliary and instrumental variables for instrumental regression will also correspond to the generalized calibration providing the most efficient weights in the sense that these weights will correct the nonresponse. As will be illustrated in Section 5.8, the best combination enabled us to retrieve almost the same GB2-adjustment that one would find without nonresponse. It is the combination where endogenous auxiliary variables are replaced by valid instruments. Thus the GB2-adjustment relying on weights obtained through generalized calibration is used as an empirical test: we graphically evaluate how the GB2-density fits the empirical density of the income distribution as in Figure 5.1, to identify weights capable of compensating the nonignorable nonresponse.

5.5 Ranks, robustness and normal scores

5.5.1 Weighted ranks

In order to robustify the income data, we will consider only their weighted ranks. In this way, extreme or *outlier* (in the sense of Chambers, 1986; Hulliger, 1999) observations will not be able to corrupt our regression models as we will see. Without loss of generality, to simplify the notation, assume that the values of y_k are distinct and sorted by order of magnitude, so that $y_k = y_{[k]}$.

Definition 5.1 *The weighted rank is defined by:*

$$R_m^w = \frac{1}{\sum_{k=1}^{n_r} w_k}\left(\sum_{k=1}^{m} \mathbf{1}_{y_m \leq y_k} w_k - \frac{w_m}{2}\right) \in [0, n_r] \subset \mathbb{R},$$

for $m = 1, 2, ..., n$, if y is defined for n_r values associated with weights w_m.

Note that if the data contains ties for y (duplicates), which often happens in practice because the respondents have rounded up some values, or because of ranged questions, then one could add a randomly chosen very small number to each value of y so that the values can be sorted uniquely.

If $\mathbf{w}_{std}$ are the standardized weights (i.e. $\sum_k w_{std,k} = n$ and $\overline{w_{std}} = 1$), then

$$R_m^w = \sum_{k=1}^{m} \mathbf{1}_{y_m \leq y_k} w_{std,k} - \frac{w_{std,m}}{2}. \tag{5.12}$$

The rank of the weighted rank is always equal to the rank since Definition 5.1 is equivalent to

$$R_m^w = \sum_{k:y_k \leq y_m} w_{std,k} - \frac{w_{std,m}}{2}.$$

5.5.2 Normal scores

We aim to perform a probit transformation of calculated ranks so that the variable we impute satisfies the usual conditions required for conducting a regression model (normal distribution). By dividing the ranks by the number of non missing values plus one, we obtain values ranging between 0 and 1 that can be considered as originating from a uniform distribution.

If we denote by $\Phi(\cdot)$ the cumulative distribution function of the standardized normal distribution $\mathcal{N}(0,1)$, Φ is a continuous, monotonically increasing function, its image is $(0,1)$:

$$u = \Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2} dt,\ z \in \mathbb{R}, \tag{5.13}$$

The probit function or quantile function is defined by $\Phi^{-1}(u) = probit(u)$. If $u \sim \mathcal{U}(0,1)$, $\Phi^{-1}(u)$ generates a random variable following a $\mathcal{N}(0,1)$ distribution. Yet the (weighted) standardized ranks are not a sequence of probabilities but they take the same range of values and we consider the corresponding transformed variable (which is analogue to normal quantiles if the ranks would follow a $\mathcal{U}(0,1)$):

$$Q^w = \Phi^{-1}\left(\frac{R^w}{n+1}\right) \tag{5.14}$$

The Q^w have the same shape as a normally distributed variable. In the literature, the Q^w are called *normal scores* and the transformation applied is known as the van der Waerden's method (Conover, 1999).

Note that instead of using the ranks, one could also operate a probit-transformation of the GB2-quantiles, $F_{GB2}(y_k) \in [0,1]$, $k \in r$, and obtain transformed variables based on GB2-adjustments on the respondents using a different set of weights. This has been tested, but we did not include these results in this article since they were not better than the ones with the above-mentioned method.

5.6 Imputation strategy

Our imputation system should fulfil the following conditions: transparency, the model(s) used must be known; robustness, extreme values must not corrupt the imputation model. As far as possible it must be able to cope with nonignorable nonresponse and respect the natural distribution of the income variable considered; it should allow the use of survey weights; it should permit variance estimation. Our imputation strategy can be summarized as follows, each of these steps is illustrated and explained more in detail in Section 5.8:

Step 1 For the respondents to y, produce adjusted weights, w_k^{cal}, through generalized calibration. The problem is to find a good combination of auxiliary and instrumental variables so that the generalized calibration produces weights capable of compensating the nonresponse that affects the variable of interest. A GB2-adjustment based on these weights is used to evaluate their relevance. We seek weights enabling us to approach the best possible GB2, that is the one we would get by fitting on all units (if all would have responded). The variable y is used, among others, as instrument: $z_{k1} = y_k$. See Section 5.8 for how the auxiliary and instrumental variables are concretely chosen.

Once the weights are computed, the poverty and inequality measures can already be estimated empirically at this stage without doing any imputation by using these weights for compensating for nonresponse. Moreover, the indicators can also be estimated parametrically through the GB2-adjustment that helped to tune the generalized calibration (see Section 5.3).

Step 2 Order the income variable by increasing weighted ranks, see Definition 5.1. The hope is that the use of weighted ranks with weights computed in Step 1 will take into account the nonignorable nonresponse. This robustifies the imputation system because we shall first impute ranks (in Step 4) and estimate the y-values later on (in Step 6).

Step 3 Compute normal scores by using the van der Waerden's method, based on ranks, see (5.14). A Q-Q-plot of the quantiles produces a straight line by construction.

Step 4 Impute by predicting the missing quantiles with the help of a classical weighted multiple linear regression model based on auxiliary variables $\mathbf{x}_{k\cdot}$ and the weights w_k^{cal} computed in Step 1 (see Section 5.8 for more details). The use of weights stemming from calibration compensates for NMAR nonresponse.

Step 5 Back transformation: obtain imputed ranks from the predicted normal quantiles of Step 4.

Step 6 Knowing the ranks, estimate values for the variable of interest by either using the adjusted GB2 from Step 1, or by some interpolation between the two nearest (with respect to ranks) known responding values.
The adjusted GB2 can be used but is not necessary to perform the imputation. It is mainly used to tune the generalized calibration or, optionally, to produce parametric estimates of the inequality indicators.

5.7 Estimating the variance

One needs some procedure to estimate the variance of statistics based on the imputed dataset. Variance may originate from several sources (see

for instance Eurostat, 2013, for a more complete description): sampling variance, nonresponse variance, imputation variance, over-coverage variance, response-variance. When one seeks to estimate the variance of complex statistics like inequality indicators, additional problems arise from the fact that the income distribution is hard to model and that indicators for poverty and social exclusion are non linear functions of y. Many estimation methods can lead to adequate estimator of variance under the commonly used sampling design. Most of the time, there are no analytic formulae to compute the variance of inequality measures. Briefly, one has the choice between two main philosophies: *replication methods* and *linearization methods*. We have opted for linearization methods because a lot of work has already been done to estimate the variance of the Swiss SILC data that we use through such methods (Massiani, 2013b). More precisely, we use generalized linearization methods, mainly introduced by Deville (1999b) and Demnati & Rao (2004c), as other authors did, see for example Osier (2009); Goga et al. (2009); Verma & Betti (2011); Langel & Tillé (2013). In this fieldwork, we revisited the results and brought some improvements to the estimation of the income density function (Graf & Tillé, 2014). There are other options for linearization, such as estimating equations as Kovacevic & Binder (1997) did.

Linearization is not only used to linearize complex statistics, but also to take the effect of calibrations applied to obtain the final survey weights into account. Indeed, Deville (1999b) shows that estimating the variance of calibrated estimators (Deville & Särndal, 1992) is equivalent to estimating the design variance of the residuals from the regression of the variable of interest (for us income) on the auxiliary variables used to calibrate. In fact the residuals of that regression are the linearized variable with respect to the calibration. Massiani (2013b) details the procedure to estimate the variance of a total for the Swiss SILC survey. The entire procedure used to produce the final survey weight is considered, taking into account the stratified random cluster sample, nonresponse corrections on several levels, weight sharing (see Lavallée, 2002), panel combinations and calibration on known totals (see Graf, 2008a, for a detailed description of the Swiss SILC weighting scheme). So our "design variance" estimates - $\hat{V}_p(\cdot)$ - take into account the full design including the complex weighting scheme of the Swiss SILC survey.

To estimate the variance of indicators expressed as functions of the four parameters, $\boldsymbol{\theta} = (a, b, p, q)'$, of the GB2 distribution, we followed the procedure proposed by Graf & Nedyalkova (2013), and applied the Delta method (see, e.g. Davison, 2003). If we denote by $\hat{A} = A(\hat{\boldsymbol{\theta}})$,the maximum likelihood estimator of a certain indicator A (e.g. A =GINI, ARPR, QSR,...), then, by the Delta method:

$$\widehat{\text{var}}(\hat{A}) = \frac{\partial \hat{A}}{\partial \hat{\boldsymbol{\theta}}}' \, \widehat{\text{var}}(\hat{\boldsymbol{\theta}}) \frac{\partial \hat{A}}{\partial \hat{\boldsymbol{\theta}}}, \tag{5.15}$$

where $\widehat{\text{var}}(\hat{\boldsymbol{\theta}})$ is the variance estimation of the parameters of the GB2 distribution. The partial derivatives of indicators with respect to the four parameters can be calculated numerically. As Graf & Nedyalkova (2013)

describe it, $\widehat{\text{var}}(\hat{\boldsymbol{\theta}})$ is the linearized sandwich variance estimator (Huber, 1967; White, 1980, 1982; Freedman, 2006; Pfefferman & Sverchkov, 2003):

$$\widehat{\text{var}}(\hat{\boldsymbol{\theta}}) \approx [\frac{\partial^2 l(\hat{\boldsymbol{\theta}})}{\partial \hat{\boldsymbol{\theta}}^2}]^{-1} \hat{V}_p(\hat{\mathbf{U}}) [\frac{\partial^2 l(\hat{\boldsymbol{\theta}})}{\partial \hat{\boldsymbol{\theta}}^2}]^{-1}, \tag{5.16}$$

where $\partial^2 l(\hat{\boldsymbol{\theta}}) / \partial \hat{\boldsymbol{\theta}}^2$ is the second derivative of the log-likelihood

$$l(\hat{\boldsymbol{\theta}}) = \sum_{k \in r} w_k \log[f_{GB2}(y_k; \hat{\boldsymbol{\theta}})],$$

and $\hat{V}_p(\hat{\mathbf{U}})$ is the estimated design variance of the scores evaluated at $\hat{\boldsymbol{\theta}}$:

$$\hat{V}_p(\hat{\mathbf{U}}) = \widehat{\text{var}}[\frac{\partial l(\hat{\boldsymbol{\theta}})}{\partial \hat{\boldsymbol{\theta}}}] = \widehat{\text{var}}[\sum_{k \in r} w_k \mathbf{u}_k(y_k; \hat{\boldsymbol{\theta}})]; \tag{5.17}$$

which is a 4×4 matrix of weighted totals of GB2 scores $\mathbf{u}_k = (u_{ka}, u_{kb}, u_{kp}, u_{kq})'$. These scores are the partial derivatives of the log-likelihood in y_k estimated at $\hat{\boldsymbol{\theta}} = (\hat{a}, \hat{b}, \hat{p}, \hat{q})$. They are themselves estimates of the scores evaluated at the true unknown parameter $\boldsymbol{\theta}$. The scores are then considered as a measured variable from which we want to estimate the variance of the weighted total. We then linearize with respect to the weights according to the sampling design and weighting scheme used in the Swiss SILC as we did above to estimate the variance of the empirically calculated indicators.

We have worked on estimating the variance due to imputation (see Graf, 2014) attempting, without performing multiple imputation, to take the imputation into account analytically and estimate its contribution to the variance. We have obtained a satisfactory solution of constructing an unbiased estimator of the total variance (design variance plus imputation variance) for quantiles based on a solution that Deville & Särndal (1994) used to estimate the variance of the total of y under imputation by regression prediction. Nevertheless, everything is still not solved for complex statistics such as poverty indicators. Another way to estimate the part of variance due to imputation is of course to perform multiple imputations (Rubin, 1987) which is what we did for this paper by generating 50 additional values to each predicted normal quantile (in Step 4, see Section 5.6). The 50 values were randomly chosen according to a normal distribution centred on the predicted value from the regression model and of variance corresponding to the estimated variance for predicted values. By the law of total variance, for a statistic A,

$$V_{tot}(A) = E_p[V_{imp}(A)|p, s] + V_p[E_{imp}(A)|p, s] \tag{5.18}$$

We took the empirical variance $s^2_{\hat{y}_{imp}}$ of the imputed values, based on 50 imputed values for each missing unit, as an estimate of the first term. The second term is the design variance of the expected value of A with respect to the imputation, which is based on the predicted values from the regression model.

5.8 Real data application

We tested our imputation system on the Swiss SILC data 2009 provided by the Swiss Federal Statistical Office conducting the survey in Switzerland. We are particularly interested in the employee's income. For that year, this mainly Computer Assisted Telephone Interview (CATI) survey could be matched with the register of the Central Compensation Office (CCO). This provided a unique opportunity, to compare the income variable collected through the CATI process to the official value of the CCO register. From year 2010 on, thanks to that matching, the CCO register has been used every year, saving in this way some questions in the CATI survey. Thus, the year 2009 was of particular interest for methodological developments since both information sources, CATI and register, were at our disposal.

The CATI variable is affected by an unknown nonresponse mechanism which is NMAR as we will see. The CCO register variable shows almost no missing values. In reality, without the register, we would be willing to impute the CATI variable, y_{CATI}, but here, we have applied the observed nonresponse mechanism affecting y_{CATI} on the corresponding register variable y_{CCO}. We then impute the register variable and can compare the result with the known truth. More precisely, y_{CCO} shows $n = 7,922$ values ($= s$) and among these, y_{CATI} has been collected for 6,884 (86.9%) respondents. In order to be able to use two other important instrumental variables (see below), the set of the common respondents to all three variables reduces to $n_r = 6,188$ (78.1%)($= r$). So 21.9% of the values have been artificially erased from y_{CCO}. This applied nonresponse mechanism is real because it has been observed through the CATI survey and does not need to be modelled. Note that on r, y_{CATI} and y_{CCO} correlate by 84.1%. The differences are due to the fact that, on the telephone, people tend to round up the quantities or do not remember precisely the amounts, or forget some components. The hope is still that a good imputation system on y_{CCO} would also work on y_{CATI}.

5.8.1 Segmentation

To verify that the nonresponse mechanism is nonignorable, it has been partially modelled by a segmentation tree (Kass, 1980) based on dichotomous auxiliary variables known on s. The segmentation method is an alternative to logistic regression allowing to form Response Homogeneity Groups (RHG). It is used in many surveys (Graf & Qualité, 2014) at the SFSO to compute nonresponse adjustment factors. Appendix E displays the segmentation tree and variables used. The segmentation produces 73 RHG's. The response rates to y_{CCO} in the 73 RHG's correlate by 39.3% with median values of y_{CCO} per RHG. This shows that the nonresponse applied to y_{CCO} (and affecting y_{CATI}) is not ignorable, it depends partly on the value of the variable of interest itself. Moreover, the high correlation, 77%, between median values of y_{CCO} per RHG for the respondents and the non-respondents is a sign that the segmentation tree creates good RHG's based on the auxiliary variables available. This can of course only be verified because we know the true values of y_{CCO} for the non-respondents.

5.8.2 Detailed imputation procedure

Step 1: generalized calibration and GB2-adjustments

To choose the auxiliary ($\mathbf{X}_s$ and $\mathbf{X}_r$) and instrumental ($\mathbf{Z}_r$) variables for generalized calibration, we have seen that the auxiliary variables must explain y_{CCO} and be available on s, and that the instrumental variables must describe the nonresponse mechanism. Therefore, we have retained as eligible auxiliary variables, those at disposal in the survey data correlating by at least 10% with y_{CCO}. As potential instrumental variables, we have first listed the variables involved in the mentioned segmentation tree modelling the nonresponse and/or which, at the same time, may explain y_{CCO}. Most of the variables are dichotomous, among these one finds regions identifiers, education levels, household types, activity status levels, individual's characteristics (age classes, sex, civil status, having children or not), employment rate levels, having managing position or not, receiving a 13th month salary or other benefits or not. A detailed list of these variables can be found in Appendix D. We have ended up with $J = 42$ variables, among which 38 are dichotomous. 39 $\mathbf{x}_{.j}$ are instrument for themselves and three have not been instrumented by themselves, instead, the following instrumental variables have been used:

- the variable of interest itself, $\mathbf{y}_{CCO}$,
- the occupation rate, TX_TP,
- the price of housing, HH_{cost},

all three known on the respondents'set only. This combination of auxiliary and instrumental variables was the one providing the best GB2 adjustment (see below). In order to control the theoretical conditions on the instruments presented in Section 5.4.1 and 5.4.2, we have run a WLS-regression and a WIV-regression. The WLS, as conventional calibration, uses only the auxiliary variable. The WIV uses the same combination as for the generalized calibration which provided the best GB2-adjustment. Table 5.1 presents the estimated values of the WLS-regression parameters $\hat{\mathbf{B}}_r$, see (5.5) and, respectively, the WIV-regression parameters $\hat{\mathbf{B}}_{rzx}$, see (5.8). One can observe that the two parameters estimates differ a lot. We then checked the condition (i), see Section 5.4.1, as well as computed the correlations between the realisations on r of the $\mathbf{x}_{.j}$ and the residuals e_k^r of the WLS-model, see (5.4), and the residuals e_k^g of the WIV-model, see (5.7). The same controls were computed with the instruments $\mathbf{z}_{.j}$. As mentioned, looking for the best GB2-adjustment has lead us to instrument three variables ($AC\_2, TXactLE80$ and $\log[\text{med}(y_{CCO}]$, see Table 5.1). Two of those were highly endogenous: $\text{corr}(AC\_2, \mathbf{e}^g) = 0.73$ and $\text{corr}(\log[\text{med}(y_{CCO})], \mathbf{e}^g) = 0.33$. But this could not be seen for variable AC_2 if one would only check the correlation with $\mathbf{e}^r$. It is not obvious that it is necessary to instrument the variable $TXactLE80$ from the point of view of compensating endogeneity. But the use of a third instrument helped to reach a better GB2-adjustment. Recall that the variable of interest itself, $\log(y_{CCO})$, was used as an instrument. Its low correlation with

the WIV-residuals indicates that it is a valid instrument, at the same time, of course its correlation with the WLS-residuals is very high.

Strictly considered, we used a logit calibration function, and the computations of Table 5.1 would more reflect what happens with a linear calibration function. Nevertheless, as we'll show next, the system is much more sensitive to the good choice of instruments as to the calibration function chosen. Thus, we argue that the comparison is possible. Note also that we applied a logarithm transformation to the interest variable as well as to the instrument "price of housing" and the median of y_{CCO} by RHG of the segmentation tree. This with the idea of limiting the undesirable effect of extreme observations in the generalized calibration. The compared regression models of Table 5.1 are then, logically, trying to explain the logarithm of y_{CCO}. Often, for generalized calibration, one would not pay too much attention to the fact of taking the logarithm transformed variables or not. But viewing the problem as an instrumental regression makes it a concern. Running the whole procedure with non log-transformed variables does not change any of our conclusions.

We tested three different pseudo-distance functions for calibration: logit, raking ratio and truncated. The linear method sometimes yielded negative weights and was therefore excluded. Table 5.2 presents and compares the weights before calibration to the ones after calibration. One can see that the weights do not differ strongly if one uses the different pseudo-distances for calibration. We have given preference to the logit method for reasons that will be explained later. At this stage, we may already compute empirical estimates of the indicators using the calibrated weights, see Table 5.4. It can be seen that none of these weighted estimates is significantly different from the true values. This means that the weights stemming from the generalized calibration are able to compensate for nonignorable nonresponse.

The two panels of Figure 5.1 show several GB2-density-fits done on the data. One can observe the following relevant points:

- the estimated empirical density on the full dataset shows two peaks. There are many low incomes. Obviously, this is far away from a normal distribution.

- by comparing the two estimated empirical densities on the full dataset and on the respondents only, one can see that there has been a stronger loss of respondents among the low incomes ($< 20,000$ swiss francs) as well as among those between, say, 60,000 and 80,000 Swiss francs,

- the GB2-density-fit on the complete dataset, with full response, is the best possible GB2-fit. It can be seen that it has to make a compromise between the two picks because the GB2 cannot be bimodal. Nevertheless it fits very well middle and high incomes. This fit is done by using the original survey weights.

- The GB2-density-fit based on the respondents only with non cor-

Table 5.1 – *Post-control of the generalized calibration by adjusting corresponding WIV-regression model and comparing it with a WLS-model.* $\boldsymbol{x}_{.j}$ *= auxiliary variables known on s,* $\hat{\boldsymbol{B}}_{rzx}$ *= estimated parameters through WIV-regression using the* $\boldsymbol{z}_{.j}$*, known at least on r, as instruments.* $\hat{\boldsymbol{B}}_r$ *= estimated parameters through WLS, i.e. all* $\boldsymbol{x}_{.j}$ *instrumented by themselves.* $corr(\boldsymbol{e}^g, \boldsymbol{x}_{.j})$ *= correlations between the observed* $\boldsymbol{x}_{.j}$*, on r, and the residuals of the adjusted WIV-regression (biserial correlation if* $\boldsymbol{x}_{.j}$ *dichotomous).* $corr(\boldsymbol{e}^r, \boldsymbol{x}_{.j})$ *= correlations between the observed* $\boldsymbol{x}_{.j}$*, on r, and the residuals of the adjusted WLS-regression.* $corr(\boldsymbol{e}^g, \boldsymbol{z}_{.j})$ *= correlations between the observed* $\boldsymbol{z}_{.j}$*, on r, and the residuals of the adjusted WIV-regression. The 5th and 7th columns reflect the test of condition (i), see Section 5.4.1. Note that* $n_r^{-\frac{1}{2}} = 0.013$*.*

39 variables instrumented by themselves						
$\mathbf{x}_{.j} = \mathbf{z}_{.j}$	$\hat{\mathbf{B}}_{rzx}$	$\hat{\mathbf{B}}_r$	$corr(\mathbf{e}^g, \mathbf{x}_{.j})$	$\frac{1}{n_r}\mathbf{x}'_{.j}\hat{\mathbf{D}}_r\mathbf{e}^g$	$corr(\mathbf{e}^r, \mathbf{x}_{.j})$	$\frac{1}{n_r}\mathbf{x}'_{.j}\hat{\mathbf{D}}_r\mathbf{e}^r$
AC_3	-1.847	-0.222	-0.03	0	-0.03	0
AC_5	-0.55	0.06	0	0	-0.01	0
ALLOCfam	-0.307	0.097	-0.01	0	0.01	0
CANTONo9_21	-0.207	-0.082	0	0	0	0
CANTONo9_25	-0.082	0.062	0.01	0	0	0
CHEF	-0.014	0.186	-0.01	0	0.01	0
COM2_09_2	0.091	0.069	0	0	-0.01	0
COM2_09_3	0.116	0.182	0	0	0	0
DIVORCE	-1.059	0.114	-0.01	0	-0.01	0
ECHOR_1	-0.088	0.052	0	0	-0.02	0
ECHOR_3	-0.162	0.043	0.01	0	-0.01	0
EDUC_12	-0.03	0.184	0	0	0	0
EDUC_14	0.111	0.291	-0.01	0	0	0
EDUC_15	0.086	0.339	-0.01	0	0	0
EDUC_16	0	0.445	0	0	0.01	0
EDUC_2	1.414	-0.2	0.08	0	-0.01	0
EDUC_6	-0.025	0.003	-0.02	0	-0.01	0
HHtypeo9_3	-0.105	0.076	-0.01	0	0	0
HHtypeo9_4	-0.172	0.039	0	0	0	0
HHtypeo9_9	0.006	-0.107	0	0	0.01	0
HOMME	0.173	0.186	0.02	0	0.02	0
KIDS	0.466	-0.063	0.02	0	-0.01	0
LOGSE2oom2	-0.026	0.028	0.01	0	-0.01	0
LOGm2MISS	0.26	0.108	-0.01	0	-0.02	0
MAISON	0.088	-0.016	0.02	0	-0.01	0
MARIE	-1.217	0.077	-0.03	0	-0.01	0
NPACT_2	0.15	0.014	-0.02	0	-0.02	0
Ncall11TO35	0.147	0.026	0	0	0	0
Ncall4TO10	0.099	0.011	0.01	0	-0.01	0
OCCUP_13	6.841	-0.505	0.04	0	0.02	0
PROPRIO	0.222	0.034	0.02	0	0	0
SANTE_1	0.114	0.043	0	0	-0.01	0
TRAVWE	0.023	-0.011	0.01	0	0	0
TXactE100	0.381	0.194	0.02	0	0.01	0
TXactL100	0.483	0.452	0	0	0	0
actINDEP	-0.035	-0.049	-0.01	0	-0.02	0
prestaCHOM	-0.587	-0.229	-0.01	0	0	0
prestaCP	-0.076	0.135	0	0	0.02	0
salair1314	-0.036	0.093	0	0	-0.03	0
3 instrumented auxiliary variables						
$\mathbf{x}_{.j}$	$\hat{\mathbf{B}}_{rzx}$	$\hat{\mathbf{B}}_r$	$corr(\mathbf{e}^g_k, \mathbf{x}_{.j})$	$\frac{1}{n_r}\mathbf{x}'_{.j}\hat{\mathbf{D}}_r\mathbf{e}^g$	$corr(\mathbf{e}^r, \mathbf{x}_{.j})$	$\frac{1}{n_r}\mathbf{x}'_{.j}\hat{\mathbf{D}}_r\mathbf{e}^r$
AC_2	-14.863	-0.125	0.73	0.31	0.01	0
TXactLE80	0.281	0.265	-0.02	-0.01	-0.01	0
$\log[\text{med}(y_{CCO})]$	1.05	0.936	-0.33	-0.45	-0.24	0
3 instruments						
$\mathbf{z}_{.j}$			$corr(\mathbf{e}^g_k, \mathbf{z}_{.j})$	$\frac{1}{n_r}\mathbf{z}'_{.j}\hat{\mathbf{D}}_r\mathbf{e}^g$	$corr(\mathbf{e}^r, \mathbf{z}_{.j})$	$\frac{1}{n_r}\mathbf{z}'_{.j}\hat{\mathbf{D}}_r\mathbf{e}^r$
TX_TP			0.02	0	0.05	1.70
$\log(HH_{cost})$			0.05	0	0.09	0.10
$\log(y_{CCO})$			0.01	0	0.56	0.59

Table 5.2 – *Weights for $\boldsymbol{y}_{CCO}$, on r, before and after calibration. The bounds for the truncated and logit method were (0.1,10).*

	Min.	Q_1	Median	Mean	Q_3	Max.
w_{1k}	85.6	302.2	372.1	428.1	499.8	4583.0
$1 + F^{logit}(\mathbf{z}_{k.}\boldsymbol{\lambda})$	1.00	1.02	1.07	1.31	1.19	11.25
$w_k^{cal,logit}$	85.8	340.7	439.4	550.1	616.3	5901.0
$F^{rak}(\mathbf{z}_{k.}\boldsymbol{\lambda})$	0.07	0.40	1.14	1.30	1.44	7.29
$w_k^{cal,rak}$	24.6	333.8	459.0	549.1	651.9	5461.0
$F^{trun}(\mathbf{z}_{k.}\boldsymbol{\lambda})$	0.10	0.93	1.20	1.30	1.56	4.24
$w_k^{cal,trun}$	25.6	327.4	472.5	548.8	671.9	5655.0

rected survey weights shows a very different shape. It naturally follows the empirical density of the respondents only.

- The GB2-density-fit based on the respondents only with corrected weights obtained through the generalized calibration conducted in Step 1 almost falls back on the best possible GB2-fit obtained with full response! This shows that the new weights are able to compensate for the nonignorable nonresponse mechanism.

- If we compare the GB2-fit obtained through generalized calibration to the one obtained through conventional calibration, we see that conventional calibration provides a fit performing better than with no calibration but worse than with generalized calibration. Clearly, generalized calibration is here better than conventional calibration thanks to the three instrumental variables we used.

- The lower panel of Figure 5.1 shows that the GB2-density-fits vary little if we change the pseudo-distance used in the generalized calibration. The methods provide similar GB2-fits. In the end of the imputation procedure, the logit method showed the best results.

At this stage we are already able to compute parametric estimates of the inequality indicators (see Table 5.4) based on the GB2-fit. As can be observed, none of the parametric estimates stemming from the GB2-fits with weights from generalized calibration, except for the ARPT, are significantly different from the true values on the full response sample. This again means that generalized calibration enabled us to find a very satisfactory fit of the true income distribution. Indeed, the estimates stemming from the parametric GB2-fit on the respondents only, without using corrected weights, as well as the empirical estimates on the respondents only, are all significantly biased. More comments on Table 5.4 are given under Step 6 below.

Steps 2 and 3: compute weighted ranks and normal scores

Weighted ranks and normal scores were computed according to the description made in Section 5.5. The weights stemming from the different

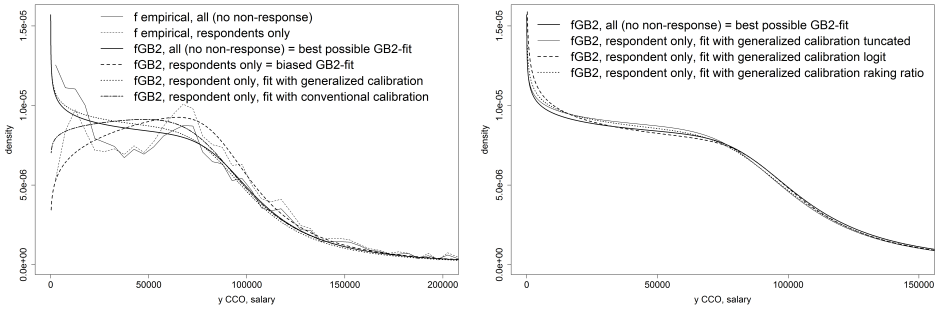


Figure 5.1 – *Step 1: Income densities and GB2-fits. Left panel, comparison between the data and the GB2-fits. Right panel, GB2 fits using weights from generalized calibration with different calibration functions: truncated (bounds=0,10), logit (bounds=0,10) and raking ratio.*

generalized calibrations were used with the hope that they will compensate for NMAR nonresponse. Thus, we compute normal scores Q^{wcal} defined on the respondents' set r.

Step 4: imputation

We adjust a classical weighted multiple linear regression model on the normally shaped Q^{wcal}. We reuse the weights from the generalized calibration. The model includes the 42 auxiliary variables of the calibration in Step 1, all of them are dichotomous except one which is continuous. The continuous variable is the median values of y_{COO} over the 73 leaves of the segmentation tree described in Section 5.8.1. The coefficient of determination of the model is over 63%. We have seen that two auxiliary variables are highly endogenous, but the use of the weights stemming from the generalized calibration compensate for this.

One performs the imputation by predicting the missing quantiles with the help of the multiple linear regression model:

$$\hat{Q}_o^{wcal} = \mathbf{X}_o \hat{\boldsymbol{\beta}}$$

with $\hat{\boldsymbol{\beta}} = (\mathbf{X}_r' \widehat{\mathbf{W}} \mathbf{X}_r)^{-1} \mathbf{X}_r' \widehat{\mathbf{W}} Q_r^{wcal}$, $\widehat{\mathbf{W}} = diag(w_k^{cal,logit})$. The values $\hat{Q}_o^{wcal}$ are expected imputed values, no residual has been added. For the 50 imputations that we also produced in order to estimate the variance due to imputation, we added 50 times a residual randomly drawn from the predicting distribution.

Step 5: back-transformation, obtain ranks

Back transformation: obtain imputed ranks from the predicted normal scores (quantiles):

$$\hat{R}_o = \Phi(\hat{Q}_o^{wcal}) \in]0,1[$$

Step 6: get imputed values

The imputed values $\hat{R}_o$ of each non-respondents intercalate between the respondent's ranks, by renumbering all these ranks from 0 to $n = n_r + n_o$, one defines the imputed ranks. For each non-respondent t, we can find two respondents $y_{L;t}$ and $y_{U;t}$ respectively having the closest rank smaller and larger than $\hat{R}_{o;t}$. The predicted income is then estimated for example by simple interpolation:

$$\hat{y}_{imp,t} = \frac{y_{L;t} + y_{U;t}}{2}, \tag{5.19}$$

which is what we have done leading to the results commented above. Note that we could also use the fitted GB2 of Step 1:

$$\hat{y}^{GB2}_{imp,t} = F^{-1}_{GB2}(\hat{R}_{o;t}). \tag{5.20}$$

Nevertheless, this leads to still good but slightly worse results.

5.8.3 Results

Figure 5.2 compares the empirical income density distributions of the imputed values to the (usually unknown) corresponding distribution of the real values, and to the distribution of the respondents. Again we can see that the respondents and the non-respondents show different distributions. Concerning the imputed values, their distribution follows very closely the one targeted for the non-respondents except for the very low (annualized) incomes below 20,000 Swiss Francs. This is also reflected in Table 5.4: the empirical estimates of the inequality indicators based on the partially imputed dataset are not significantly different from the true values except for the QSR! The latter indicator is by definition more sensitive to the lower incomes and thus to the precision of our imputation in that region of the income distribution. But these inequality measures depend only on the shape of the distribution. What about the precision at the unit level? The four bottom panels of Figure 5.2 show that for the half of the imputed values we commit an absolute error comprised between $[-9,896; 19,881]$ Swiss Francs, and for 90% of the imputed values, the absolute error lays within $[-38,954; 70,716]$ Swiss Francs. The two lower panels of Figure 5.2 confirm, as can be suspected by observing the densities of the top panel, that we are relatively less precise in the lower tail of the distribution, that means for low income. This originates from the fact that the low incomes were hard to model with the information at our disposal. For middle to high income, we have very good results. Indeed, to compare our performance, we have conducted other imputations using IVEware. The same auxiliary variables were provided in input, and we have imputed the log-transformed y_{CCO} variable as currently done at the SFSO. Note that in the SFSO's imputation procedure, the imputation is split into imputation classes[2] and some bounds restricting the allowed

[2]See Appendix F for more details on the creation of imputation classes. Note that the appendix is not part of the original publication.

predicting range are set. For example, the employed income of an individual cannot exceed the declared total income of his or her household, if available. We did not implement these possibilities in order to compare the two imputation strategies with investing similar efforts in both, so to have a fair evaluation of the respective performances. Figure 5.3 presents the result of this attempt. The right panels summarise what is obtained with IVEware by taking as imputed value, for each observation, the mean over 50 imputations. The left panels illustrates what comes out when one chooses, for each missing value, the first out of the 50 imputations as imputed value. To take the mean on 50 imputations completely destroys the original income distribution and ends up roughly imputing the median income plus or minus some variations. The imputation errors are much bigger than with our system and the results on the inequality indicators based on data partially imputed data in this manner are not acceptable (see Table 5.4). If one chooses the first imputation out of 50 as an imputed value, the empirical density of the imputed values looks better, but it is still much worse than the one provided by our system. The imputation errors are still huge and computed inequality indicators are still (all, except one) significantly different from the true values (see Table 5.4).

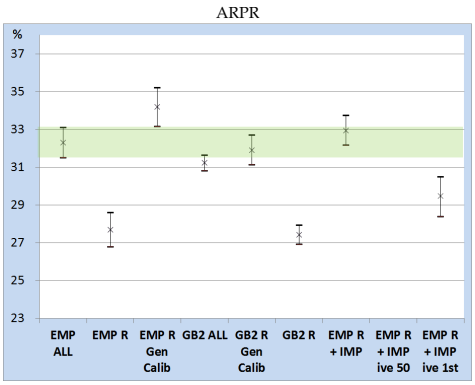
Finally, for our imputation system, Table 5.3 shows the order of magnitude of the estimated design variance and imputation variance, see (5.18). Our imputations are very stable since the coefficient of determination of the model used to predict the normal quantiles (Step 4) is over 63%. Consequently, we observe that for the considered inequality measures, the imputation variance is very low compared to the design variance (less than 6% of the total variance for all indicators) and could therefore almost be neglected.

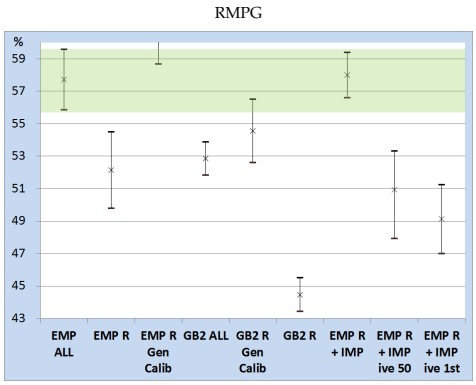
Table 5.3 – *Imputation-variance estimation from our imputation strategy. One can see that the imputation variance is very low compared to the other component, reaching at the maximum 6.6% of the total variance for the RMPG. EMP R + IMP = empirical estimation on respondents + non-respondents imputed by our imputation procedure (original survey weights).*

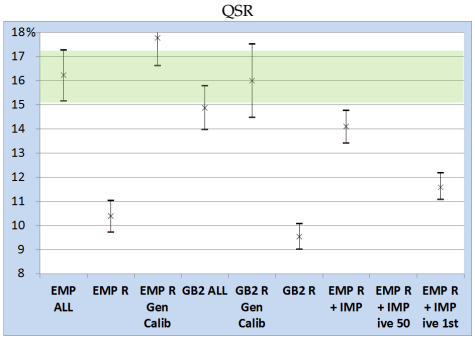
	EMP R + IMP	Imputation variance s^2_{imp}	Design variance V_p	Total variance V_{tot}	$\frac{s^2_{imp}}{V_{tot}}$	Relative bias of point estimate
ARPT	34'099	418	156162	156580	0.3%	-3%
ARPR	32.96	0.00	0.16	0.17	2.6%	2%
RMPG	58.00	0.04	0.51	0.55	6.6%	0%
QSR	14.10	0.00	0.12	0.12	0.4%	-13%
GINI	41.79	0.00	0.17	0.17	0.1%	-2%

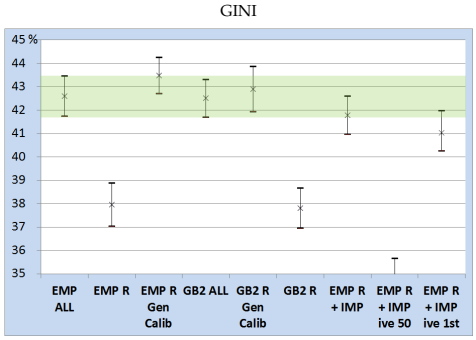
Table 5.4 – *Inequality indicators. All the estimations are weighted. 95% confidence estimated intervals. EMP ALL = empirical estimation on all the individuals (no nonresponse, original survey weights), EMP R = empirical estimation on respondents only (original survey weights, no nonresponse correction), EMP R Gen Calib = empirical estimation on respondents only using weights stemming from the generalized calibration, GB2 ALL = parametric estimation out of the GB2 fitted on all individuals (no nonresponse, original survey weights), GB2 R Gen Calib = parametric estimation out of the GB2 fitted on the respondents only using weights stemming from the generalized calibration, GB2 R = parametric estimation out of the GB2 fitted on the respondents using only the original survey weights (no nonresponse correction), EMP R + IMP = empirical estimation on respondents + non-respondents imputed by our imputation procedure (original survey weights). All variance estimations are done through linearization techniques (see Section 5.7) and take the sampling design, several nonresponse corrections, panel combinations and calibrations into account. The variance due to imputation is evaluated through multiple imputation.*

	EMP ALL			EMP R			EMP R GEN CALIB		
	Est.	ciL	ciU	Est.	ciL	ciU	Est.	ciL	ciU
ARPT	35,081	34,389	35,774	38,942	38,283	39,600	33,977	32,856	35,098
ARPR	32.31	31.51	33.12	27.68	26.78	28.59	34.20	33.17	35.22
RMPG	57.72	55.88	59.57	52.14	49.80	54.49	60.19	58.70	61.68
QSR	16.23	15.17	17.28	10.39	9.73	11.05	17.78	16.63	18.92
GINI	42.60	41.73	43.46	37.96	37.03	38.88	43.49	42.71	44.26
	GB2 ALL			GB2 R			GB2 R GEN CALIB		
	Est.	ciL	ciU	Est.	ciL	ciU	Est.	ciL	ciU
ARPT	33,210	32,618	33,802	37,737	37,114	38,359	32,246	30,902	33,589
ARPR	31.23	30.81	31.65	27.43	26.92	27.94	31.92	31.14	32.7
RMPG	52.85	51.82	53.87	44.47	43.43	45.52	54.56	52.61	56.51
QSR	14.88	13.98	15.79	9.54	9.01	10.07	16.01	14.48	17.53
GINI	42.51	41.70	43.32	37.80	36.95	38.66	42.91	41.94	43.88
	Our imputation			IVEware Imputations					
	EMP R + IMP			mean of 50 imp.			first imputation		
	Est.	ciL	ciU	IMP_iveT	ciL	ciU	IMP_ive2T	ciL	ciU
ARPT	34,099	33,323	34,874	34,131	NA	NA	37'332	NA	NA
ARPR	32.96	32.16	33.76	19.10	18.16	19.95	29.48	28.38	30.48
RMPG	58.00	56.55	59.45	50.93	47.95	53.32	49.13	46.99	51.23
QSR	14.10	13.43	14.78	7.52	7.19	7.94	11.59	11.08	12.18
GINI	41.79	40.98	42.60	34.79	34.10	35.65	41.02	40.25	41.98









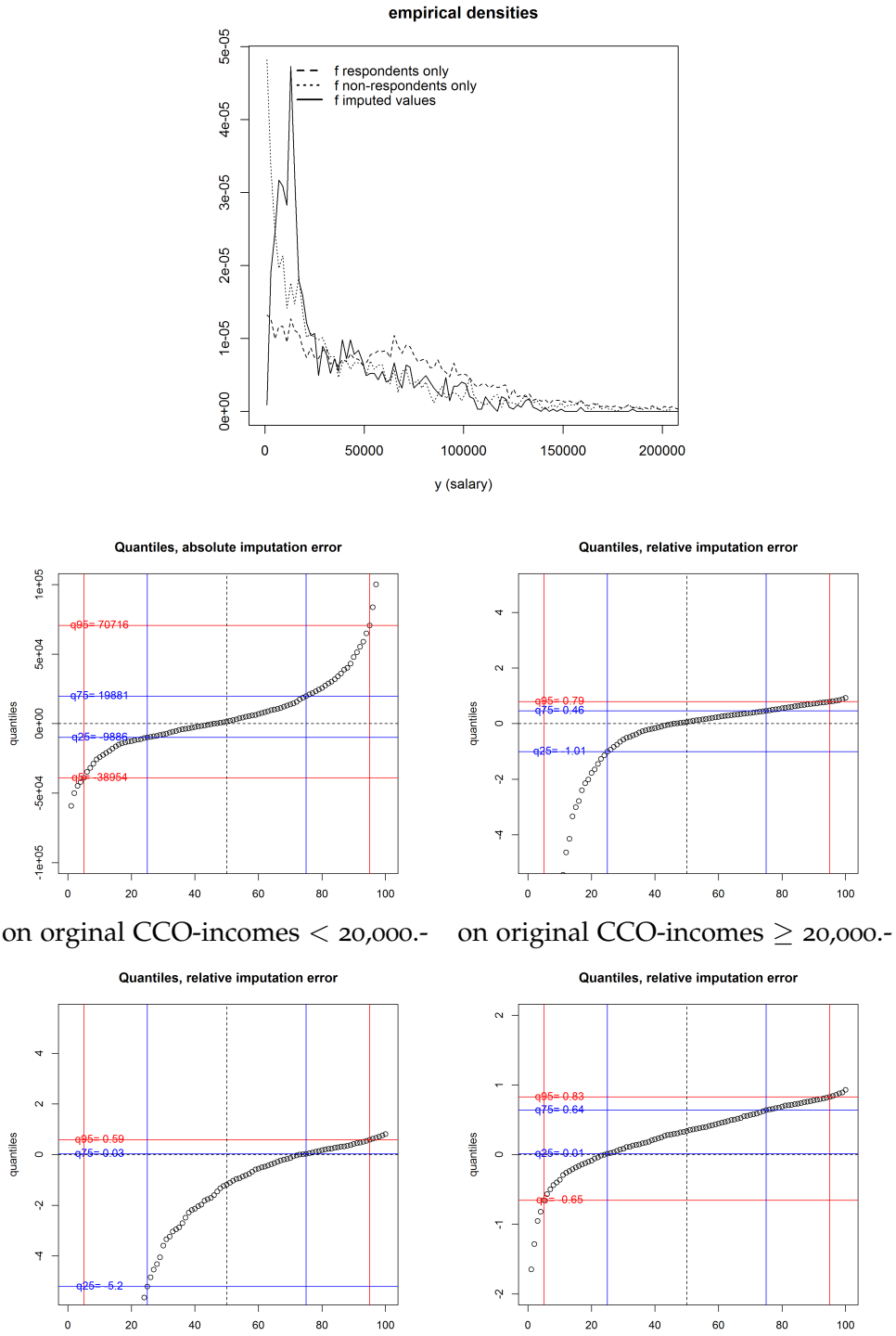


Figure 5.2 – *Results. Top panel: income empirical densities. Middle left panel: quantiles of the absolute imputation error which corresponds to the real value minus the imputed value, $y_{CCO,k} - \hat{y}_{imp,k}$, $k \in o$. Middle right and two bottom panels, quantiles of the relative imputation error, $(y_{CCO,k} - \hat{y}_{imp,k})/y_{CCO,k}$, $k \in o$; bottom left, zoom on original income below 20,000.-, bottom right, zoom on original income over 20,000.- CHF.*

IVEware imputations, mean over 50 imputed values for each missing observation.

IVEware imputations, one single imputed value for each observation.

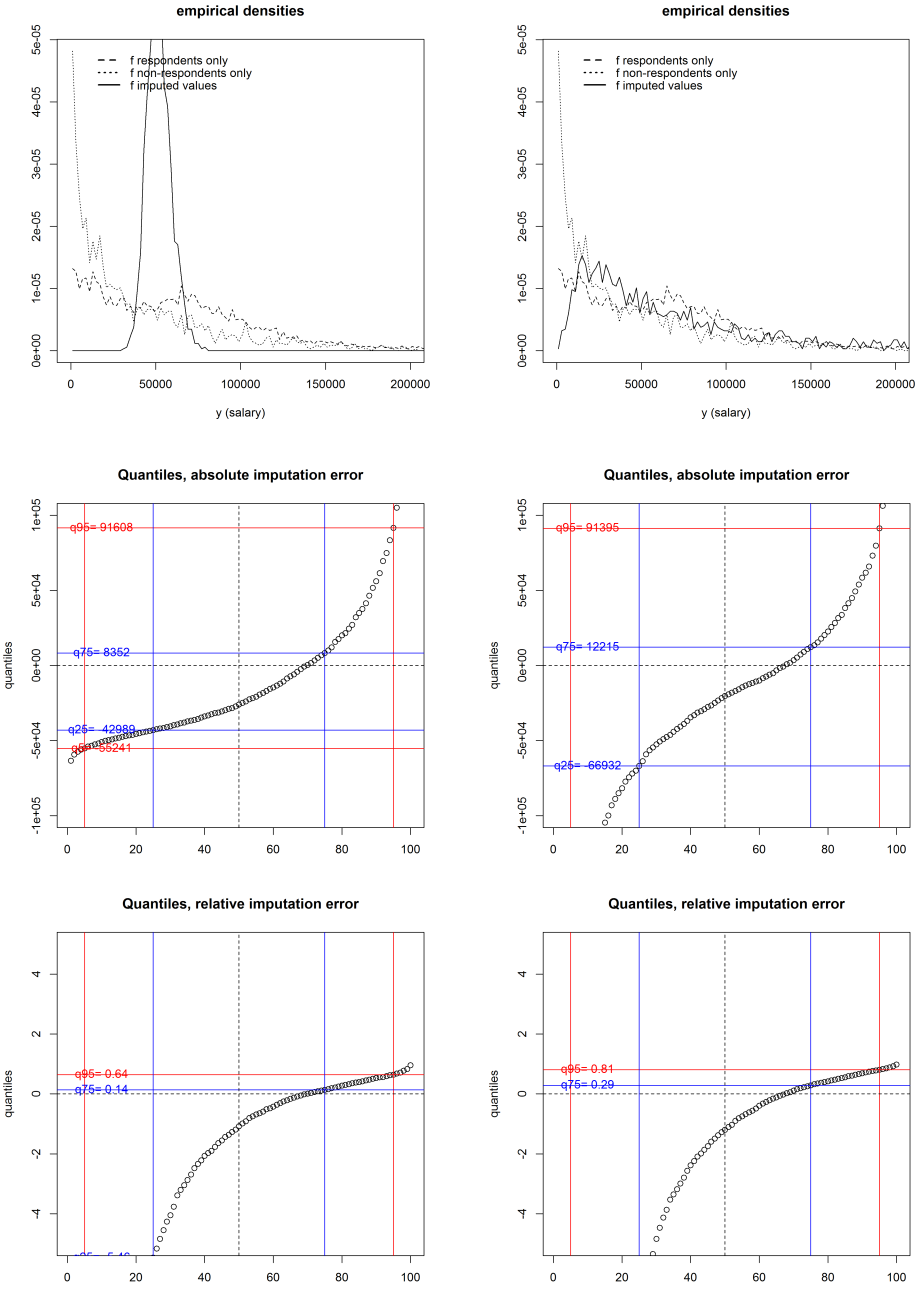


Figure 5.3 – *Results from imputations using IVEware, to be compared with Figure 5.2. On the left side, the imputed value corresponds to the mean of 50 imputations, on the right side, only one imputed value has been generated for each missing observation. Top panels: income empirical densities. Four lower panels: quantiles of the absolute, $y_{CCO,k} - \hat{y}_{imp,k}$, (middle ones) and relative, $(y_{CCO,k} - \hat{y}_{imp,k})/y_{CCO,k}$, $k \in o$ (bottom ones) imputation error, $k \in o$.*

5.9 Conclusion

The data from the Swiss Survey on Income and Living Conditions (SILC) 2009 could be matched with the register of the Central Compensation Office (CCO). Thereby some of the survey variables could be compared with their register equivalents. We have in particular applied the nonignorable nonresponse (NMAR) (*Non Missing At Random*, in the sense of Little & Rubin, 2002) mechanism affecting the survey to a register variable, the salaried annual income. We then impute the register variable by regression imputation after having transformed the variable of interest (robust normal scores). Weights stemming from a generalized calibration are of crucial importance in our imputation system. We have used adjustments of a Generalized Beta distribution of the second kind (GB2) to tune the generalized calibration. That is, once the sampled survey is fixed, the output of the generalized calibration (new weights), depends on the combination of auxiliary and instrumental variables which is chosen. GB2-adjustments relying on these new weights were used to check if the combination of auxiliary and instrumental variables is satisfactory. It is satisfactory if the GB2-density fit is approaching the shape one would get by adjusting a GB2 on the complete dataset, without nonresponse. This adherence was used as a graphical test in order to identify the best combination of auxiliary and instrumental variables defining the generalized calibration. When this is the case, it means that the weights stemming form generalized calibration are able to compensate for the NMAR nonresponse.

Our variance estimates for the inequality indicators (formerly called *Laeken indicators*, see Eurostat, 2003, 2005a) are obtained by using the generalized linearization method (Deville, 1999b; Demnati & Rao, 2004c). They take all the survey steps into account, from the sampling design, to the calibrations, going through nonresponse corrections by segmentation models, weight sharing and panel combination as has been discribed in Massiani (2013b). For the GB2-parametric estimates, we use a Sandwich variance estimator (as in the FP7 project European Union, 2011a) and linearize the weights and take the different survey steps into account. The imputation variance of our system was evaluated by performing multiple imputations (Rubin, 1987) but its magnitude was very low compared to the estimated design variance since the coefficient of determination of the model used to predict the missing value was very high (over 63%).

For the five inequality indicators mentioned, the results show no significant difference between the empirical values estimated out of the register data (no nonresponse, no imputation), and, empirical estimates based on the respondents only weighted by weights obtained by generalized calibration. At the same time, the same estimates weighted by the original survey weights are all significantly different. This means that, in this survey, the weights obtained by generalized calibration allow us to compensate for nonignorable nonresponse. These very weights allow us to almost retrieve the GB2 distribution adjustment that one would obtain by fitting a GB2 on the complete dataset (without missing values). Furthermore, the GB2 fitted on the respondent only without using the weights stem-

ming from the generalized calibration (only the original survey weights) has a very different shape and leads to biased estimates. Consequently, the parametric estimates of the inequality indicators from these GB2-fits are not significantly different from one another. In addition these estimates are, for almost all indicators, very near to the true ones, those obtained empirically on the complete dataset (see Table 5.4), only At-Risk-of-Poverty-Threshold (ARPT) shows some small differences as well as the Relative-Median-Poverty-Gap (RMPG) when the GB2 is fitted on the complete dataset.

Except for the ARPT, the empirical estimates, based on the partially imputed data resulting from our imputation system, are not significantly different from the true values. The imputation errors at the unit level are satisfactory. By comparing our imputations to several obtained by running IVEware software (Raghunathan et al., 2001) on the logarithm of the variable of interest, we conclude that we are more precise at the unit level as well as at the income distribution level. The empirical estimates obtained after the IVEware imputations were almost all biased, which means significantly different from the true values.

Thus our imputation method reveals the importance and usefulness of a set of weights capable of compensating for nonignorable nonresponse. It adapts itself and respects much better the natural distribution of income variables than a procedure needing the hypothesis that income would be (log)normally distributed. It is robust against outliers (in the sense of Chambers, 1986; Hulliger, 1999) because it is based on the ranks of the observations transformed into normal scores (by the socalled van der Waerden's method, see Conover, 1999). Our method allows, through the weights stemming from a generalized calibration and/or a GB2 fit based on the latter ones, to deliver empirical or parametric estimates of inequality measures without or before imputing. Its precision on the unit level depends on the explanatory power of the auxiliary variables at our disposal and from the capability of the weights obtained through generalized calibration to compensate for the nonresponse mechanism.

The choice of instruments in the generalized calibration is crucial. Defining a systematic method identifying valuable instruments among the variables at our disposal and allowing us to test which variable needs to be instrumented would be of great help. For the time being, we have used a graphical evaluation of GB2-adjustments. We are working on defining tests resembling the Hausman-Durbin-Wu tests (Durbin, 1954; Wu, 1973; Hausman, 1978) but in a survey context. These tests should allow one to find which auxiliary variables should be instrumented if one has valid instruments at one's disposal, i.e. instruments satisfying conditions (i) and (ii) (see Section 5.4.1).

5.10 Acknowledgement

The author wishes to warmly thank Prof. Yves Tillé for his support and advice during the research. The present work is of direct relevance to National Statistical Institutes and is endorsed by a very productive collaboration with the Swiss Federal Statistical Office (SFSO).

Discussion

6

In this section we expose some additional material which could not be included in the article because of length restrictions or because the results were judged to be not mature enough to be published. In Section 6.1 we first give some interesting complements on the Generalized Beta Distribution of the second kind. Section 6.2 presents an exploitative cogitation on the famous Durbin-Wu-Hausman test. It raises more questions than brings answers, but we consider it within the problem of tuning instrumental calibration. Section F deals with forming imputation classes for variables of income nature.

6.1 Notes on the GB2

This section aims to give the reader more insights on the interesting GB2 distribution. It is inspired by the works of McDonald (1984); Kleiber & Kotz (2003); European Union (2011a); Graf & Nedyalkova (2011, 2013).

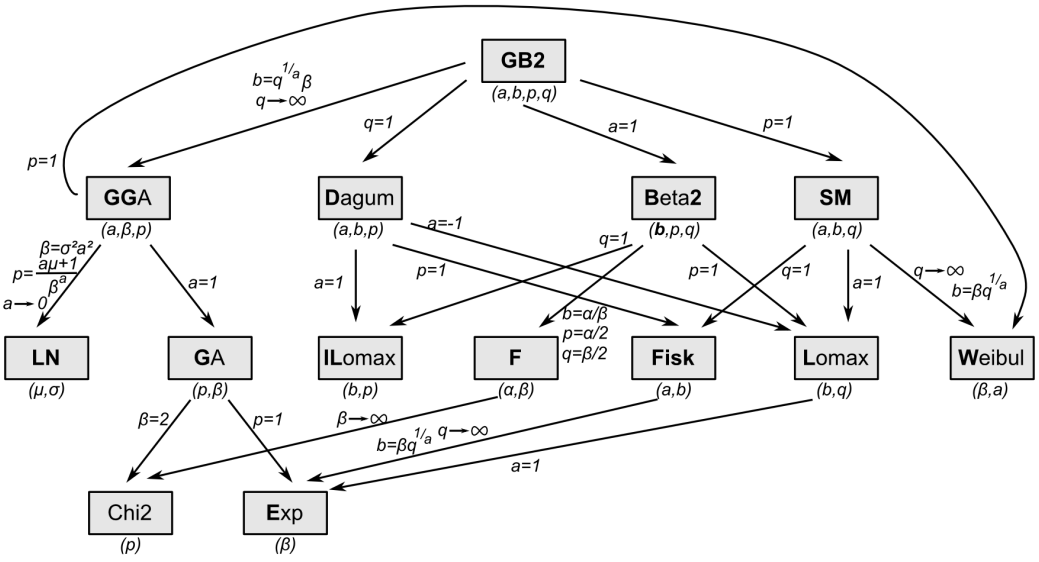


Figure 6.1 – *Links between the GB2 and other well-known distributions.*

As mentioned in Section 5.3, many authors agree on the fact that the GB2(a, b, p, q) distribution is very flexible and is well adapted to be fitted on income data. Figure 6.1 illustrates this flexibility. One can see that, with

various reparametrizations or by anchoring some parameters, the GB2 can be seen as a generalized version of many other well-known distributions often used to model income. In particular, by setting $p = q = 1$ one obtains the Fisk or log-logistic distribution, the Dagum with $q = 1$, the Singh-Maddala with $p = 1$, the Beta2 with $a = 1$, the Lomax with $p = a = 1$ and the Lomax inverse with $q = a = 1$. By letting q tend to infinity, and by applying the variable substitutions shown in Figure 6.1, one can derive the Generalized Gamma and Lognormal. If in addition $a = 1$, one obtains the Gamma, with $p = 1$ the Weibull and with $a = p = 1$ the Exponential.

6.1.1 GB2 distribution function

The density funtion of a random variable following a GB2 is already given in Section 5.3. Now, if $B(p,q) = \int_0^1 t^{p-1}(1-t)^{q-1}dt$ is the beta function (defined only for $p > 0$ and $q > 0$), let $I_z(p,q)$ be the incomplete beta ratio function, given by:

$$I_z(p,q) = \frac{1}{B(p,q)} \int_0^z t^{p-1}(1-t)^{q-1}dt, 0 \leq z \leq 1.$$

Then the distribution function of a random variable following a GB2 can be expressed as follows:

$$F_{GB2}(y;a,b,p,q) = \mathbf{P}[Y < y] = I_z(p,q) \tag{6.1}$$

where $z = \frac{(y/b)^a}{1+(y/b)^a}$, or, more precisely, $z = u/(1+u)$ with $u = (y/b)^a$.

Since the incomplete Beta function, $B_z(p,q) = \int_0^z t^{p-1}(1-t)^{q-1}dt$, is connected with the hypergeometric function[1] $B_z(p,q) = \frac{z^p}{p} {}_2F_1[p, 1-q; p+1; z]$, we can also express (6.1) as follows

$$F_{GB2}(y;a,b,p,q) = \frac{z^p}{pB(p,q)} {}_2F_1 \left[\begin{array}{ll} p, & 1-q \\ p+1 & \end{array} ;z \right].$$

6.1.2 Moments of the GB2 distribution

Let z_α be the α-quantile of the $Beta(p,q)$ distribution, and let $u_\alpha = z_\alpha/(1-z_\alpha)$, then the α-quantile of the $GB2(a,b,p,q)$ is given by:

$$y_\alpha = b\, u_\alpha^{1/\alpha}. \tag{6.2}$$

If Z is a random variable following a standard $Beta(p,q)$ distribution, and writing $U = Z/(1-Z)$, then $Y = b\, U^{1/\alpha}$ follows a $GB2(a,b,p,q)$.

Let Y be a random variable following a GB2. Then, reminding that $B(p,q) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}$, the k-th order moment is defined by:

$$E(Y^k) = b^k \frac{\Gamma(p+k/a)\Gamma(q-k/a)}{\Gamma(p)\Gamma(q)} = b^k \frac{B(p+k/a, q-k/a)}{B(p,q)} \tag{6.3}$$

[1] The hypergeometric function is defined by:

$${}_pF_q \left[\begin{array}{llll} a_1, & a_2, & \ldots & a_p \\ b_1, & b_2, & \ldots & b_q \end{array} ;x \right] = \sum_{n=0}^{\infty} \frac{(a_1)_n \ldots (a_p)_n}{(b_1)_n \ldots (b_q)_n} \frac{x^n}{n!},$$

with $(a)_n = a(a+1)(a+2)\ldots(a+n-1)$, $(a)_0 = 1$.

where the moments exist for $-ap < k < aq$, and $\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}dt = 2\int_0^\infty e^{-t^2}t^{2z-1}dt$ is the gamma function, which is defined only for $z > 0$. One notes in particular that the expectation, if it exists, depends linearly on b. On the other hand, the variance of any distribution can be expressed as $E(Y^2) - E(Y)^2$. If it exists, by (6.3), it tends to 0 for $a \to \infty$. In this case, as it has been mentioned in Section 5.3, the expectation tends towards b. Therefore, for large values of a, the mass of probability of GB2-density is concentrated around b.

The incomplete moment of order k is given by

$$\frac{E(Y^k|Y < y)}{E(Y^k)} = F_{GB2(k)}(y; a, b, p, q) = F_{GB2}(y; a, b, p + \frac{k}{a}, q - \frac{k}{a}). \quad (6.4)$$

6.1.3 Indicators of poverty and social exclusion in the EU-SILC and GB2 Distribution

As mentioned in Section 5.3, an advantage of a parametric estimation of income distribution, in particular with a GB2, is that there are explicit formulae for the inequality measures as functions of the four parameters of the law fitted on the data (McDonald, 1984; European Union, 2011a; Graf & Nedyalkova, 2011, 2013). The paramteric expressions of the five inequality measures defined in Eurostat (2003) and used in our article are now presented. The empirical definitions of these indicators are given in Sections 3.2.1 to 3.2.7.

The at-risk-of-poverty threshold (ARPT)

Let q_{50} be the median of the $GB2(a, b, p, q)$, $F_{GB2}(q_{50}) = 0.5$, computed as detailed in (6.2). Then, the ARPT is given by:

$$ARPT(a, b, p, q) = 0.6\, q_{50}$$

The at-risk-of-poverty rate (ARPR)

The risk of poverty is the proportion of the population below the poverty line. As it is scale-independent, the parameter b can be arbitrarily chosen, e.g. set $= 1$. Graf & Nedyalkova (2013) show that for large values of a (approx. ≥ 9) the ARPR becomes insensitive to the values taken by q. We have often been in this situation with the SILC 2009 data used. If Y is a random variable following a $GB2(a, b, p, q)$:

$$\begin{aligned} ARPR(a, p, q) &= \mathbf{P}(Y < 0.6\, q_{50}) = \mathbf{P}(Y < ARPT) \\ &= F_{GB2}(0.6\, q_{50}) = F_{GB2}(ARPT; a, 1, p, q), \end{aligned}$$

where F_{GB2} is the distribution function, see (6.1). Note that $ARPT = q_{GB2}(ARPR; a, 1, p, q)$ where q_{GB2} is the quantile function of the considered GB2.

Relative median at-risk-of-poverty gap (RMPG)

It is the relative difference between the poverty line and the median of the poor (= those below the threshold): let $m_p = F_{GB2}^{-1}(ARPR/2) = q_{GB2}(ARPR/2)$ be the median of the poor,

$$RMPG(A,a,p,q) = \frac{0.6q_{50} - m_p}{0.6q_{50}} = \frac{ARPT - m_p}{ARPT},$$

where $A = ARPR(a,p,q)$ is the at-risk-of-poverty rate. The RMPG is defined as one minus the ratio between the median income of people having an income below the ARPT and 60% of the median income of the population.

$$RMPG(A,a,p,q) = 1 - \frac{q_{GB2}(A/2,a,1,p,q)}{q_{GB2}(A,a,1,p,q)} = 1 - \frac{q_{GB2}(A/2,a,1,p,q)}{ARPT}$$

where $q_{GB2} = F_{GB2}^{-1}$ is the quantile function of considered GB2.

Quintile share ratio (QSR or S_{80}/S_{20})

Let q_{80} et q_{20} be the 80$^{\text{th}}$ and 20$^{\text{th}}$ percentiles of the GB2 distribution function. The quintile share ratio is the ratio of the cumulated incomes of the 20% richest, over the cumulated incomes of the 20% poorest:

$$QSR = \frac{\mathbf{E}(Y|Y > q_{80})}{\mathbf{E}(Y|Y < q_{20})}.$$

It can be expressed using the incomplete moments of order one, see (6.4) with $k = 1$:

$$QSR(a,p,q) = \frac{1 - F_{GB2(1)}(q_{80};a,1,p,q)}{F_{GB2(1)}(q_{20};a,1,p,q)} = \frac{1 - F_{GB2}(q_{80};a,1,p+\frac{1}{a},q-\frac{1}{a})}{F_{GB2}(q_{20};a,1,p+\frac{1}{a},q-\frac{1}{a})}.$$

Gini Index

There are several possible definitions. If X and Y are two random variables with distribution F,

$$GINI(F) = \frac{\mathbf{E}(|X-Y|)}{2\mathbf{E}(X)}.$$

It is an inequality index measuring the expectation of the absolute difference of two independently selected income values relative to the median income. The Gini index of a GB2 distribution is given in McDonald (1984):

$$GINI(a,p,q) = \frac{B(2p+1/a, 2q-1/a)}{B(p,q)B(p+1/a,q-1/a)} \left\{ \frac{1}{p} G_1 - \frac{1}{p+1/a} G_2 \right\}$$

where

$$G_1 = {}_3F_2 \left[\begin{array}{ccc} 1, & p+q, & 2p+1/a \\ & p+1, & 2(p+q) \end{array} ; 1 \right]$$

and

$$G_2 = {}_3F_2 \left[\begin{array}{ccc} 1, & p+q, & 2p+1/a \\ & p+1+1/a, & 2(p+q) \end{array} ; 1 \right],$$

where ${}_3F_2$ is the hypergeometric function already defined in Section 6.1.1. The Gini index is defined only if the expectation exists, i.e. when $q - 1/a > 0$, see (6.3)). Direct application of the above formula for the Gini may lead to convergence problems. An efficient algorithm that may be used to calculate the Gini index is given in Graf (2009c) and implemented in the `R`-package GB2 (function `gb2.gini`).

In some cases, this index takes a simpler form (see European Union, 2011a; Kleiber & Kotz, 2003)

- Distribution Beta-2 ($a = 1$): $GINI(p, q) = \frac{B(2p,2q-1)}{2pB^2(p,q)}$,
- Distribution Dagum ($q = 1$): $GINI(a, p) = \frac{\Gamma(p)\Gamma(2p+1/a)}{\Gamma(2p)\Gamma(p+1/a)} - 1$,
- Distribution Singh-Maddalah ($p = 1$): $GINI(a, q) = 1 - \frac{\Gamma(q)\Gamma(2q-1/a)}{\Gamma(2q)\Gamma(q-1/a)}$.
- Distribution Lognormal: $GINI(\sigma) = 2\Phi(\sigma/\sqrt{2}) - 1$, Φ being the cumulative distribution function of the standard normal distribution.

6.1.4 Note on the variance-covariance matrix of totals of the GB2 estimated scores

This is a short add-on to complete the explanation of formula (5.17) given in Section 5.7. As said, $\hat{V}_p(\hat{\mathbf{U}})$ is the estimated design variance of the totals of the scores evaluated at $\hat{\boldsymbol{\theta}}$, it can be developed as follows:

$$\begin{aligned}
\hat{V}_p(\hat{\mathbf{U}}) &= \widehat{\text{var}}[\sum_{k\in r} w_k \mathbf{u}_k(y_k;\hat{\boldsymbol{\theta}})] \\
&= \widehat{\text{var}}\begin{pmatrix} \sum_{i=1}^n w_i u_{ia}(y_i;\hat{\theta}_n) \\ \sum_{i=1}^n w_i u_{ib}(y_i;\hat{\theta}_n) \\ \sum_{i=1}^n w_i u_{ip}(y_i;\hat{\theta}_n) \\ \sum_{i=1}^n w_i u_{iq}(y_i;\hat{\theta}_n) \end{pmatrix} := \widehat{\text{var}}\begin{pmatrix} \hat{U}_a \\ \hat{U}_b \\ \hat{U}_p \\ \hat{U}_q \end{pmatrix} \\
&= \begin{pmatrix} \hat{V}_p(\hat{U}_a) & \widehat{\text{cov}}_p(\hat{U}_a,\hat{U}_b) & \widehat{\text{cov}}_p(\hat{U}_a,\hat{U}_p) & \widehat{\text{cov}}_p(\hat{U}_a,\hat{U}_q) \\ \widehat{\text{cov}}_p(\hat{U}_b,\hat{U}_a) & \hat{V}_p(\hat{U}_b) & \widehat{\text{cov}}_p(\hat{U}_b,\hat{U}_p) & \widehat{\text{cov}}_p(\hat{U}_b,\hat{U}_q) \\ \widehat{\text{cov}}_p(\hat{U}_p,\hat{U}_a) & \widehat{\text{cov}}_p(\hat{U}_p,\hat{U}_b) & \hat{V}_p(\hat{U}_p) & \widehat{\text{cov}}_p(\hat{U}_p,\hat{U}_q) \\ \widehat{\text{cov}}_p(\hat{U}_q,\hat{U}_a) & \widehat{\text{cov}}_p(\hat{U}_q,\hat{U}_b) & \widehat{\text{cov}}_p(\hat{U}_q,\hat{U}_p) & \hat{V}_p(\hat{U}_q) \end{pmatrix}
\end{aligned}$$

Note that we can obtain the covariances by simply calculating variances, since for two random variables X and Y, we can write

$$\text{var}(X - Y) = \text{var}(X) + \text{var}(Y) - 2\text{cov}(X, Y),$$

thus

$$\text{cov}(X, Y) = \frac{\text{var}(X) + \text{var}(Y) - \text{var}(X - Y)}{2}.$$

It follows that to obtain the estimated variance-covariance matrix of GB2-scores evalutated at $\hat{\boldsymbol{\theta}}$, $\hat{V}_p(\hat{\mathbf{U}})$, it is enough to compute the diagonal terms $\hat{V}_p(\hat{U}_a)$, $\hat{V}_p(\hat{U}_b)$, $\hat{V}_p(\hat{U}_p)$, $\hat{V}_p(\hat{U}_q)$ as well as $\hat{V}_p(\hat{U}_a - \hat{U}_b)$, $\hat{V}_p(\hat{U}_a - \hat{U}_p)$, $\hat{V}_p(\hat{U}_a - \hat{U}_q)$, $\hat{V}_p(\hat{U}_b - \hat{U}_p)$, $\hat{V}_p(\hat{U}_b - \hat{U}_q)$, $\hat{V}_p(\hat{U}_p - \hat{U}_q)$. All these variances of totals are estimated via linearization as described in Massiani (2013a) in order to take the variability of the survey weights into account.

6.2 Some remarks about the Durbin-Wu-Hausman test

This section aims to share some research we have done on the Durbin-Wu-Hausman tests. Some interesting questions and first conclusions could be formulated, which could lead, later on, to concrete results. This section has to be read as a complement to Sections 5.4.1 and 5.4.2. It details also what has been written in the conclusions of Section 5.9. We take over the notations from Sections 5.4.1 and 5.4.2 and do not re-introduce them here.

It was found that, in practice, it is decisive for the success of generalized calibration to 1) to have good exogenous instruments $\mathbf{Z}_1$ and 2) to properly sort among the $\mathbf{X} = [\mathbf{X}_1|\mathbf{X}_2]$ the endogenous auxiliary variables, $\mathbf{X}_1$, and the exogenous ones, $\mathbf{X}_2$, which can be instruments for themselves. We must be able to classify the columns of matrix $\mathbf{X}$ in $\mathbf{X}_1$ and $\mathbf{X}_2$ and, in addition, have instruments $\mathbf{Z}_1$ that are able to compensate for the endogeneity of $\mathbf{X}_1$, i.e. satisfying the conditions stated under Section 5.4.1.

If we can do it, generalized calibration will provide much better weights[2] than conventional calibration, otherwise the generalized calibration can be very risky (i.e. lead to worse GB2 fits than those obtained using weights stemming from conventional calibration).

We would actually like to have a way to test if the auxiliary and instrumental variables are endogenous or not. In the field of econometrics, this problem is well known and explored in the context of solving simultaneous equations, for example. The method for doing this is to use **Durbin-Wu-Hausman tests** (DHW). These tests were developed by Durbin (1954); Wu (1973); Hausman (1978).

For the time being, we will forget about sampling theory and use a conventional notation of classical statistics. Indeed, to our knowledge, the DHW tests have not been adapted to a sampling framework yet. So, let us consider the linear model

$$y = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}. \tag{6.5}$$

There has been considerable work done on the DWH tests since then (see Ruud, 1984, for a review and many more recent ones). These tests are often interpreted as testing for endogeneity or exogeneity of the columns of $\mathbf{X}$ not belonging to the subspace generated by $\mathbf{Z} = [\mathbf{Z}_1|\mathbf{X}_2]$ and it is also with this intention that we would like to apply them. However what is truly being tested is not the endogeneity or exogeneity of the columns of $\mathbf{X}$ but rather, **the effect of a possible endogeneity on the components β** estimated by regressing $\mathbf{X}$ on y.

The basic idea of DWH tests is to build a test on a vector of contrasts, i.e. the difference between two vectors of estimated parameters. One will be consistent under less restrictive conditions than the other.

It is assumed that the model to be tested is (6.5) with $\epsilon \sim iid\,\mathcal{N}(0, \sigma^2\mathbb{1})$ where there are n observations and J regressors. In this context the principle of the DHW test is to compare two estimators $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_A$.

The null hypothesis is, H_0:

[2] Best in the sense that, in our framework, they can lead to fit and find the GB2 distribution that would be obtained by fitting on the complete data.

- There is a consistent estimator $\hat{\boldsymbol{\beta}}_0$, asymptotically normal and efficient[3],
- There is another estimator $\hat{\boldsymbol{\beta}_A}$ which is consistent under H_0 and H_A but which is inefficient.

If H_0 is true, both estimators are unbiased and therefore we prefer the one that is efficient.

Under H_A:

- $\hat{\boldsymbol{\beta}}_0$ is biased and inconsistent (the 0-model is miss-specified),
- $\hat{\boldsymbol{\beta}_A}$ remains consistent,

Consequently, under H_A, the vector of contrasts $\hat{\mathbf{q}} = \hat{\boldsymbol{\beta}}_A - \hat{\boldsymbol{\beta}}_0$ is large in absolute value relative to its standard deviation. In the case that interests us, $\boldsymbol{\beta}_0$ is the Ordinary Least Squares (OLS) estimator

$$\hat{\boldsymbol{\beta}}_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'y \tag{6.6}$$

and $\boldsymbol{\beta}_A$ is another linear estimator[4], for example, the Instrumental Variable (IV) estimator:

$$\hat{\mathbf{b}}_{IV} = (\mathbf{X}'\mathbf{P}_Z\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_Z'y. \tag{6.7}$$

The $\mathbf{P}_Z$ matrix projects orthogonally onto the hyperplane L_z, where $\mathbf{Z} = [\mathbf{Z}_1|\mathbf{X}_2]$ is the matrix of instruments.

To build his test, Hausman (1978) shows that, if H_0 is true, $cov(\hat{\mathbf{q}}, \hat{\boldsymbol{\beta}}_0) = 0$, equivalently $cov(\hat{\boldsymbol{\beta}}_A, \hat{\boldsymbol{\beta}}_0) = var(\hat{\boldsymbol{\beta}}_0)$. We convince ourselves easily from this fact in the case of comparison between the OLS-estimator and the IV-estimator. Indeed, in this case, under H_0 and with $\epsilon \sim iid\,\mathcal{N}(0, \sigma^2\mathbb{1})$:

$$\begin{aligned}
cov(\hat{\mathbf{b}}_{IV}, \hat{\boldsymbol{\beta}}_{OLS}) &= cov(\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'y, \mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'y) \\
&= cov((\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\epsilon, (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\epsilon) \\
&= (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\epsilon\epsilon'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= (\mathbf{X}'\mathbf{P}_Z\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_Z'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\sigma_\epsilon^2 \\
&= \sigma_\epsilon^2(\mathbf{X}'\mathbf{X})^{-1} = var(\hat{\boldsymbol{\beta}}_{OLS}).
\end{aligned}$$

Hausman also shows that the matrix $\mathbf{C} = var(\hat{\boldsymbol{\beta}}_1) - var(\hat{\boldsymbol{\beta}}_0) \geq 0$ is defined non negative. $\mathbf{C}$ is called the *precision matrix*, its rank is the number of endogenous variables tested, i.e. the number of variables in $\mathbf{X}_2$, say k. The test statistic is defined by:

$$H = \hat{\mathbf{q}}'\mathbf{C}^-\hat{\mathbf{q}} = \|\mathbf{q}\|^2_{\mathbf{C}^-}$$

where $\mathbf{C}^-$ is the Moore-Penrose generalized inverse of $\mathbf{C}$. H follows a χ^2_k distribution if σ^2 is known. In the case where this error variance must be estimated, $H/k \sim F_{k,n-J}$. A too large value of H will result in rejecting H_0.

[3]Efficient: the estimator reaches the Cramer-Rao bound on the sample of finite size.

[4]We can take any other estimator $\hat{\boldsymbol{\beta}}_A = (\mathbf{X}'\mathbf{A}\mathbf{X})^{-1}\mathbf{X}'\mathbf{A}y$ where $\mathbf{A}$ is a symmetric matrix of size $n \times n$ which is assumed, to simplify, having a rank at least equal to the number of variables in $\mathbf{X}$ (otherwise, we could not estimate all components of $\hat{\boldsymbol{\beta}}_A$, and we could compare only its estimated part to the corresponding sub-vector $\hat{\boldsymbol{\beta}}_{OLS}$). The projection matrix $\mathbf{P}_Z$, see Section 5.4.1 does the job.

Intuitively we will remind ourselves that if $\mathbf{X} \sim \mathcal{N}(\mu, \mathbf{\Sigma}))$, then the square of the $\mathbf{\Sigma}^{-1}$-norm of the vector of contrasts of variables in $\mathbf{X}$ relatively to their expectation μ, $\|\mathbf{X} - \mu\|^2_{\mathbf{\Sigma}^{-1}}$, follows a chi squared distribution: $(\mathbf{X} - \mu)'\mathbf{\Sigma}^{-1}(\mathbf{X} - \mu) \sim \chi^2_r$, where $r = rang(\mathbf{\Sigma})$, moreover a Fisher distribution is obtained by the ratio of two chi-square distributions divided by their respective degrees of freedom.

When comparing the OLS regression to IV, the test is written:

$$H = \frac{1}{k}(\hat{\mathbf{b}}_{IV} - \hat{\boldsymbol{\beta}}_{OLS})'[\mathrm{Var}(\hat{\mathbf{b}}_{IV}) - \mathrm{Var}(\hat{\boldsymbol{\beta}}_{OLS})]^{-}(\hat{\mathbf{b}}_{IV} - \hat{\boldsymbol{\beta}}_{OLS}) \sim F_{k,n-J}. \quad (6.8)$$

Hausman also shows that the above-described test is equivalent to test $H_0 : \boldsymbol{\alpha} = 0$ or $H_0 : \boldsymbol{\delta} = 0$ in one of the following two regressions:

$$\begin{aligned} y &= [\mathbf{X}_1|\mathbf{X}_2]\boldsymbol{\beta} + \mathbf{P}_Z\mathbf{X}_1\boldsymbol{\alpha} + r_1 \quad (6.9)\\ &= \mathbf{X}\boldsymbol{\beta} + \hat{\mathbf{X}}_1\boldsymbol{\alpha} + r_1 \end{aligned}$$

or

$$\begin{aligned} y &= [\mathbf{X}_1|\mathbf{X}_2]\breve{\boldsymbol{\beta}} + \mathbf{Q}_Z\mathbf{X}_1\boldsymbol{\delta} + r_2 \quad (6.10)\\ &= \mathbf{X}\breve{\boldsymbol{\beta}} + \mathbf{v}\boldsymbol{\delta} + r_2 \end{aligned}$$

where $\mathbf{Q}_Z = \mathbb{1} - \mathbf{P}_Z$, i.e. that $\mathbf{v}$ are the residuals of the projection of the endogenous variables $\mathbf{X}_1$ onto the hyperplane L_z spanned by the instrumental variables $\mathbf{Z} = [\mathbf{Z}_1|\mathbf{X}_2]$: $\mathbf{X}_1 = \hat{\mathbf{X}}_1 + \mathbf{v} = \mathbf{P}_Z\mathbf{X}_1 + \mathbf{Q}_Z\mathbf{X}_1$.

One then performs a Student-test t on α or δ if one single variable is tested, or a Fisher-test on $\boldsymbol{\alpha}$ or $\boldsymbol{\delta}$ if several potentially endogenous variables are tested at once.

With the theorem of Frisch-Waugh-Lovell (FWL) (see for example Davidson & MacKinnon, 1993), it can be shown that $\|r_1\|^2 = \|r_2\|^2$, thus tests on $\boldsymbol{\alpha}$ ou $\boldsymbol{\delta}$ will be numerically identical. These tests may be easier to conduct in some cases. However, in our study we used the Hausman test (6.8) on the estimated parameters from OLS and IV regressions and did not do testing on the significance of the residuals $\mathbf{v}$, see (6.11), or of $\hat{\mathbf{X}}_1$, see (6.10), obtained from the projection of the endogenous variables onto the hyperplane L_z.

To compute (6.8) we use the following approximation:

$$H = \frac{1}{k}(\hat{\mathbf{b}}_{IV} - \hat{\boldsymbol{\beta}}_{OLS})'[\hat{V}(\hat{\mathbf{b}}_{IV}) - \hat{V}(\hat{\boldsymbol{\beta}}_{OLS})]^{-}(\hat{\mathbf{b}}_{IV} - \hat{\boldsymbol{\beta}}_{OLS}), \quad (6.11)$$

where $\hat{V}(\hat{\mathbf{b}}_{IV})$ et $\hat{V}(\hat{\boldsymbol{\beta}}_{OLS})$ are consistent estimates of $\mathrm{Var}(\hat{\mathbf{b}}_{IV})$ and respectively $\mathrm{Var}(\hat{\boldsymbol{\beta}}_{OLS})$.
$\hat{\mathbf{b}}_{IV}$ is given by (6.7) and $\hat{\boldsymbol{\beta}}_{OLS}$ by (6.6). Moreover,

$$\begin{aligned} \mathrm{Var}(\hat{\mathbf{b}}_{IV}) &= \sigma^2_{IV}(\mathbf{X}'\mathbf{P}_Z\mathbf{X})^{-1}, \quad (6.12)\\ \mathrm{Var}(\hat{\boldsymbol{\beta}}_{OLS}) &= \sigma^2_{OLS}(\mathbf{X}'\mathbf{X})^{-1}, \quad (6.13) \end{aligned}$$

may be estimated by:

$$\begin{aligned} \hat{\sigma}^2_{IV} &= \frac{1}{n-J}\hat{\epsilon}'_{IV}\hat{\epsilon}_{IV},\\ \hat{\sigma}^2_{OLS} &= \frac{1}{n-J}\hat{\epsilon}'_{OLS}\hat{\epsilon}_{OLS}, \end{aligned}$$

where $\hat{\epsilon}_{IV} = y - \mathbf{X}\hat{\mathbf{b}}_{IV}$ et $\hat{\epsilon}_{OLS} = y - \mathbf{X}\hat{\boldsymbol{\beta}}_{OLS}$.

However, under H_0, $\hat{\sigma}^2_{IV}$ and $\hat{\sigma}^2_{OLS}$ are both consistent estimators of σ^2. It is advisable in practice to choose one of the two to ensure a positive test statistic. We can write

$$\mathbf{C}^- = \sigma^2[(\mathbf{X}\mathbf{P}_Z\mathbf{X})^{-1} - (\mathbf{X}\mathbf{X})^{-1}].$$

and use σ^2_{OLS} as an estimator of σ^2.

6.2.1 DWH tests with weighted observations

Hausman tests previously reported are not weighted in nearly all of the versions presented in the literature. However, we need a version of the test that can take into account the survey weights w_k. That is to reformulate the model (6.5) as follows (still considering ourselves in the framework of classical statistics, not in sampling theory):

$$y = \mathbf{X}\boldsymbol{\beta} + \epsilon,\ \epsilon \sim \mathcal{N}(0, \sigma^2\boldsymbol{\Omega}),$$

i.e. with heteroscedastic errors: $var(\epsilon) = \sigma^2\boldsymbol{\Omega} = \sigma^2 diag(\frac{1}{w_k})$.

$$\boldsymbol{\Omega} = \begin{pmatrix} 1/w_1 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & 1/w_n \end{pmatrix}$$

We find ourselves in a special case of generalized least squares and the Weighted Least Squares (WLS) estimator $\hat{\boldsymbol{\beta}}_{WLS}$ is obtained by minimizing the $\boldsymbol{\Omega}^{-1}$-norm between y and $\mathbf{X}\boldsymbol{\beta}$:

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{WLS} &= \underset{\boldsymbol{\beta}\in\mathbb{R}^J}{\operatorname{argmin}} \|y - \mathbf{X}\boldsymbol{\beta}\|^2_{\boldsymbol{\Omega}^{-1}} \\ &= \underset{\boldsymbol{\beta}\in\mathbb{R}^J}{\operatorname{argmin}} \{(y - \mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Omega}^{-1}(y - \mathbf{X}\boldsymbol{\beta})\} \\ &= \underset{\boldsymbol{\beta}\in\mathbb{R}^J}{\operatorname{argmin}} \{\sum_{k=1}^{n} w_k(y_k - \sum_{j=1}^{J}\beta_j x_{ij})\}. \end{aligned}$$

The solution is given by:

$$\hat{\boldsymbol{\beta}}_{WLS} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}y$$

with $\mathbf{W} = \boldsymbol{\Omega}^{-1} = diag(w_k)$ the diagonal matrix of weights. Its variance is given by

$$var(\hat{\boldsymbol{\beta}}_{WLS}) = \sigma^2_{WLS}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1},$$

while consistent estimation of σ^2_{WLS} is given by

$$\hat{\sigma}^2_{WLS} = \frac{1}{n-J}\|y - \mathbf{X}\hat{\boldsymbol{\beta}}_{WLS}\|^2_{\boldsymbol{\Omega}^{-1}} = \frac{1}{n-J}\hat{\epsilon}'_{WLS}\mathbf{W}\hat{\epsilon}_{WLS}. \tag{6.14}$$

As mentioned in Section 5.4.1, the Weighted Instrumental Variable (WIV) estimator is given by

$$\begin{aligned}\hat{\mathbf{b}}_{WIV} &= \underset{\beta\in\mathbb{R}^J}{\operatorname{argmin}} \|y-\mathbf{X}\beta\|^2_{\mathbf{P}_Z\mathbf{\Omega}^{-1}} \\ &= \underset{\beta\in\mathbb{R}^J}{\operatorname{argmin}} \left\{(y-\mathbf{X}\beta)'\mathbf{P}_Z\mathbf{\Omega}^{-1}(y-\mathbf{X}\beta)\right\}\end{aligned} \tag{6.15}$$

The solution is given by:

$$\hat{\mathbf{b}}_{WIV} = (\mathbf{Z}'\mathbf{W}\mathbf{X})^{-1}\mathbf{Z}'\mathbf{W}y.$$

Now to calculate the variance-covariance matrix of $\hat{\mathbf{b}}_{WIV}$, by substituting $y = \mathbf{X}\beta + \epsilon$ in (6.16), we find:

$$\begin{aligned}\hat{\mathbf{b}}_{WIV} &= (\mathbf{Z}'\mathbf{W}\mathbf{X})^{-1}\mathbf{Z}'\mathbf{W}(\mathbf{X}\beta+\epsilon) \\ &= \beta + (\mathbf{Z}'\mathbf{W}\mathbf{X})^{-1}\mathbf{Z}'\mathbf{W}\epsilon,\end{aligned}$$

thus,

$$\hat{\mathbf{b}}_{WIV} - \beta = (\mathbf{Z}'\mathbf{W}\mathbf{X})^{-1}\mathbf{Z}'\mathbf{W}\epsilon.$$

Next, the variance-covariance matrix is given by:

$$\begin{aligned}var(\hat{\mathbf{b}}_{WIV}) &= E[(\hat{\mathbf{b}}_{WIV}-\beta)(\hat{\mathbf{b}}_{WIV}-\beta)'] \\ &= E[((\mathbf{Z}'\mathbf{W}\mathbf{X})^{-1}\mathbf{Z}'\mathbf{W}\epsilon)((\mathbf{Z}'\mathbf{W}\mathbf{X})^{-1}\mathbf{Z}'\mathbf{W}\epsilon)'] \\ &= E[(\mathbf{Z}'\mathbf{W}\mathbf{X})^{-1}\mathbf{Z}'\mathbf{W}\epsilon\epsilon'\mathbf{W}\mathbf{Z}(\mathbf{X}'\mathbf{W}\mathbf{Z})^{-1}],\end{aligned}$$

Considering the columns of $\mathbf{X}$ and $\mathbf{Z}$ as fixed (non-stochastic)[5] and remembering that $\epsilon\epsilon' = \sigma^2\mathbf{\Omega} = \sigma^2\mathbf{W}^{-1}$:

$$\begin{aligned}var(\hat{\mathbf{b}}_{WIV}) &= (\mathbf{Z}'\mathbf{W}\mathbf{X})^{-1}\mathbf{Z}'\mathbf{W}E(\epsilon\epsilon')\mathbf{W}\mathbf{Z}(\mathbf{X}'\mathbf{W}\mathbf{Z})^{-1} \\ &= (\mathbf{Z}'\mathbf{W}\mathbf{X})^{-1}\mathbf{Z}'\mathbf{W}\sigma^2\mathbf{W}^{-1}\mathbf{W}\mathbf{Z}(\mathbf{X}'\mathbf{W}\mathbf{Z})^{-1} \\ &= \sigma^2(\mathbf{Z}'\mathbf{W}\mathbf{X})^{-1}\mathbf{Z}'\mathbf{W}\mathbf{Z}(\mathbf{X}'\mathbf{W}\mathbf{Z})^{-1}\end{aligned} \tag{6.16}$$

It is also possible to express everything with respect to $\hat{\mathbf{X}} = \mathbf{P}_z\mathbf{X}$:

$$var(\hat{\mathbf{b}}_{WIV}) = \sigma^2(\hat{\mathbf{X}}'\mathbf{W}\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\mathbf{W}^{-1}\hat{\mathbf{X}}(\hat{\mathbf{X}}'\mathbf{W}\hat{\mathbf{X}})^{-1}.$$

With both formulations, we fall correctly back to (6.12) if $\mathbf{W} = \mathbb{1}$. As for the unweighted case, $\hat{\sigma}^2_{WLS}$ and $\hat{\sigma}^2_{WIV}$ are both consistent estimators of σ^2 under H_0. We will thus save the trouble of deriving a consistent estimator of $\hat{\sigma}^2_{WIV}$ at this stage and apply the following statistical test using $\hat{\sigma}^2_{WLS}$:

$$\begin{aligned}H_w &= \frac{1}{k}(\hat{\mathbf{b}}_{WIV}-\hat{\beta}_{WLS})'[\hat{V}(\hat{\mathbf{b}}_{WIV}) - \hat{V}(\hat{\beta}_{WLS})]^-(\hat{\mathbf{b}}_{WIV}-\hat{\beta}_{WLS}) \\ &= \frac{1}{k}(\hat{\mathbf{b}}_{WIV}-\hat{\beta}_{WLS})'[\hat{\sigma}^2_{WLS}((\mathbf{Z}'\mathbf{W}\mathbf{X})^{-1}\mathbf{Z}'\mathbf{W}\mathbf{Z}(\mathbf{X}'\mathbf{W}\mathbf{Z})^{-1}) - (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1})]^- \\ &\qquad (\hat{\mathbf{b}}_{WIV}-\hat{\beta}_{WLS})\end{aligned} \tag{6.17}$$

[5] Which may be questionable in a sampling theory framework. Especially since the columns of $\mathbf{Z}$ are only known on the respondent's sample when considering the process of nonresponse as a stochastic process.

with $\hat{\sigma}^2_{WLS}$ given by (6.14).

If we want to compare two WIV-regressions, the one with fewer variables, the following statistical test will be used:

$$H^{12}_w = \frac{1}{k}(\hat{\mathbf{b}}^2_{WIV} - \hat{\mathbf{b}}^1_{WIV})'[\hat{V}(\hat{\mathbf{b}}^2_{WIV}) - \hat{V}(\hat{\mathbf{b}}^1_{WIV})]^-(\hat{\mathbf{b}}^2_{WIV} - \hat{\mathbf{b}}^1_{WIV}),$$

where $\hat{V}(\hat{\mathbf{b}}^i_{WIV})$, $i = 1, 2$ is given by (6.16) with $\hat{\sigma}^{2,i}_{WIV} = \frac{1}{n-J}\hat{\epsilon^i}'_{WIV}\hat{\epsilon^i}_{WIV}$ and $\hat{\epsilon^i}_{WIV} = y - \mathbf{X}\hat{\mathbf{b}}^i_{WIV}$.

6.2.2 Lessons learned so far with DWH tests

We have implemented the DWH test (6.17), as well as equivalent tests as (6.10) and (6.11), and tested them on the SILC 2009 data. The goal was to see if the optimal combination of auxiliary and instrumental variables found by fitting GB2 distributions as explained in Section 5.4.2 could be retrieved. We did not fully reach that goal because of unsolved problems, which are described below.

Estimation of σ^2

We deal with income data which contain extreme values by nature, moreover, the survey weights themselves can vary by a factor of 30 or more between the smaller and the larger. These tow factors lead to very unstable estimates of σ^2. Thus it is necessary to use robust estimations of σ^2 or to apply some treatment to the data (e.g. MAD-rule, see Section 2.5.3) and possibly to the weights before estimating σ^2. But the value of $\hat{\sigma}^2$ has a direct influence on the value of the test, so the test itself is very sensitive to the robustifying treatment one applies.

Income data are far from following normal distribution, even if we take their logarithm. This aspect is also not negligible and is producing outliers while WLS or WIV models are fitted.

Negative values for DWH test

Users often report that the application of DWH tests can lead to negative test statistics, caused by estimated parameter variance differences that are not positive semi-definite. Or that the test statistic can become misleadingly small while still positive. This happened to us in our investigations and it is not reassuring, given that the test statistic should be χ^2 under the null hypothesis. Schreiber (2008) even states that the statistic can be asymptotically negative with large samples. He writes that this weakens the consensus (textbook) reaction of attributing such annoying results to peculiar random data constellations in given finite samples.

Dichotomous auxiliary or instrumental variables

How to handle dichotomous variables instrumented by continuous variables in a DWH test framework is also not clear. Geometrically, how

should the projection of such dichotomous variables through $\mathbf{P}_z$ look like? Shall we allow that it becomes a continuous variable or should it be dichotomized? If it has to be dichotomized, it will not belong to the hyperplane L_z any more. What are the effects then? Lollivier (2001) has done some work on the endogeneity of dichotomous variables, but it does not satisfy all our needs.

Conclusions on DWH tests

These tests only work if one can be sure that the instrumental variables used are exogenous. This means a set of clearly exogenous variables must be available. Only then is there a chance of detecting other endogenous auxiliary variables. This is a strong restriction. Thus the strategy is to start with a small set of variables for which one has good reasons to think that they are all exogenous. Then, one adds a supplementary variable to the WLS model and, provided that one more exogenous instrument is available[6], one can run the DWH test to check if the added variable is exogenous or not. And so on, all the variables at our disposal may be tested.

We successfully identified one endogenous variable at the time. But if more than one was present among the auxiliary variables, the problems mentioned in the previous paragraphs where creating too much "noise" and the results of the tests where not convincing, unstable and sometimes in contradiction to one another.

Intuitively, we do believe that a convincing adaptation of the DWH test in a survey framework should be possible, but the time was too short to come up with a satisfactory solution. Such tests would allow one to tune the choice of instruments when performing generalized calibration. It still assumes that the statistician has a secure way to be provided with an exogenous instrument, which may not be easy, but if so, it would already be of great help for the use of generalized calibration in practice. As we have shown in our article that generalized calibration can provide us with weights that are able to compensate for even nonignorable nonresponse, this field of research is surely worth of investing more time.

[6]If not, an exogenous auxiliary variable must be removed from the model so it can be used as an instrument.

General conclusion

The Phd thesis proposed here is of direct relevance to national statistical institutes and is endorsed by a productive collaboration with the Swiss Federal Office of Statistics (SFSO) where the author previously worked for almost nine years. The theories developed have been applied to real data to assess their validity.

The imputation strategy described in Chapter 5 is the main product of this thesis and constitutes the realization of my initial motivation when I left the SFSO. The other chapters result from problems I encountered in my quest for a new imputation method based on the use of GB2 distribution and generalized calibration or from topics taking their roots in my previous work at the SFSO.

Chapter 2 was co-written with Dr. Lionel Qualité and gives a good overview of the background and set-up with which I started my research. As a methodologist at a national statistical institute having participated in the implementation of the SILC survey in Switzerland, I had the chance to become familiar with several EU-projects (European Union, 2003, 2004, 2007, 2011a) often using SILC data among others. Evaluating the precision of the produced estimations has always been a crucial point. After having experimented with the use of bootstrap samples of weights for the Swiss Household Panel, we turned to linearization procedures because they were quicker and easier to handle. My interest in linearization procedures to estimate variances of complex statistics persisted during my thesis work and Chapter 3 reflects the substantial improvements that we brought. These improvements are of direct relevance for the SFSO which computes such variance estimates every year.

Chapter 4 needs to be completed. It is inspired by the interesting and complicated work of Deville & Särndal (1994) who present a variance estimate for a total through linearization taking regression imputation into account. I attempted to extend it to poverty indicators. One could then use macros to estimate the variance without doing multiple imputations. So far I reach some results for quantiles. Many inequality indicators are based on quantiles, which was a first step. But it is unfinished work.

In our imputation strategy presented in Chapter 5, except the data transformations which must be robust against extreme values and avoid normal distribution assumptions, we have focused on the advantages of generalized calibration tuned with the help of GB2 adjustment on the income data. Lesage & Haziza (2013b,a) show in their studies that the choice of the calibration function reflects assumptions on the form of the nonresponse mechanism, and therefore is not irrelevant. A bad choice would reflect a poor model of nonresponse far away from the true one and could lead to important bias and variance increases of the calibrated estimator.

We have seen that, with the data we tested, the choice of a good combination of instrumental and auxiliary variables is much more important than the choice of the calibration function. If the choice of the instrumental and auxiliary variables is good, the choice of the calibration function is secondary. In our case, the logit method gave the best results, but imputations derived from calibration weights with another calibration function also produced good results. Most of the estimated inequality indicators with the imputed data are not significantly different if one changes the calibration function used.

We also tried to impute via instrumental regression (in Step 4 of the proposed imputation strategy) applying the same combination of variables than the one found to be the best for defining the weights through generalized calibration. Although the obtained distribution of the imputed values became slightly better, the coherence within the data was poor. For the inequality indicators, the results were not bad since these statistics depend on the shape of the distribution. But on the unit level, the imputation errors were too large.

Of course, our imputation strategy is not directly applicable since we tuned it knowing the values for the nonrespondents. But the methodology could be applied in the following way. If one has two sets of units at one's disposal, a first set of "easy" respondents and another of "difficult" respondents, the tuning of the calibration could be done an attempt to impute the "difficult" respondents. Once the combination of instrumental and auxiliary variables is determined, the same combination could be used to really impute the nonrespondents. This is assuming that the nonrespondents are behaving similary to the "difficult" respondents. To identify the "difficult" respondents, the mandataries can conduct a nonresponse survey aiming to gain some more respondents who could not be convinced to answer to the main survey. Or using information from the survey fieldwork, "easy" respondents could be defined as those accepting to answer without going through a refusal conversion process.

The more we study a data file, the better we "feel" it, the better we are able to refine and improve our statistical models and treatments. It is probably an endless phenomenon. At a certain point, one needs to content oneself with the achieved quality in the allotted time. Thenceforth, as the methodologist declares a data file ready to be delivered to users, he will never state it thinking that nothing could be improved, but that the ratio quality/invested time is good. Similarly this thesis presents various researches where the ratio quality/invested time was found good. But that ratio is an indicator that never reaches its maximum. The author is convinced to have considerably improved the quality of the presented imputations (see Chapter 5). And this was at the cost of three years of work. The plot of quality against time invested, as the square root plot, has no asymptote, and, after a first phase, one must invest more and more time to increase the quality a bit more.

I hope that the work presented will inspire future methodologists and survey practitioners either by applying it directly or by improving it further.

Proof of Proposition 4.3

A

Proof of (4.20):

Proof. As, from (4.1), $y_k = \mathbf{x}_{k.}\hat{\boldsymbol{\beta}} + \varepsilon_k$ and $y_\ell = \mathbf{x}_{l.}\hat{\boldsymbol{\beta}} + \varepsilon_l$, since ε_i are iid, and still neglecting that $\hat{Q}_\alpha$ is random:

$$\begin{aligned}
\mathrm{E}_\xi(z_k^{Q_\alpha} z_l^{Q_\alpha}) &= \frac{1}{f^2(\hat{Q}_\alpha)}\frac{1}{\hat{N}^2}\mathrm{E}_\xi[\mathbf{1}_{[y_k\le\hat{Q}_\alpha]}\mathbf{1}_{[y_\ell\le\hat{Q}_\alpha]} - \alpha(\mathbf{1}_{[y_k\le\hat{Q}_\alpha]} + \mathbf{1}_{[y_\ell\le\hat{Q}_\alpha]}) + \alpha^2] \\
&= \frac{1}{f^2(\hat{Q}_\alpha)}\frac{1}{\hat{N}^2}\left\{\mathrm{E}_\xi(\mathbf{1}_{[y_k\le\hat{Q}_\alpha]}\mathbf{1}_{[y_\ell\le\hat{Q}_\alpha]}) - \alpha[\mathrm{E}_\xi(\mathbf{1}_{[y_k\le\hat{Q}_\alpha]}) + \mathrm{E}_\xi(\mathbf{1}_{[y_\ell\le\hat{Q}_\alpha]})] + \alpha^2\right\} \\
&= \frac{1}{f^2(\hat{Q}_\alpha)}\frac{1}{\hat{N}^2}[\Phi_{kl} - \alpha(\Phi_k + \Phi_\ell) + \alpha^2]. \qquad \text{(A.1)}
\end{aligned}$$

with (since, if $k = \ell$, $\mathbf{1}_{[y_k\le\hat{Q}_\alpha]}\mathbf{1}_{[y_k\le\hat{Q}_\alpha]} = \mathbf{1}_{[y_k\le\hat{Q}_\alpha]}$):

$$\Phi_{k\ell} = \begin{cases} \Phi_k\Phi_\ell & \text{si } k \neq \ell \\ \Phi_k & \text{si }, k = \ell \end{cases}$$

□

We therefore deduce in particular that for $k = l$:

$$\mathrm{E}_\xi[(z_k^{Q_\alpha})^2] = \frac{1}{f^2(\hat{Q}_\alpha)}\frac{1}{\hat{N}^2}[\Phi_k(1-2\alpha) + \alpha^2].$$

Proof of (4.21):

Proof.

$$\begin{aligned}
\mathrm{E}_\xi(\hat{z}_k^{Q_\alpha}) &= -\frac{1}{f(\hat{Q}_\alpha)}\frac{1}{\hat{N}}\left[\mathrm{E}_\xi(\mathbf{1}_{[\hat{y}_k\le\hat{Q}_\alpha]}) - \alpha\right] \\
&= -\frac{1}{f(\hat{Q}_\alpha)}\frac{1}{\hat{N}}\left[\mathrm{E}_\xi(\mathbf{1}_{[\mathbf{x}_{k.}\hat{\boldsymbol{\beta}}\le\hat{Q}_\alpha]}) - \alpha\right]. \\
\mathrm{E}_\xi(\mathbf{1}_{[\mathbf{x}_{k.}\hat{\boldsymbol{\beta}}\le\hat{Q}_\alpha]}) &= P_\xi(\mathbf{x}_{k.}\hat{\boldsymbol{\beta}} \le \hat{Q}_\alpha) \\
&= P_\xi(\mathbf{x}_{k.}\boldsymbol{\beta} + \mathbf{x}_{k.}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon} \le \hat{Q}_\alpha).
\end{aligned}$$

Defining $\boldsymbol{\eta}_k' := \mathbf{x}_{k\cdot}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, we have

$$\mathrm{E}_{\tilde{\xi}}(\hat{z}_k^{Q_\alpha}) = P_{\tilde{\xi}}\left(\sum_{i\in r}\eta_{ki}\varepsilon_i \leq \hat{Q}_\alpha - \mathbf{x}_{k\cdot}\boldsymbol{\beta}\right).$$

Since $\varepsilon_i\, iid. \sim N(0,\sigma^2)$, $\eta_{ki}\varepsilon_i \sim N(0,\sigma^2\eta_{ki}^2)$ and thus $\sum_{i\in r}\eta_{ki}\varepsilon_i \sim N(0,\sigma^2\sum_{i\in r}\eta_{ki}^2)$. Furthermore, $\sum_{i\in r}\eta_{ki}^2 = \boldsymbol{\eta}_k'\boldsymbol{\eta}_k = \mathbf{x}_{k\cdot}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_{k\cdot}' = \mathbf{x}_{k\cdot}\mathbf{T}^{-1}\mathbf{x}_{k\cdot}'$, it follows that:

$$\begin{aligned}
\mathrm{E}_{\tilde{\xi}}(\mathbf{1}_{[\mathbf{x}_{k\cdot}\hat{\boldsymbol{\beta}}\leq\hat{Q}_\alpha]}) &= P_{\tilde{\xi}}\left(\frac{\sum_{i\in r}\eta_{ki}\varepsilon_i}{\sigma\sqrt{\mathbf{x}_{k\cdot}\mathbf{T}^{-1}\mathbf{x}_{k\cdot}'}} \leq \frac{\hat{Q}_\alpha - \mathbf{x}_{k\cdot}\boldsymbol{\beta}}{\sigma\sqrt{\mathbf{x}_{k\cdot}\mathbf{T}^{-1}\mathbf{x}_{k\cdot}'}}\right)\\
&= \Phi\left(\frac{\hat{Q}_\alpha - \mathbf{x}_{k\cdot}\boldsymbol{\beta}}{\sigma\sqrt{\mathbf{x}_{k\cdot}\mathbf{T}^{-1}\mathbf{x}_{k\cdot}'}}\right)\\
&:= M_k.
\end{aligned}$$

Thus,

$$\mathrm{E}_{\tilde{\xi}}(\hat{z}_k^{Q_\alpha}) = -\frac{1}{f(\hat{Q}_\alpha)}\frac{1}{\hat{N}}\left[M_k - \alpha\right]. \tag{A.2}$$

□

Proof of (4.22):

Proof.

$$\begin{aligned}
\mathrm{E}_{\tilde{\xi}}(\hat{z}_k^{Q_\alpha}\hat{z}_l^{Q_\alpha}) &= \frac{1}{f^2(\hat{Q}_\alpha)}\frac{1}{\hat{N}^2}\mathrm{E}_{\tilde{\xi}}[\mathbf{1}_{[\hat{y}_k\leq\hat{Q}_\alpha]}\mathbf{1}_{[\hat{y}_\ell\leq\hat{Q}_\alpha]} - \alpha(\mathbf{1}_{[\hat{y}_k\leq\hat{Q}_\alpha]} + \mathbf{1}_{[\hat{y}_\ell\leq\hat{Q}_\alpha]}) + \alpha^2]\\
&= \frac{1}{f^2(\hat{Q}_\alpha)}\frac{1}{\hat{N}^2}\left\{\mathrm{E}_{\tilde{\xi}}(\mathbf{1}_{[\hat{y}_k\leq\hat{Q}_\alpha]}\mathbf{1}_{[\hat{y}_\ell\leq\hat{Q}_\alpha]}) - \alpha[\mathrm{E}_{\tilde{\xi}}(\mathbf{1}_{[\hat{y}_k\leq\hat{Q}_\alpha]}) + \mathrm{E}_{\tilde{\xi}}(\mathbf{1}_{[\hat{y}_\ell\leq\hat{Q}_\alpha]})] + \alpha^2\right\}
\end{aligned}$$

$$\begin{aligned}
\mathrm{E}_{\tilde{\xi}}(\mathbf{1}_{[\hat{y}_k\leq\hat{Q}_\alpha]}\mathbf{1}_{[\hat{y}_\ell\leq\hat{Q}_\alpha]}) &= \mathrm{E}_{\tilde{\xi}}(\mathbf{1}_{[\mathbf{x}_{k\cdot}\boldsymbol{\beta}+\boldsymbol{\eta}_k'\boldsymbol{\varepsilon}\leq\hat{Q}_\alpha]}\mathbf{1}_{[\mathbf{x}_{l\cdot}\boldsymbol{\beta}+\boldsymbol{\eta}_k'\boldsymbol{\varepsilon}\leq\hat{Q}_\alpha]})\\
&= P_{\tilde{\xi}}\left(\frac{\boldsymbol{\eta}_k'\boldsymbol{\varepsilon}}{\sigma\sqrt{\mathbf{x}_{k\cdot}\mathbf{T}^{-1}\mathbf{x}_{k\cdot}'}} \leq \frac{\hat{Q}_\alpha - \mathbf{x}_{k\cdot}\boldsymbol{\beta}}{\sigma\sqrt{\mathbf{x}_{k\cdot}\mathbf{T}^{-1}\mathbf{x}_{k\cdot}'}} \bigcap \frac{\boldsymbol{\eta}_l'\boldsymbol{\varepsilon}}{\sigma\sqrt{\mathbf{x}_{l\cdot}\mathbf{T}^{-1}\mathbf{x}_{l\cdot}'}} \leq \frac{\hat{Q}_\alpha - \mathbf{x}_{l\cdot}\boldsymbol{\beta}}{\sigma\sqrt{\mathbf{x}_{l\cdot}\mathbf{T}^{-1}\mathbf{x}_{l\cdot}'}}\right)
\end{aligned}$$

Both events are distributed according to a standard normal distribution. Their intersection thus follows a two-dimensional normal distribution:

$$\mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & \hat{\rho}_{kl}^M\\ \hat{\rho}_{kl}^M & 1\end{pmatrix}\right)$$

$\hat{\rho}_{kl}^M$ is computable, thus this probability can be estimated numerically and we denote it M_{kl}, with

$$\begin{aligned}
\hat{\rho}_{kl}^{M} &= cov_{\xi}\left(\frac{\boldsymbol{\eta}_k'\boldsymbol{\varepsilon}}{\sigma\sqrt{\mathbf{x}_{k.}\mathbf{T}^{-1}\mathbf{x}_{k.}'}}, \frac{\boldsymbol{\eta}_l'\boldsymbol{\varepsilon}}{\sigma\sqrt{\mathbf{x}_{l.}\mathbf{T}^{-1}\mathbf{x}_{l.}'}}\right) \\
&= \frac{\boldsymbol{\eta}_k'}{\sigma\sqrt{\mathbf{x}_{k.}\mathbf{T}^{-1}\mathbf{x}_{k.}'}} cov_{\xi}(\boldsymbol{\varepsilon},\boldsymbol{\varepsilon}) \frac{\boldsymbol{\eta}_l}{\sigma\sqrt{\mathbf{x}_{l.}\mathbf{T}^{-1}\mathbf{x}_{l.}'}} \\
&= \frac{\boldsymbol{\eta}_k'}{\sigma\sqrt{\mathbf{x}_{k.}\mathbf{T}^{-1}\mathbf{x}_{k.}'}} \sigma^2\mathbf{1} \frac{\boldsymbol{\eta}_l}{\sigma\sqrt{\mathbf{x}_{l.}\mathbf{T}^{-1}\mathbf{x}_{l.}'}} \\
&= \frac{\boldsymbol{\eta}_k'\boldsymbol{\eta}_l}{\sqrt{\mathbf{x}_{k.}\mathbf{T}^{-1}\mathbf{x}_{k.}'}\sqrt{\mathbf{x}_{l.}\mathbf{T}^{-1}\mathbf{x}_{l.}'}} \\
&= \frac{\mathbf{x}_{k.}\mathbf{T}^{-1}\mathbf{x}_{l.}'}{\sqrt{(\mathbf{x}_{k.}\mathbf{T}^{-1}\mathbf{x}_{k.}')(\mathbf{x}_{l.}\mathbf{T}^{-1}\mathbf{x}_{l.}')}}.
\end{aligned}$$

Thus:

$$\mathrm{E}_{\xi}(\hat{z}_k^{Q_\alpha}\hat{z}_l^{Q_\alpha}) = \frac{1}{f^2(\hat{Q}_\alpha)}\frac{1}{\hat{N}^2}\left[M_{kl} - \alpha(M_k + M_l) + \alpha^2\right]. \qquad \text{(A.3)}$$

□

We note in particular that if $k = l$, $\hat{\sigma}_{kk} = 1$ and so $M_{kk} = M_k$, it follows that:

$$\mathrm{E}_{\xi}[(\hat{z}_k^{Q_\alpha})^2] = \frac{1}{f^2(\hat{Q}_\alpha)}\frac{1}{\hat{N}^2}\left[M_k(1-2\alpha) + \alpha^2\right].$$

Proof of (4.23):

Proof. By the same methods as in the previous demonstration, we have:

$$\begin{aligned}
\mathrm{E}_{\xi}(\mathbf{1}_{[y_k \le \hat{Q}_\alpha]}\mathbf{1}_{[\hat{y}_\ell \le \hat{Q}_\alpha]}) &= \mathrm{E}_{\xi}(\mathbf{1}_{[\mathbf{x}_{k.}\boldsymbol{\beta}+\varepsilon_k \le \hat{Q}_\alpha]}\mathbf{1}_{[\mathbf{x}_{l.}\boldsymbol{\beta}+\boldsymbol{\eta}_k'\boldsymbol{\varepsilon} \le \hat{Q}_\alpha]}) \\
&= P_{\xi}\left(\frac{\varepsilon_k}{\sigma} \le \frac{\hat{Q}_\alpha - \mathbf{x}_{k.}\boldsymbol{\beta}}{\sigma} \bigcap \frac{\boldsymbol{\eta}_l'\boldsymbol{\varepsilon}}{\sigma\sqrt{\mathbf{x}_{l.}\mathbf{T}^{-1}\mathbf{x}_{l.}'}} \le \frac{\hat{Q}_\alpha - \mathbf{x}_{l.}\boldsymbol{\beta}}{\sigma\sqrt{\mathbf{x}_{l.}\mathbf{T}^{-1}\mathbf{x}_{l.}'}}\right)
\end{aligned}$$

Both events are distributed according to a standard normal distribution. Their intersection thus follows a two-dimensional normal distribution:

$$\mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \hat{\rho}_{kl}^{\Phi M} \\ \hat{\rho}_{kl}^{\Phi M} & 1 \end{pmatrix}\right)$$

$\hat{\rho}_{kl}^{\Phi M}$ is computable, thus this probability can be estimated numerically and we denote it ΦM_{kl}, with

$$\begin{aligned}
\hat{\rho}_{kl}^{\Phi M} &= cov_{\xi}\left(\frac{\varepsilon_k}{\sigma}, \frac{\boldsymbol{\eta}_l'\boldsymbol{\varepsilon}}{\sigma\sqrt{\mathbf{x}_{l.}\mathbf{T}^{-1}\mathbf{x}_{l.}'}}\right) \\
&= \frac{1}{\sigma} cov_{\xi}(\varepsilon_k, \boldsymbol{\varepsilon}') \frac{\boldsymbol{\eta}_l}{\sigma\sqrt{\mathbf{x}_{l.}\mathbf{T}^{-1}\mathbf{x}_{l.}'}} \\
&= \frac{\mathbf{x}_{k.}\mathbf{T}^{-1}\mathbf{x}_{l.}'}{\sqrt{\mathbf{x}_{l.}\mathbf{T}^{-1}\mathbf{x}_{l.}'}},
\end{aligned}$$

thus

$$\mathrm{E}_{\xi}(z_k^{Q_\alpha}\hat{z}_\ell^{Q_\alpha}) = \frac{1}{f^2(\hat{Q}_\alpha)}\frac{1}{\hat{N}^2}\left[\Phi M_{k\ell} - \alpha(\Phi_k + M_\ell) + \alpha^2\right]. \quad \text{(A.4)}$$

□

Proof of Proposition 4.4

B

$$V_{imp\tilde{\xi}} = \mathrm{E}_{\tilde{\xi}}\{(\hat{Z}_\bullet - \hat{Z})^2|s,r\} + 2\mathrm{E}_{\tilde{\xi}}[(\hat{Z}_\bullet - \hat{Z})(\hat{Z} - Z)|s,r]$$

Proof. For the left-hand side of $V_{imp\tilde{\xi}}$:

$$\begin{aligned}\mathrm{E}_{\tilde{\xi}}[(\hat{Z}_\bullet - \hat{Z})^2] &= \mathrm{E}_{\tilde{\xi}}[(\hat{Z}^o_\bullet - \hat{Z}^o)^2] \\ &= \mathrm{E}_{\tilde{\xi}}[(\hat{Z}^o_\bullet)^2] - 2\mathrm{E}_{\tilde{\xi}}(\hat{Z}^o\hat{Z}^o_\bullet) + \mathrm{E}_{\tilde{\xi}}[(\hat{Z}^o)^2]\end{aligned}$$

Observe that:

$$\begin{aligned}\mathrm{E}_{\tilde{\xi}}[(\hat{Z}^o_\bullet)^2] &= \sum_{k\in o}\sum_{\ell\in o} w_k w_\ell \mathrm{E}_{\tilde{\xi}}(\hat{z}_k\hat{z}_\ell) \\ &\overset{(4.22)}{=} \frac{1}{f^2(\hat{Q}_\alpha)}\frac{1}{\hat{N}^2}\left\{\sum_{k\in o}\sum_{\ell\in o} w_k w_\ell[M_{k\ell} - \alpha(M_k + M_\ell) + \alpha^2]\right\}\end{aligned}$$

$$\begin{aligned}\mathrm{E}_{\tilde{\xi}}(\hat{Z}^o\hat{Z}^o_\bullet) &= \sum_{k\in o}\sum_{\ell\in o} w_k w_\ell \mathrm{E}_{\tilde{\xi}}(z_k\hat{z}_\ell) \\ &\overset{(4.23)}{=} \frac{1}{f^2(\hat{Q}_\alpha)}\frac{1}{\hat{N}^2}\left\{\sum_{k\in o}\sum_{\ell\in o} w_k w_\ell[\Phi M_{k\ell} - \alpha(\Phi_k + M_l) + \alpha^2]\right\}\end{aligned}$$

$$\begin{aligned}\mathrm{E}_{\tilde{\xi}}[(\hat{Z}^o)^2] &= \sum_{k\in o}\sum_{\ell\in o} w_k w_\ell \mathrm{E}_{\tilde{\xi}}(z_k z_\ell) \\ &\overset{(4.20)}{=} \frac{1}{f^2(\hat{Q}_\alpha)}\frac{1}{\hat{N}^2}\left\{\sum_{k\in o}\sum_{\ell\in o} w_k w_\ell[\Phi_{k\ell} - \alpha(\Phi_k + \Phi_\ell) + \alpha^2]\right\}\end{aligned}$$

Thus:

$$\mathrm{E}_{\tilde{\xi}}[(\hat{Z}_\bullet - \hat{Z})^2] = \frac{1}{f^2(\hat{Q}_\alpha)}\frac{1}{\hat{N}^2}\left\{\sum_{k\in o}\sum_{\ell\in o} w_k w_\ell[M_{k\ell} - 2\Phi M_{k\ell} + \Phi_{k\ell}]\right\} \quad \text{(B.1)}$$

For the right-hand side of $V_{imp\xi}$:

$$\begin{aligned}
\mathrm{E}_{\xi}[(\hat{Z}_{\bullet}-\hat{Z})(\hat{Z}-Z)|s,r] &= \mathrm{E}_{\xi}[(\hat{Z}_{\bullet}-\hat{Z})(\hat{Z}-Z)|s,r] \\
&= \mathrm{E}_{\xi}(\hat{Z}^{o}_{\bullet}\hat{Z}) - \mathrm{E}_{\xi}(\hat{Z}^{o}_{\bullet}Z) - \mathrm{E}_{\xi}(\hat{Z}^{o}\hat{Z}) + \mathrm{E}_{\xi}(\hat{Z}^{o}Z) \\
&= \frac{1}{f^2(\hat{Q}_{\alpha})}\frac{1}{\hat{N}^2}\left\{\sum_{k\in o}\sum_{\ell\in s} w_k w_\ell[\Phi M_{\ell k} - \alpha(\Phi_\ell + M_k) + \alpha^2]\right. \\
&\quad - \sum_{k\in o}\sum_{\ell\in U} w_k[\Phi M_{\ell k} - \alpha(\Phi_\ell + M_k) + \alpha^2] \\
&\quad - \sum_{k\in o}\sum_{\ell\in s} w_k w_\ell[\Phi_{k\ell} - \alpha(\Phi_k + \Phi_\ell) + \alpha^2] \\
&\quad \left. + \sum_{k\in o}\sum_{\ell\in U} w_k[\Phi_{k\ell} - \alpha(\Phi_k + \Phi_\ell) + \alpha^2]\right\} \\
&= \frac{1}{f^2(\hat{Q}_{\alpha})}\frac{1}{\hat{N}^2}\left\{\sum_{k\in o}\sum_{\ell\in s} w_k w_\ell \Phi M_{\ell k} - \sum_{k\in o}\sum_{\ell\in U} w_k \Phi M_{\ell k}\right. \\
&\quad \left. + \sum_{k\in o}\sum_{\ell\in U} w_k \Phi_{k\ell} - \sum_{k\in o}\sum_{\ell\in s} w_k w_\ell \Phi_{k\ell}\right\}. \qquad \text{(B.2)}
\end{aligned}$$

The calculation of (B.2) requires information on the population U which is not available. So we will approximate the sums on U by weighted sums on s. By doing so all vanishes. (B.2) was calculated for some samples of the mentioned simulations. Indeed, in the case of simulations the entire population is known. The calculation is possible but long! The term (B.2) represents less than half a percent of the term (B.1). In other words, the variance due to imputation is given by (B.1):

$$V_{imp\xi} = \frac{1}{f^2(\hat{Q}_{\alpha})}\frac{1}{\hat{N}^2}\left\{\sum_{k\in o}\sum_{\ell\in o} w_k w_\ell (M_{k\ell} - 2\Phi M_{k\ell} + \Phi_{k\ell})\right\}$$

□

Expression for $C^{Q_\alpha,sas}$

C

One obtains (4.28) by the following: starting from (4.24),

$$
\begin{aligned}
C^{Q_\alpha,sas} &= \frac{1}{f^2(\hat{Q}_\alpha)}\frac{1}{\hat{N}^2}\left\{\sum_{k\in o}\sum_{\ell\in o}A_{k\ell}[\Phi_{kl}-M_{kl}-\alpha(\Phi_k+\Phi_\ell-M_k-M_l)]\right.\\
&+ \left. 2\sum_{k\in r}\sum_{\ell\in o}A_{k\ell}[\Phi_{kl}-\Phi M_{k\ell}-\alpha(\Phi_\ell-M_\ell)]\right\}\\
&= \frac{1}{f^2(\hat{Q}_\alpha)}\frac{1}{N^2}\left\{\sum_{k\in o}\sum_{\ell\in o}\frac{N(N-n)}{n^2}(\Phi_k-M_k)(1-2\alpha)\right.\\
&\quad +\sum_{k\in o}\sum_{\ell\in o,\ell\neq k}-\frac{N(N-n)}{n^2(n-1)}[\Phi_{k\ell}-\Phi M_{k\ell}-2\alpha(\Phi_k-M_k)]\\
&\quad \left.+2\sum_{k\in r}\sum_{\ell\in o}-\frac{N(N-n)}{n^2(n-1)}[\Phi_{k\ell}-\Phi M_{k\ell}-\alpha(\Phi_\ell-M_l)]\right\}\\
&= \frac{1}{f^2(\hat{Q}_\alpha)}\frac{N-n}{N}\left\{\frac{1}{n^2}(\Phi_o-M_o)(1-2\alpha)\right.\\
&\quad -\frac{1}{n^2(n-1)}[\Phi_{\overline{oo}}-M_{\overline{oo}}-2\alpha(n_o-1)(\Phi_o-M_o)]\\
&\quad \left.-\frac{1}{n^2(n-1)}[2\Phi_{ro}-2\Phi M_{ro}-2\alpha n_r(\Phi_o-M_o)]\right\}\\
&= \frac{1}{f^2(\hat{Q}_\alpha)}\frac{N-n}{N}\left\{\frac{1}{n^2}(\Phi_o-M_o)(1-2\alpha)\right.\\
&\quad \left.-\frac{1}{n^2(n-1)}[\Phi_{\overline{oo}}-M_{\overline{oo}}+2(\Phi_{ro}-\Phi M_{ro})-2\alpha(n-1)(\Phi_o-M_o)]\right\}\\
&= \frac{1}{f^2(\hat{Q}_\alpha)}\frac{N-n}{N}\left\{\frac{1}{n^2}(\Phi_o-M_o)(1-2\alpha)\right.\\
&\quad \left.-\frac{1}{n^2(n-1)}\left[\Phi_{\overline{oo}}-M_{\overline{oo}}+2(\Phi_{ro}-\Phi M_{ro})\right]+\frac{2\alpha}{n^2}(\Phi_o-M_o)\right\}\\
&= \frac{1}{f^2(\hat{Q}_\alpha)}\frac{N-n}{Nn^2}\left\{\Phi_o-M_o-\frac{1}{n-1}\left[\Phi_{\overline{oo}}-M_{\overline{oo}}+2(\Phi_{ro}-\Phi M_{ro})\right]\right\}. \qquad \square
\end{aligned}
$$

Definition of the variables used

D

Table D.1 – *The first column lists the variables at our disposal, the $x_{.j}$-column is hooked if the variables were retained as auxiliary variables. Similarly the $z_{.j}$-column is hooked if they were retained as instrumental variables. The variables with an asterisk are continuous, all others are dichotomic. The fourth column is hooked if the variables appear in the segmentation tree (see Appendix E) modelling nonresponse. The last column is hooked if the variable correlates at least to 10% with the interst variable y_{CCO}.*

Variable name	$\mathbf{x}_{.j}$	$\mathbf{z}_{.j}$	Explains NR to y_{CCO}	Explains y_{CCO}	Description
AC_2	√		√	√	Age class 16-19 y. old
AC_3	√	√	√	√	Age class 20-29 y. old
AC_5	√	√	√	√	Age class 40-49 y. old
ALLOCfam	√	√	√	√	receives family allowances
CANTONo9_21	√	√	√		canton of Ticino
CANTONo9_25	√	√	√		canton of Geneva
CHEF	√	√	√	√	has managing position
COM2_09_2	√	√	√		suburban municipalities
COM2_09_3	√	√		√	high-income municipalities
DIVORCE	√	√	√		divorced
ECHOR_1	√	√	√		sampled in wave 2007
ECHOR_3	√	√	√		sampled in wave 2009
EDUC_12	√	√	√	√	education level: prof./techn. school
EDUC_14	√	√		√	education level: teaching high school
EDUC_15	√	√		√	education level: specialised high school
EDUC_16	√	√		√	education level: university level
EDUC_2	√	√		√	education level: obligatory school
EDUC_6	√	√	√	√	education level: 3 years apprenticeship
*log(HH_{cost})		√		√	log. of housing costs
HHtypeo9_3	√	√	√		HH of two adults below 65 y. old
HHtypeo9_4	√	√	√		HH of two adults, one at least over 65 y. old
HHtypeo9_9	√	√	√		HH of two adults + at least 3 children
HOMME	√	√	√	√	man
KIDS	√	√	√		kids in HH
LOGSE200m2	√	√	√	√	housing surface $\geq 200m^2$
LOGm2MISS	√	√	√	√	housing surface missing
MAISON	√	√	√		living in a house
MARIE	√	√	√	√	married
NPACT_2	√	√	√		2 persons active in HH
Ncall11TO35	√	√	√		11 to 35 phone-calls attempts were neede to reach HH
Ncall4TO10	√	√	√		4 to 10 phone-calls attempts were neede to reach HH
OCCUP_13	√	√	√	√	apprentice
PROPRIO	√	√	√		owner of housing
SANTE_1	√	√	√		very good health status
TRAVWE	√	√	√		is working on weekends
*TX_TP		√		√	occupation rate in %
TXactE100	√	√	√	√	occupation rate = 100%
TXactL100	√	√	√		occupation rate < 100%
TXactLE20			√	√	occupation rate ≤ 20%
TXactLE40			√	√	20% < occupation rate ≤ 40%
TXactLE60			√	√	40% < occupation rate ≤ 60%
TXactLE80	√		√		60% < occupation rate ≤ 80%
actINDEP	√	√	√		self employed
*log[med(y_{CCO})]	√			√	logarithm of median income per RHG
prestaCHOM	√	√	√		receives unemployment allowance
prestaCP	√	√	√		receives complementary allowance
salair1314	√	√	√		receives 13th and/or 14th salary
*log(y_{CCO})		√			logarithm of income y_{CCO}

Segmentation tree for y_{CCO}

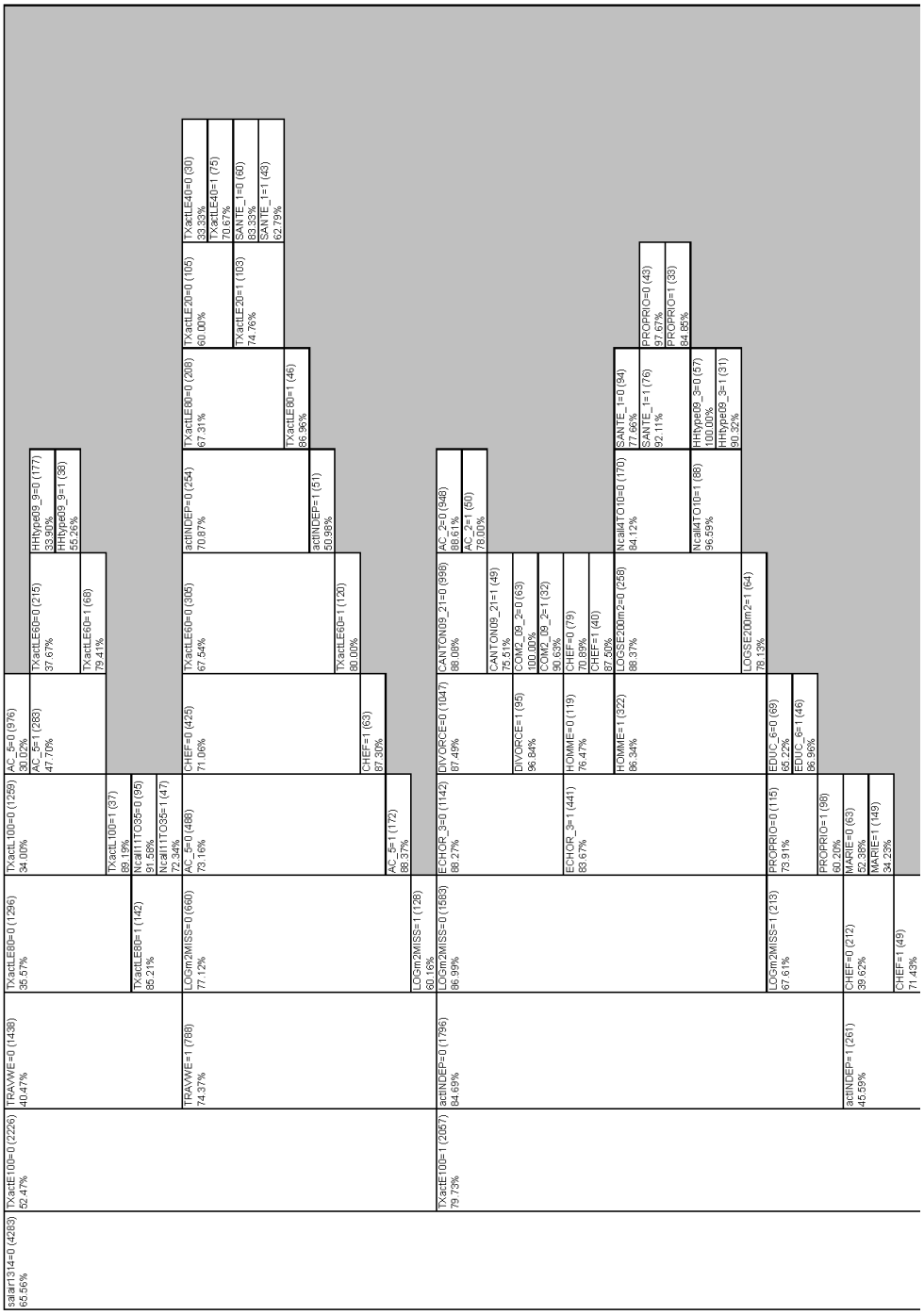


Figure E.1 – *Segmentation tree first part. Refer to Table D.1 for the definitions of variables.*

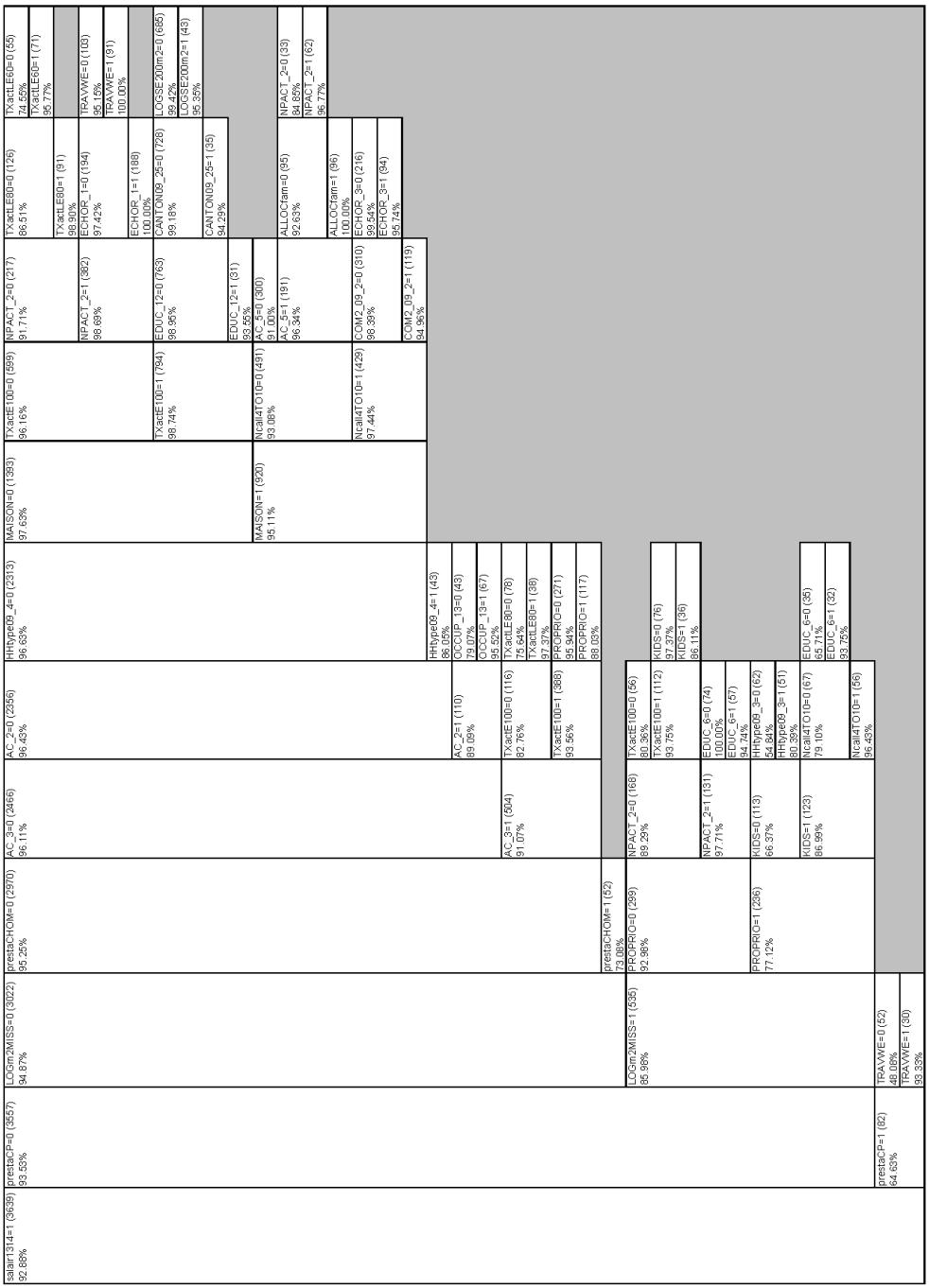


Figure E.2 – *Segmentation tree second part. Refer to Table D.1 for the definitions of variables.*

Imputation classes with segmentation trees

F

This appendix completes the Sections 4.7.3 and 5.8.3 giving some details about how we create imputation classes for income variables.

At the SFSO, imputation of several income variables was implemented in the SILC and Swiss Labour Force Survey (SLFS). As mentioned in Section 2.5.4, after a long evaluation procedure, the SFSO opted for the IVEware software to perform the imputations. One of the drawbacks of IVEware is that it assumes that the variables to be imputed follow a normal distribution. It is far from being the case for the variables to be imputed (employed or self-employed salary, household equivalent income, different rents and pensions, taxes): most of these variables have a distribution with heavy and long right-tail and a much thicker left tail (many very small incomes with respect to the mean or median) than a normal distribution. After having worked a few years with such data, we realized that, in many cases, low / middle / high incomes tend to behave differently. Sometimes this is easily explainable. For example, in the employed salary category, people following an apprenticeship or students having summer jobs have very small annual incomes compared to other workers. Nevertheless their income is reported as the ones of the other workers.

We therefore decided to form imputation classes attempting to separate the incomes in three categories: low, medium and high incomes. IVEware was then used independently to fit imputation regression models in each imputation class.

These three imputation classes were formed with the help of segmentation trees (see Sections 2.4.2 and 5.8.1) aiming to identify the characteristics of poor and rich units. Among the respondents, dichotomous variables identifying poor (defined with having and income $\leq Q_{25}$) and rich people (defined with having and income $\geq Q_{75}$) were created. All variables potentially able to explain status poor or rich were used (and split up in one or many dichotomous variables when necessary). One segmentation tree was constructed to model the fact of being poor, and one for being rich. Both segmentation trees prompted us to define two times three categories of units depending on the response rate of each Response Homogeneity Group (RHG):

- Segmentation tree for low incomes:

  - Poor units: response rate per RHG $\geq$ 66%,
  - undefined poor units: response rate per RHG between 33% and 66%,
  - Non poor units: response rate per RHG $\leq$ 33%.

- Segmentation tree for high incomes:
  - Rich units: response rate per RHG $\geq$ 66%,
  - undefined rich units: response rate per RHG between 33% and 66%,
  - Non rich units: response rate per RHG $\leq$ 33%.

Three imputation classes were then defined as follows:

- low incomes: identified as poor in the tree for low incomes and as non rich or undefined rich units in the tree for high incomes,
- high incomes: identifed as rich in the tree for high incomes and non poor or undefined poor units in the tree for low incomes,
- middle incomes: all other categories.

This system of forming imputation classes seems to work fairly well in separating the very low and very high from the middle class containing over 80% of the data.

Bibliography

Antal, E., Langel, M. & Tillé, Y. (2011). Variance estimation of inequality indices in complex sampling designs. In *Proc. 58th World Statistical Congress*. Dublin. Session IPS056. (Cited on pages 75 and 76)

Antal, E. & Tillé, Y. (2011). A direct bootstrap method for complex sampling designs from a finite population. *Journal of the American Statistical Association* **106**, 534–543. (Cited on page 67)

Ardilly, P. & Osier, G. (2007). Cross-sectional variance estimation fot the french "labour force survey". *Survey Research Methods* **1**, 75–83. (Cited on page 75)

Beaumont, J.-F., Haziza, D. & Ruiz-Gazen, A. (2013). A unified approach to robust estimation in finite population sampling. *Biometrika* **100**, 555–569. (Cited on page 80)

Brewer, K. R. W., Early, L. J. & Hanif, M. (1984). Poisson, modified Poisson and collocated sampling. *Journal of Statistical Planning and Inference* **10**, 15–30. (Cited on page 39)

Brewer, K. R. W., Early, L. J. & Joyce, S. F. (1972). Selecting several samples from a single population. *Australian Journal of Statistics* **3**, 231–239. (Cited on pages 37 and 39)

Cameron, A. C. & Pravin, K. T. (2005). *Microeconometrics: Methods and Applications*. Cambridge University Press. (Cited on page 123)

Canty, A. J. & Davison, A. C. (1998). Variance estimation for two complex surveys in switzerland. document interne, Office Fdral de la Statistique. (Cited on page 68)

Cauchon, C. (2006). Swiss household panel and silc. bootstrap samples and variances. Tech. rep., Statistics Canada, Ottawa. (Cited on page 68)

Chambers, R. (1986). Outlier robust finite population estimation. *Journal of the American Statistical Association* **81**, 1063–1069. (Cited on pages 23, 24, 82, 128, and 145)

Chauvet, G. (2007). *Méthodes de Bootstrap en Population Finie*. Thèse de doctorat, Université Rennes 2, Rennes, France. (Cited on page 67)

Conover, W. J. (1999). *Practical nonparametric statistics*. New York: J. Wiley, cop., 3rd ed. (Cited on pages 129 and 145)

Cotton, F. & Hesse, C. (1992a). Co-ordinated selection of stratified samples. In *Proceedings of Statistics Canada Symposium*. (Cited on page 37)

Cotton, F. & Hesse, C. (1992b). Tirages coordonnés d'échantillons. Document de travail de la Direction des Statistiques Économiques E9206. Tech. rep., INSEE, Paris. (Cited on page 37)

Croux, C. (1998). Limit behaviour of the empirical influence function of the median. *Statistics & Probability Letters* **37**, 331–340. (Cited on pages 80, 85, and 115)

Dastrup, S. R., Hartshorn, R. & McDonald, J. B. (2007). The impact of taxes and transfer payments on the distribution of income: A parametric comparison. *Journal of Economic Inequality* **5**, 353–369. (Cited on pages 70 and 119)

David, I. P. & Sukhatme, B. V. (1974). On the bias and mean square error of the ratio estimator. *Journal of the American Statistical Association* **69**, 464–466. (Cited on page 124)

Davidson, R. & MacKinnon, J. K. (1989). Testing for consistency using artificial regressions. *Econometric Theory* **5**, 363–384. (Cited on page 123)

Davidson, R. & MacKinnon, J. K. (1993). *Estimation and inference in econometrics*. Oxford University Press, Inc. (Cited on pages 123, 126, and 154)

Davison, A. (2003). *Statistical Models*. Cambridge: Cambridge University Press. (Cited on page 131)

Davison, A. C. & Hinkley, D. V. (1997). *Bootstrap Methods and Their Application*. Cambridge: Cambridge University Press. (Cited on page 67)

De Ree, S. J. M. (1983). A system of co-ordinated sampling to spread response burden of enterprises. In *44th Session of the ISI Madrid*. (Cited on page 37)

Demnati, A. & Rao, J. N. K. (2004a). Estimateurs de variance par linéarisation pour des données d'enquête. *Techniques d'enquête* **30**, 17–27. (Cited on pages 69, 75, and 76)

Demnati, A. & Rao, J. N. K. (2004b). Linearization variance estimators for survey data. *Survey Methodology* **30**, 17–26. (Cited on pages 25, 92, and 115)

Demnati, A. & Rao, J. N. K. (2004c). Linearization variance estimators for survey data (with discussion). *Survey Methodology* **30**, 17–34. (Cited on pages 131 and 144)

Deng, L.-Y. & Chhikara, R. S. (1991). On asymptotically design-unbiased estimators of a finite population mean. *Biometrika* **78**, 189–195. (Cited on page 124)

Deville, J.-C. (1999a). Estimation de variance pour des statistiques et des estimateurs complexes : techniques de résidus et de linéarisation. *Techniques d'enquête* **25**, 219–230. (Cited on pages 68 and 69)

Deville, J.-C. (1999b). Variance estimation for complex statistics and estimators: linearization and residual techniques. *Survey Methodology* **25**, 193–204. (Cited on pages 25, 73, 74, 75, 76, 80, 82, 83, 87, 92, 100, 101, 115, 131, and 144)

Deville, J.-C. (2000). Generalized calibration and application to weighting for non-response. In *Compstat - Proceedings in Computational Statistics: 14th Symposium Held in Utrecht, The Netherlands*. New York: Springer. (Cited on page 123)

Deville, J.-C. (2002). La correction de la nonréponse par calage généralisé. In *Actes des Journées de Méthodologie Statistique*. Paris: INSEE. (Cited on pages 54, 55, 69, 121, 123, and 127)

Deville, J.-C. & Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association* **87**, 376–382. (Cited on pages 39, 48, 54, 55, 74, 100, 121, and 131)

Deville, J.-C. & Särndal, C.-E. (1994). Variance estimation for the regression imputed Horvitz-Thompson estimator. *Journal of Official Statistics* **10**, 381–394. (Cited on pages 19, 21, 25, 91, 92, 93, 94, 95, 96, 97, 98, 100, 110, 115, 132, and 159)

Deville, J.-C., Särndal, C.-E. & Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association* **88**, 1013–1020. (Cited on pages 55, 121, and 122)

Dufour, J., Gagnon, F., Morin, Y., Renaud, M. & Särndal, C.-E. (1998). Measuring the impact of alternative weighting schemes for longitudinal data. In *Survey Research Methods Section*. American Statistical Association. (Cited on page 53)

Durbin, J. (1954). Errors in variables. *Review of the International Statistical Institute* **22**, 23–32. (Cited on pages 70, 145, and 152)

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics* **7**, 1–26. (Cited on page 67)

Ernst, L. R. (1996). Maximizing the overlap of sample units for two designs with simultaneous selection. *Journal of Official Statistics* **12**, 33–45. (Cited on page 37)

European Union (2003). Project EUREDIT : The development and evaluation of new methods for editing and imputation. (Cited on pages 60 and 159)

European Union (2004). Project DACSEIS : DAta quality in Complex Surveys within the new European Information Society. FP5. (Cited on pages 69 and 159)

European Union (2007). Project EDIMBUS : Recommended practices for EDiting and IMputation in cross-sectional BUsiness Surveys. Luzi, O. et al. (Cited on pages 60 and 159)

European Union (2008). Project KEI : Knowledge Economy Indicators. FP6. (Cited on page 69)

European Union (2011a). Project AMELI : Advanced Methodology for European Laeken Indicators. FP7. (Cited on pages 24, 69, 70, 119, 121, 144, 147, 149, 151, and 159)

European Union (2011b). Project SAMPLE : Small Area Methods for Poverty and Living Condition Estimates. FP7. (Cited on page 69)

European Union (2014). Project EU-SILC : European Union Statistics on Income and Living Conditions. (Cited on page 92)

Eurostat (2003). Laeken indicators - detailed calculation methodology. Doc.e2/ipse/2003, Directorate E: Social Statistics, Unit E-2: Living Conditions, Luxembourg. (Cited on pages 24, 92, 108, 118, 144, and 149)

Eurostat (2004a). Common cross-sectional eu indicators based on eu-silc; the gender pay gap. Working papers and studies, Office for Official Publications of the European Communities, Luxembourg. EU-SILC 131-rev/04. (Cited on pages 69, 79, and 80)

Eurostat (2004b). Cross sectional weighting: first year of each subsample. Working papers and studies, Office for Official Publications of the European Communities, Luxembourg. EU-SILC 134/04. (Cited on page 35)

Eurostat (2004c). Theoretical study of the gini index. Working papers and studies, Office for Official Publications of the European Communities, Luxembourg. EU-SILC 131-A/04. (Cited on pages 69 and 78)

Eurostat (2005a). The continuity of indicators during the transition between ECHP and EU-SILC. Working papers and studies, Office for Official Publications of the European Communities, Luxembourg. (Cited on pages 24, 69, 73, 89, 92, 108, 115, 118, and 144)

Eurostat (2005b). Cross-sectional weighting: from second year of the survey onwards. Working papers and studies, Office for Official Publications of the European Communities, Luxembourg. EU-SILC 157/05. (Cited on page 35)

Eurostat (2013). Handbook on precision requirements and variance estimation for ESS household surveys. (Cited on pages 69 and 131)

Fellegi, P. & Holt, D. (1976). A systematic approach to automatic edit and imputation. *Journal of the American Statistical Association* **71**, 17–35. (Cited on pages 63 and 64)

Feusi Widmer, R. (2004). L'Enquête Suisse sur la Population Active (ESPA). Tech. rep., Office Fédéral de la Statistique, Neuchâtel. Statistique de la Suisse. (Cited on page 50)

Freedman, D. A. (2006). On the so-called "Hubert sandwich estimator" and "robust standard errors". *The American Statistician* **60**, 299–302. (Cited on page 132)

Fuller, W. A. (1987). *Measurement Error Models*. New York: John Wiley & Sons. (Cited on page 123)

Goga, C., Deville, J.-C. & Ruiz-Gazen, A. (2009). Use of functionals in linearization and composite estimation with application to two-sample survey data. *Biometrika* **96**, 691–709. (Cited on pages 74 and 131)

Graf, E. (2006). SILC, sample sizes requirements for Switzerland according to the rules defined by EUROSTAT for the European Union and rotational design. Internal document, Swiss Federal Statistical Office, Neuchâtel. (Cited on pages 27 and 35)

Graf, E. (2008a). Pondérations du SILC pilote SILC_I vague 2, SILC_II vague 1, SILC_I et SILC_II combinés. Rapport de méthodes, Office Fédéral de la Statistique, Neuchâtel. (Cited on pages 27, 28, 35, 50, 57, and 131)

Graf, E. (2008b). Pondérations du silc pilote SILC_I vague 2, SILC_II vague 1, SILC_I et SILC_II combinés. Methodology report, 338-0051, Swiss Federal Statistical Office, Neuchâtel, Switzerland. (Cited on page 115)

Graf, E. (2009a). Pondérations du Panel Suisse de Ménages. Rapport de méthodes, Office Fédéral de la Statistique, Neuchâtel. (Cited on pages 54 and 57)

Graf, E. (2009b). Weightings of the Swiss Household Panel. Methodology report, 338-0058, Swiss Federal Statistical Office, Neuchâtel. (Cited on page 115)

Graf, E. (2010). Enquête suisse sur la santé 2007. Plan d'échantillonnage pondérations et analyses pondérées des données. Rapport de méthodes, Office Fédéral de la Statistique, Neuchâtel. (Cited on pages 54 and 55)

Graf, E. (2014). Variance Estimation for Regression Imputed Quantiles and Inequality Indicators. Tech. rep. (Cited on page 132)

Graf, E. & Kilchmann, D. (2010). La formation spécialisée en statistique: Données manquantes et données éronnées. Document interne, Office Fédéral de la Statistique. Cours B4.2. (Cited on page 60)

Graf, E. & Qualité, L. (2014). Sondage dans un registre de population et de ménages en suisse: coordination d'échantillons, pondération et imputation. a paraître. *Journal de la Société Française de Statistique* **155**, 95–133. (Cited on pages 92 and 133)

Graf, E. & Tillé, Y. (2012). Imputations de données de revenu à l'aide de calage généralisé et de lois GB2: illustration sur les données silc suisses 2009. In *Colloque Francophone sur les Sondages*. Rennes. (Cited on pages 67, 69, and 70)

Graf, E. & Tillé, Y. (2014). Variance estimation through linearization for poverty and social exclusion indicators. *Survey Methodology* **40**, 61–79. (Cited on pages 92, 99, 101, 108, 115, and 131)

Graf, M. (2009c). An efficient algorithm for the computation of the gini coefficient of the generalized beta distribution of the second kind. In *JMS Proceedings*. American Statistical Association. (Cited on page 151)

Graf, M. (2013). A simplified approach to linerarization variance for surveys. Tech. rep., University of Neuchâtel, Neuchâtel. Unpublished. (Cited on page 75)

Graf, M. & Nedyalkova, D. (2011). *GB2: Generalized Beta Distribution of the Second Kind: properties, likelihood, estimation.* R package version 1.0. (Cited on pages 85, 121, 147, and 149)

Graf, M. & Nedyalkova, D. (2013). Modeling of income and indicators of poverty and social exclusion using the generalized beta distribution of the second kind. *Review of Income and Wealth* . (Cited on pages 121, 131, 147, and 149)

Hájek, J. (1981). *Sampling from a Finite Population*. New York: Marcel Dekker. (Cited on page 39)

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association* **69**, 383–393. (Cited on pages 73, 75, and 80)

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica* **46**, 1251–1271. (Cited on pages 70, 145, 152, and 153)

Hidiroglou, M. A. & Berthelot, J. M. (1986). Statistical editing and imputation for periodic business surveys. *Survey Methodology* **12**, 73–83. (Cited on page 64)

Horvitz, D. G. & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* **47**, 663–685. (Cited on pages 39 and 48)

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1. (Cited on page 132)

Hulliger, B. (1999). Simple and robust estimators for sampling. In *Survey Research Methods Section*. American Statistical Association. (Cited on pages 23, 24, 64, 82, 128, and 145)

Hyndman, R. J. & Fan, Y. (1996). Sample quantiles in statistical packages. *The American Statistician* **50**, 361–365. (Cited on pages 79, 80, and 101)

Jenkins, S. P. (2007). Inequality and the GB2 income distribution. *Discussion Paper IZA No. 2831* . (Cited on pages 70 and 119)

Jones, A., Lomas, J. & Rice, N. (2011). Applying beta-type size distributions to healthcare cost regressions. Tech. rep., University of York. Health, Econometrics and Data Group. (Cited on pages 70 and 119)

Kalton, G. & Kasprzyk, D. (1986). Le traitement des données d'enquête manquantes. *Techniques d'enquête* **12**, 1–17. (Cited on page 54)

Kass, G. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics* **29**, 119–127. (Cited on pages 52 and 133)

Keyfitz, N. (1951). Sampling with probabilities proportional to size: adjustmentfor changes in the probabilities. *Journal of the American Statistical Association* **46**, 105–109. (Cited on page 37)

Kish, L. & Scott, A. (1971). Retaining units after changing strata and probabilities. *Journal of the American Statistical Association* **66**, 461–470. (Cited on page 37)

Kleiber, C. & Kotz, S. (2003). *Statistical Size Distributions in Economics and Actuarial Sciences*. New York: Wiley. (Cited on pages 70, 119, 147, and 151)

Kott, P. S. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology* **32**, 133–142. (Cited on pages 55, 69, and 121)

Kovacevic, M. S. & Binder, D. A. (1997). Variance estimation for measures of income inequality and polarization - the estimating equations approach. *Journal of Official Statistics* **13**, 41–58. (Cited on pages 75 and 131)

Kröger, H., Särndal, C.-E. & Teikari, I. (1999). L'échantillonnage mixte de poisson : une famille de plans permettant une sélection coordonnée à l'aide de nombres aléatoires permanents. *Techniques d'enquête* **25**, 3–12. (Cited on page 37)

Langel, M. & Tillé, Y. (2011). Statistical inference for the quintile share ratio. *Journal of Statistical Planning and Inference* **141**, 2976–2985. (Cited on pages 78 and 79)

Langel, M. & Tillé, Y. (2013). Variance estimation of the Gini index: Revisiting a result several times published. *Journal of the Royal Statistical Society* **A176**, 521–540. (Cited on pages 78 and 131)

LaRoche, S. (2007). Pondérations longitudinale et transversale de l'Enquête sur la dynamique du travail et du revenu. Année de référence 2003. Tech. rep., Statistique Canada, Ottawa. (Cited on pages 52 and 53)

Lavallée, P. (2002). *Le sondage indirect ou la méthode généralisée du partage des poids*. Paris: Ellipses. (Cited on pages 48, 57, and 131)

Le Guennec, J. & Sautory, O. (2002). Calmar 2: Une nouvelle version de la macro calmar de redressement d'chantillon par calage. In *Journées de Méthodologie Statistique*. Paris: INSEE. (Cited on pages 55, 56, 69, and 121)

Lee, H., Rancourt, E. & Särndal, C.-E. (1994). Experiments with variance estimation from survey data with imputed values. *Journal of Official Statistics* **10**, 231–243. (Cited on pages 91, 92, and 115)

Lesage, E. & Haziza, D. (2013a). A discussion of weighting procedures for unit nonresponse. *Journal of Official Statistics* To appear. (Cited on pages 118, 121, 123, 126, 128, and 159)

Lesage, E. & Haziza, D. (2013b). On the problem of bias and variance amplification of the instrumental calibration estimator in the presence of unit nonresponse. *Journal of Survey Statistics and Methodology* To appear. (Cited on pages 118, 121, 123, 126, 128, and 159)

Lévesque, I. & Franklin, S. (2000). Pondérations longitudinale et transversale de l'enquête sur la dynamique du travail et du revenu. année de référence 1997. Tech. rep., Statistique Canada, Ottawa. (Cited on page 58)

Little, R. J. A. & Rubin, D. B. (2002). *Statistical Analysis With Missing Data*. New York: Wiley, 2nd ed. (Cited on pages 30, 65, 95, 119, and 144)

Lollivier, S. (2001). Endogénéité d'une variable explicative dichotomique dans le cadre d'un modèle probit bivarié. une application au lien entre fécondité et activité féminine. *Annales d'Economie et de Statistique* **62**, 251–269. (Cited on page 158)

Lorenz, M. O. (1905). Methods of measuring the concentration of wealth. *American Statistical Association* **9**, 209–219. (Cited on page 77)

Massiani, A. (2013a). Estimation de la variance d'indicateurs transversaux pour l'enquête SILC en Suisse. *Techniques d'enquêtes* **39**, 139–167. (Cited on pages 57, 68, 69, and 151)

Massiani, A. (2013b). Estimation of the variance of cross-sectional indicators for the SILC survey in Switzerland. *Survey Methodology* **39**, 121–148. (Cited on pages 28, 131, and 144)

McDonald, J. B. (1984). Some generalized functions for the size distribution of income. *Econometrica* **52**, 647–663. (Cited on pages 70, 121, 147, 149, and 150)

McDonald, J. B., Sorensen, J. & Turley, P. A. (2013). Skewness and kurtosis properties of income distribution models. *Review of Income and Wealth* **2**, 360–374. (Cited on pages 70 and 119)

McFadden, D. (1999). Chapter 4: Instrumental variables. Lecture: Eonometrics/Statistics. University of California, Berkeley. (Cited on page 123)

Merkouris, T. (2001). Estimation transversale dans le cas des enquêtes auprs des ménages à panels multiples. *Techniques d'enquête* **27**, 189–200. (Cited on page 58)

Müller, A. und Schoch, T. (2014). Vermögenslage der privaten haushalte. vermögensdefinitionen, datenlage und datenqualität. 20 wirtschaftliche und soziale situation der bevölkerung, Budesamt fr Statistik, Neuchâtel. (Cited on page 23)

Nakache, J.-P. & Confais, J. (2003). *Statistique explicative appliquée : analyse discriminante, modèle logistique, segmentation par arbre.* Paris: Editions Technip. (Cited on page 53)

OFS (2012). L'enquête suisse sur la population active dès 2010. concepts - bases méthodologiques - considérations pratiques. Encyclopédie statistique de la suisse, Office Fédéral de la Statistique, Neuchâtel. (Cited on pages 35 and 50)

Ohlsson, E. (1995). Coordination of samples using permanent random numbers. In *Business Survey Methods*, B. G. Cox, D. A. Binder, B. N. Chinnappa, A. Christianson, M. J. Colledge & P. S. Kott, eds. New York: Wiley. (Cited on page 37)

Osier, G. (2009). Variance estimation for complex indicators of poverty and inequality using linearization techniques. *Survey Research Methods* **3**, 167–195. (Cited on pages 19, 20, 24, 25, 69, 73, 75, 78, 79, 80, 81, 82, 92, 101, 115, and 131)

Osier, G. (2013). Dealing with non-ignorable non-response using generalised calibration: Simulation study based on the Luxemburgish household budget survey. *Economie et Statistiques: Working papers du STATEC* **65**, 1–25. (Cited on pages 69, 118, and 121)

Park, S. & Kim, J. K. (2014). Instrumental-variable calibration estimation in survey sampling. *Statistica Sinica* **24**, 1001–1015. (Cited on page 121)

Patterson, H. D. (1950). Sampling on successive occasions with partial replacement of units. *Journal of the Royal Statistical Society* **B12**, 241–255. (Cited on page 37)

Pfefferman, D. & Sverchkov, M. Y. (2003). *Fitting generalized linear model under informative sampling*, chap. 12. New York: Wiley, pp. 175–195. Skinner, C. and Chambers, R., editors. (Cited on page 132)

Qualité, L. (2009). *Unequal probability sampling and repeated surveys*. Thèse de doctorat, Université de Neuchâtel, Neuchâtel, Suisse. (Cited on pages 40 and 42)

Quenouille, M. H. (1949). Approximation tests of correlation in time series. *Journal of the Royal Statistical Society* **B11**, 18–84. (Cited on page 68)

Raghunathan, T. E., Lepkowski, J. M., van Hoewyk, J. & Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology* **27**, 85–95. (Cited on pages 24, 66, 92, 117, 118, and 145)

Rakotomalala, R. (2005). Arbres de décision. *Revue MODULAD* **33**, 163–187. (Cited on page 53)

Renfer, J.-P. (2009). Etude de la précision des estimateurs de l'enquête continue suisse sur la population active selon le choix du schéma de rotation des échantillons. In *Journées de Méthodologie Statistique*. Paris: INSEE. (Cited on pages 35, 50, and 58)

Rivière, P. (1998). Description of the chosen method: Deliverable 2 of supcom 1996 project (part "co-ordination of samples"). Report, EUROSTAT. (Cited on page 37)

Rivière, P. (1999). Coordination of samples: the microstrata methodology. In *13th International Roundtable on Business Survey Frames*. Paris: INSEE. (Cited on page 37)

Rivière, P. (2001a). Coordinating samples using the microstrata methodology. In *Proceedings of Statistics Canada Symposium*. (Cited on page 37)

Rivière, P. (2001b). Random permutations of random vectors as a way of co-ordinating samples. Tech. rep., University of Southampton, Southampton. (Cited on page 37)

Rosén, B. (1997a). Asymptotic theory for order sampling. *Journal of Statistical Planning and Inference* **62**, 135–158. (Cited on page 37)

Rosén, B. (1997b). On sampling with probability proportional to size. *Journal of Statistical Planning and Inference* **62**, 159–191. (Cited on page 37)

Rubin, D. B. (1987). *Multiple Imputation for Nonreponse in Surveys*. New York: Wiley. (Cited on pages 132 and 144)

Ruud, P. A. (1984). Tests of specification in econometrics. *Econometric Reviews* **3**, 211–242. (Cited on pages 70 and 152)

Sande, I. G. (1981). Imputations in surveys: Coping with reality. *Survey Methodology* **7**, 21–43. (Cited on page 66)

Särndal, C.-E. & Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. New York: Wiley. (Cited on pages 54 and 55)

Särndal, C.-E., Swensson, B. & Wretman, J. H. (1992). *Model Assisted Survey Sampling*. New York: Springer. (Cited on page 68)

Sautory, O. (2003). Calmar 2: A new version of the calmar calibration adjustment program. In *Proceedings of Statistics Canada Symposium*. (Cited on pages 55, 69, and 121)

Sautory, O. & Le Guennec, J. (2005). La macro calmar2: redressement d'un échantillon par calage sur marges. Tech. rep., INSEE. (Cited on page 127)

Schreiber, S. (2008). The hausman test statistic can be negative even asymptotically. *Journal of Economics and Statistics* **4**, 394–405. (Cited on page 157)

Sepanski, J. H. & Kong, J. (2008). A family of generalized beta distributions for income. *Advances and Applications in Statistics* **10**, 75–84. (Cited on pages 70 and 119)

SFSO (2014). Swiss pendent of the European Union statistics on Income and Living Conditions (EU-SILC). (Cited on page 106)

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall. (Cited on pages 82, 83, and 84)

StatCan (2010). *BOOTVAR, Guide de l'usager*. Statistique Canada. Bootvar 3.2, version SAS. (Cited on page 68)

Stock, J. H. & Trebbi, F. (2003). Who invented instrumental variable regression? *Journal of Economic Perspectives* **17**, 177–194. (Cited on page 123)

Strand, M. M. (1979). Estimation of a population total under a "Bernoulli sampling" procedure. *The American Statistician* **33**, 81–84. (Cited on page 39)

Tambay, J. L., Schiopu-Kratina, I., Mayda, J., Stukel, D. & Nadon, S. (1998). Traitement de la non-réponse du cycle de deux enquêtes sur la santé de la population. *Techniques d'enquête* **24**, 159–169. (Cited on page 54)

Theil, H. (1953). Repeated least square applied to complete equation systems. Tech. rep., Central Planning Bureau, The Hague. (Cited on page 69)

Tillé, Y. (2006). *Sampling Algorithms*. New York: Springer. (Cited on pages 39 and 48)

Van Huis, L. T., Koeijers, C. A. J. & de Ree, S. J. M. (1994a). EDS, sampling system for the central business register at statistics netherlands. Tech. rep., Statistics Netherlands, The Hague. (Cited on page 37)

Van Huis, L. T., Koeijers, C. A. J. & de Ree S. J. M. (1994b). Response burden and co-ordinated sampling for economic surveys. Tech. rep., Statistics Netherlands, The Hague. (Cited on page 37)

Verma, V. & Betti, G. (2011). Taylor linearization sampling errors and design effects for poverty measures and other complex statistics. *Journal of Applied Statistics* **38**, 1549–1576. (Cited on pages 74, 82, and 131)

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity (STMA V22 942). *Econometrica* **48**, 817–838. (Cited on page 132)

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50**, 1–26. (Cited on page 132)

Wright, P. G. (1928). *The Tariff on Animal and Vegetable Oils*. New York: Macmillan. (Cited on page 123)

Wu, D.-M. (1973). Alternative tests of independence between stochastic regressors and distrubances: Finite sample results. *Econometrica* **42**, 529–546. (Cited on pages 70, 145, and 152)

Zeileis, A. (2012). *ineq: Measuring Inequality, Concentration, and Poverty*. R package version 0.2-10. (Cited on page 85)

Notations and abbreviations

The most used abbreviations in the text are listed below, followed by some of the notations used. More detailed notations are introduced, when needed, in each chapter.

ABBREVIATIONS

AMELI	Advanced Methodology for European Laeken (EU-Project);
ARPR	At Risk of Poverty Rate;
ARPT	At Risk of Poverty Threshold;
BIT	Bureau International du Travail;
CATI	Computer Assisted Telephone Interview;
Castem	Cadre de sondage pour le tirage d'échantillons de ménages;
CCO	Central Compensation Office;
COICOP	Classification Of Individual COnsumption by Purpose;
DACSEIS	DAta Quality in Complex Surveys within the New European Information Society (EU-Project);
DPP	Data Preparation Process;
EBM	Enquête sur le Budget des Ménages;
EDIMBUS	Recommended Practices for EDiting and IMputation in Cross-Sectional BUsiness Surveys (EU-Project);
EUREDIT	The development and evaluation of new methods for editing and imputation (EU-Project);
Espa	Enquête Suisse sur la Population Active;
GB2	Generalized Beta distribution of the second kind;
GINI	Gini index;
GHR	Groupes Homogènes de Réponse;
INS	Instituts Nationaux de Statistique;
IVEware	Imputation and Variance Estimation softWARE;
MAD	Median Absolute Deviation;
MAR	Missing At Random;
MCAR	Missing Completely at Random;
MEDP	Median of the poor (units below ARPT);
NMAR	Not Missing At Random;
PPSD	Processus de Préparation Statistique des Données;
OFS	Office Fédéral de la Statistique;

OSM	Original Sample Member;
QSR	Quintile Share Ratio;
RHG	Response Homogeneity Groups;
RMPG	Relative Median at-risk-of Poverty Gap;
SFSO	Swiss Federal Statistical Office;
SILC	Survey on Income and Living Conditions;
SRPH	Stichprobenrahmen für Personen und Haushaltserhebungen;
SRS	Simple Random Sampling;
Statpop	statistique de la population et des ménages;
Symic	Système d'information central sur la migration;
WIV	Weighted Instrumental Variable (regression or parameters);
WLS	Weighted Least Squares (regression or parameters);

NOTATIONS

$U = \{1, \ldots, N\}$	Population of size N;
$s \subset U$	Sample of size n;
$r \subset s$	Respondent sample of size n_r for variable of interest y;
$I_k(s)$	Indicator variable of unit k in sample s;
$p(\cdot)$	Sampling design;
$q(\cdot\|s)$	Nonresponse mechanism from s to r;
$\pi_k = \mathrm{E}(I_k)$	First order inclusion probabilities, $\mathrm{E}(\cdot)$ is the expectation under $p(\cdot)$;
$\pi_{k\ell} = \mathrm{E}(I_k I_\ell)$	Second order inclusion probabilities;
$\Delta_{k\ell} = \pi_{k\ell} - \pi_k \pi_\ell$	Covariance of I_k and I_ℓ;
$\widehat{Y}_{HT}$, $\widehat{Y}$	Horvitz-Thompson estimator;
Q_α	the α^{th} quantile;
$\mathbf{x}_{k\cdot} = (x_{k1}, ..., x_{kj}, ..., x_{kJ})$	Row vector of auxiliary variables;
$\mathbf{z}_{k\cdot} = (z_{k1}, ..., z_{kj}, ..., z_{kJ})$	Row vector of instrumental variables;
$\mathbf{X}_U$, $\mathbf{X}_s$, $\mathbf{X}_r$	Matrices auxiliary variables, of sizes $N \times J$, $n \times J$, respectively $n_r \times J$;
$\mathbf{Z}_r$	Matrix of instrumental variables, of size $n_r \times J$;
y	Variable of interest, of income nature.

This document has been prepared using the text editor Texmaker and typesetting system LaTeX 2_ε.