universität wien

# MAGISTERARBEIT

Titel der Magisterarbeit

„Weighting procedures to counter unit nonresponse bias of estimators for sample surveys”

Verfasser

Mag. rer. soc. oec. Thomas Glaser

angestrebter akademischer Grad

Magister der Sozial- und Wirtschaftswissenschaften (Mag. rer. soc. oec.)

Wien, im September 2012

| | |
|---|---|
| Studienkennzahl lt. Studienblatt | A 066 951 |
| Studienrichtung lt. Studienblatt | Magisterstudium Statistik |
| Betreuer | Ao. Univ.-Prof. Mag. Dr. Andreas Futschik |

# Abstract

The main topic of the present thesis is bias of estimators from sample surveys with voluntary participation that may occur because of unit nonresponse. If not all units selected in the sample take part in the survey, this may lead to bias in estimators if the event of unit nonresponse does not happen completely at random. Weighting methods can be used to compensate for this bias. Therefore, suitable methods like using the inverse of estimated response probabilities to adjust for unit nonresponse or calibration to external marginal distributions are presented.

In order to investigate potential bias due to unit nonresponse, simulations were carried out to make a comparison of different estimators applying various weighting methods that presumably counter bias. The data for these simulations were taken from the survey EU-SILC (European Union Statistics on Income and Living Conditions) in Austria. EU-SILC is a probability sample survey of private households in Austria that employs a rotating panel design. Participation in the survey is voluntary. The survey variable of interest in this thesis is a component of the household income (income from employment and old-age benefits). The sum of this variable over all households in the population is the parameter that was chosen to be estimated from the respondents of the selected sample. The response process was simulated for the selected sample of the first wave of EU-SILC 2010 and the expectation of different estimators calculated from the simulated response sets was compared to the design-weighted estimate of the selected sample. Results show that no adjustment for unit nonresponse leads to a biased estimator. Weighting methods that use auxiliary information to adjust for unit nonresponse bias can reduce this bias. The most precise estimators result from calibration, because this method reduces bias and also facilitates a reduction of variance.

# Abstract in German

Das Hauptthema der vorliegenden Arbeit sind Verzerrungen von Schätzern aus Stichprobenerhebungen mit freiwilliger Teilnahme. Antwortausfälle aufgrund von nicht an der Erhebung teilnehmenden Erhebungseinheiten können zu verzerrten Schätzern führen, wenn der Ausfall nicht komplett zufällig stattfindet. Verschiedene Gewichtungsmethoden können verwendet werden, um dieser Verzerrung entgegenzuwirken. Dazu zählen beispielsweise die Verwendung der Inversen der geschätzten Antwortwahrscheinlichkeiten oder Kalibrierung an externe Randverteilungen.

Um mögliche Verzerrungen aufgrund von Antwortausfällen zu untersuchen, wurden Simulationen durchgeführt, die einen Vergleich der verschiedenen Schätzer und der verwendeten Gewichtungsmethoden, die vermutlich Verzerrungen entgegenwirken, erlauben. Die Datengrundlage der Simulationen bildet die Erhebung EU-SILC (European Union Statistics on Income and Living Conditions) in Österreich. EU-SILC ist eine Erhebung mit freiwilliger Teilnahme basierend auf einer Wahrscheinlichkeitsstichprobe privater Haushalte in Österreich und einem rotierenden Paneldesign. Die wichtigste Untersuchungsvariable für die vorliegende Arbeit ist ein Bestandteil des gesamten Haushaltseinkommens (Einkommen aus unselbständiger Tätigkeit und Alterspensionen). Die Summe dieser Variable über alle Haushalte in der Population wurde als zu schätzender Parameter gewählt. Der Antwortprozess wurde für die gezogene Stichprobe der Erstbefragung EU-SILC 2010 simuliert und der Erwartungswert der verschiedenen Schätzungen, basierend auf den teilnehmenden Haushalten, wurde mit der designgewichteten Schätzung der gezogenen Stichprobe verglichen. Die Ergebnisse zeigen, dass ohne eine Anpassung der Gewichte an den Antwortausfall der Schätzer für den Wert in der Grundgesamtheit verzerrt ist. Gewichtungsmethoden, die Hilfsvariablen zur Kompensation des Antwortausfalls verwenden, können Verzerrungen verkleinern. Kalibrierungen an externe Randverteilungen resultierten schließlich in Schätzern mit der höchsten Präzision, da sie nicht nur vergleichsweise kleine Verzerrungen aufweisen, sondern auch zu einer verringerten Varianz führen.

# Contents

# List of Tables

# List of Figures

# 1 Introduction

## 1.1 Motivation

During the last years I have been working at Statistics Austria with surveys that rely on probability samples, for example EU-SILC (European Union Statistics on Income and Living Conditions). EU-SILC is the main source of data about household income in Austria. It is also used for estimating EU-indicators on social inclusion (e.g. the at-risk-of-poverty rate). EU-SILC is a probability sample survey with voluntary participation and a rotating design of four rotations. During my work on EU-SILC topics like sampling and weighting are of importance. Therefore over-/undercoverage, sampling error and nonresponse are issues with which I have been confronted on a regular basis. In order to make inference on the population, to counter potential bias due to unit nonresponse and to guarantee comparability with statistics from other sources, EU-SILC facilitates a sophisticated weighting procedure which is composed of three main steps: (1) Design weights (inverse of selection probabilities), (2) nonresponse weights (inverse of estimated response probabilities) and (3) calibration on known external marginal distributions.[1]

The first part of the weighting procedure is straightforward. Since EU-SILC is a probability sample of addresses of private household as sampling units drawn from the central residence register (Zentrales Melderegister – ZMR), the probabilities of selection for all units belonging to the sampling frame are known prior to sample selection. Calibration, the third part of the weighting process is applied with CALMAR ("CALage sur MARges")[2], a macro for the statistical software SAS developed by the national statistical institute of France (INSEE). However, the second part of the weighting procedure poses certain challenges, because the probabilities of taking part in the survey have to be estimated for those who are willing to respond as well as for those who do not re-

[1] Cf. Glaser & Till (2010).
[2] Cf. Sautory (1993).

spond. Since only a limited amount of data is available from the sampling frame before the actual interview of the first year wave of a rotational subsample, it is very difficult to estimate the response probabilities. Another challenge is that nonresponse is usually highest in the first year wave of the survey. For example, in EU-SILC 2010 the response rate amounted to about 62% for the first wave rotation, yielding a nonresponse rate of 38% (Statistics Austria 2011, p. 26). A logistic regression model was used to estimate response probabilities, but only variables from the sampling frame and the type of building from the gross sample could be used as predictors. The pseudo r-square of the model was smaller than 5%. This could result from a nonresponse mechanism that acts almost completely at random, or, is due to insufficient auxiliary data available for the first year of the survey. For the second, third and fourth year of the survey information from the past years can be used to estimate response probabilities. Therefore the corresponding models to estimate the response probabilities have much more explanatory variables at hand compared to the model for the first year. The goodness of fit of the corresponding logistic regression models is higher for subsequent survey years (pseudo r-square is about 20%). The response rate for follow-up rotations ranged from 83% to 90% in EU-SILC 2010.[3]

## 1.2 Main topic and research questions

A comparatively high nonresponse rate may be the source of some major bias of estimators which are used to calculate estimates based on survey data. A survey can be defined as "a systematic method for gathering information from (a sample of) entities for the purposes of constructing quantitative descriptors of the attributes of the larger population of which the entities are members" (Groves et al. 2004, p. 4). In this thesis the term "survey" will always refer to "sample survey" and thus describing the process of drawing a sample from a (big) population and subsequently collecting information only from the elements in the sample. The data gathered from the sample can then be used to estimate parameters of certain quantities of the population.

Estimators from sample surveys encounter many kinds of errors that are usually described with the concept of "total survey error" (Groves et al. 2004, p. 49) which incorporates all sources of a deviation of the parameter of interest from its estimator based on survey data. Of these potential errors of the survey process only "coverage

[3]Cf. Statistics Austria (2010).

error, sampling error and nonresponse error" (Särndal & Lundström 2005, p. 6) will be of interest in this thesis, where the main focus will lie on the last one. The first error mentioned above refers to the error that occurs if not all elements of the population are eligible in the sampling frame, or if some elements in the sampling frame do not belong to the population. Sampling error occurs, because instead of the whole population, only a sample is used for data collection. If not all units selected in the sample take part in the survey, error due to this so-called "unit nonresponse" (Groves et al. 2004, p. 45) arises. If the above mentioned errors occur, they may increase the variance and/or distort the estimators that are based on the survey data. The latter case is usually called "bias" (Groves et al. 2004, p. 45). The bias of a parameter estimator is defined as the deviation of that estimator from the true parameter in the population.[4] Besides the error due to (an increased) variance there are three types of biases that correspond to the errors mentioned at the beginning of this section: coverage bias, sampling bias and nonresponse bias (Groves et al. 2004, ch. 2.3). Coverage bias refers to the difference of a parameter based on the sampling frame to the parameter in the whole survey population. Sampling bias occurs if certain elements of the sampling frame have a zero probability of selection or a probability of selection that is not or incorrectly represented by design weights. Bias due to unit nonresponse arises, if estimators based on the respondents differ from those based on the whole sample. Nonresponse bias has to be assumed as one of the biggest challenges related to estimators based on sample surveys, because often there is not enough information at hand to counter unit nonresponse in an effective way. Bias due to nonresponse may considerably distort estimators. Furthermore, if unit nonresponse introduces bias to an estimator that is sample-unbiased, not only the point estimate becomes biased, but also the corresponding confidence interval of the estimator will be rendered invalid.[5] Särndal and Lunström state that "it is the nonresponse bias that constitutes by far the most important obstacle to correct statistical conclusions in a survey" (Särndal & Lundström 2005, p. 14). So the increase in variance due to unit nonresponse can be considered a minor obstacle as long as unbiasedness can be (approximately) maintained although unit nonresponse occurs. As described in the previous section, sample surveys which rely on voluntary participation, e.g. EU-SILC, will almost certainly encounter some form of unit nonresponse. Although incentives to encourage participation (e.g. financial recompensation or vouchers) may lead to a higher response rate, in most cases there will be a percentage of units selected

[4]Cf. Särndal et al. (2003), p. 40.
[5]Cf. Särndal (2007), p. 144.

in the sample who are not willing to participate. If this selection process does not occur completely at random, the findings of the survey may be biased for certain variables of interest. If, for example, a measure of income inequality as the gini coefficient is estimated with data from a survey, a bias occurs if people belonging to the highest income decile are underrepresented, because many of those persons are difficult to reach or are not willing to take part in the survey. In order to counter this bias caused by unit nonresponse, weighting the respondents with the inverse of their estimated (group-)response probability is a suitable method.[6] If it is possible to estimate the response probabilities of certain groups of selected sampling units, the remaining respondents of that group may compensate for the loss because of unit nonresponse by applying a weighting factor that is the inverse of the estimated response probability. As mentioned above it is usually quite difficult to correctly estimate the unknown response probability of each selected sampling unit. Usually, only a limited amount of information about the selected sampling unit is known before the time of the interview. In this case the usage of register information in addition to characteristics of the sampling frame may considerably ameliorate the scope of possible explanatory variables for estimating the response probability. Another challenge is the actual method for countering unit nonresponse. Using the inverse of the estimated response probabilities (e.g. estimated by a logistic regression model) may be a useful tool, but other methods like calibration to known external marginal distributions have to be taken into account too. Although weighting may be a suitable method for countering bias, another important issue is the enlargement of the variance of weighted estimates. The ideal solution is a method that reduces bias and at the same time does not enlarge the variance of weights too much.

The main topic of this thesis is bias that may exist for estimators which are based on data from sample surveys in which unit nonresponse occurs. The main assumption is that the major source of bias in such surveys is nonresponse that does not occur at random in the case of the EU-SILC survey. That means the probability to take part in the survey differs for all sampled elements in a way that has a distorting influence on the estimators of parametes of interest, based on survey variables.

[6] This method is described as appropriate for the weighting procedure for EU-SILC, cf. Osier et al. (2006).

The main research questions which shall be dealt with in this thesis using data from the Austrian EU-SILC survey can be summarized as follows:

- Is there a considerable amount of bias due to unit nonresponse that has to be dealt with?
- Which methods to adjust estimators in order to counter unit nonresponse exist and how can they be improved?
- May a reduction of bias have undesirable consequences, for example a considerable enlargement of variance?

These three questions are the foundation of the following chapters and will be answered in this thesis. Chapter 2 introduces the topic of estimation of a certain parameter of interest when only a sample of the concerning population is at hand. Also the most important terms and definitions of the present thesis, e.g. "bias" or "nonresponse", will be explained in this chapter. Chapter 3 follows with a presentation of different weighting methods that result in unbiased estimators or estimators with can reduce bias due to unit nonresponse, if it occurs. Chapter 2 and chapter 3 provide the theoretical foundation of the practical application of methods to adjust estimators for unit nonresponse in chapter 4. These simulations with data from the Austrian EU-SILC survey are used to to evaluate the methods presented in chapter 3. Finally, chapter 5 sums up the findings of the previous chapters and gives an outlook on possible future applications for sample survey.

# 2 Estimation and Bias

## 2.1 Estimation of the parameter of interest based on sample data

Many studies, for instance EU-SILC, are carried out in order to analyze characteristics of large populations, e.g. all private households in Austria and the persons who have their main residence at the address of the household. Since it would be very expensive to collect information about, for example, income from employment for the whole population, only a fraction of the population is interviewed. If the selection of such a sample is the result of a random experiment, there exists statistical theory (Lehmann & Caselly 1998, ch.1) with which inference on the whole population can be carried out. The variable of interest $y_k$ in this thesis is the sum of income from employment and from old-age benefits at household level of households $k$ in the population $U$ of interest, which consists of all private households in Austria. Let $\theta(y_k)_{k\in U}$ denote the sum of income of all private households $k$ belonging to the population $U$. The size of the population shall be referred to as $N$ and the index $k$, $(k = 1, ..., N)$ identifies a specific element, i.e. a private household in the population.

$$\theta = \theta(y_k) = \sum_{k\in U} y_k \tag{2.1}$$

$\theta = \theta(y_k)$ is the parameter that has to be estimated, sometimes also called the "estimand" (Lehmann & Caselly 1998, ch. 1.1). Since it would be extremely cumbersome to gather income information from the whole population, a sample $s$ is drawn from the population $U$ in order to estimate the population total shown in (2.1).

Let $S$ denote a random variable which refers to the random experiment of drawing a sample $s$ from the set of all possible samples.[1] The probability $\pi_k$ that an element $k$ of

[1]The nomenclature in this section is inspired by Särndal et al. (2003).

the population is included in the sample is defined as follows (Särndal et al. 2003, ch. 2.4):

$$\pi_k = P(k \in s) = \sum_{s \ni k} p(s) \tag{2.2}$$

where $p(s) = P(S = s)$ denotes the probability of selecting a specific sample $s$ and the sum on the right hand side of (2.2) is over all samples $s$ that contain the element $k$. If the probabilities of selection $\pi_k$ are known and $\pi_k > 0$ holds for all $k \in U$ then a so-called "probability sample" (Särndal et al. 2003, p. 32) can be drawn by conducting a random experiment for all elements $k$ in the population. The resulting sample $s$ can then be used to collect information about the variable $y$ of interest for all elements $k \in s$. For the sake of convenience, from now on "$s$" will be used to describe the random variable $S$ as well as the realized sample $s$. This simplification is also justified by statistical literature.[2] The re-identification of an element $k$ of the population within a specific sample $s$ can easily be done with the "sample membership indicator" (Särndal et al. 2003, p. 30), here called $T_k$, which is defined as follows:

$$T_k = \begin{cases} 1 & \text{if } k \in s \\ 0 & \text{if } k \notin s \end{cases} \tag{2.3}$$

$T_k$ in (2.3) is a random variable, because it results from the random experiment of drawing a sample. Obviously, the expectation of the sample membership indicator is the probability of selection, because

$$E(T_k) = 1 \cdot \pi_k = \pi_k \tag{2.4}$$

If the size $n$ of the sample $s$ is much smaller than the size of the population ($n << N$), it is obviously much more convenient to use the sample as a source of gathering data on the variable of interest instead of the whole population. However, since only a few randomly chosen elements of the population now remain in the sample, the parameter $\theta$ has to be estimated by using the values $y_k$ given in the sample. The "estimator" of $\theta$ is called $\hat{\theta}$ and depends on the random experiment of drawing a sample, so $\hat{\theta} = \hat{\theta}(s)$.[3] Since only a sample is used to estimate the parameter of interest, the outcome of the estimation process, the "estimate", can differ from the value obtained from the whole

[2] Cf. Särndal et al. (2003), p. 41.
[3] Cf. Särndal et al. (2003), p. 39.

population. This difference $\hat{\theta} - \theta$ is called the "sampling error" (Särndal & Lundström 2005, p. 45). Ideally, this deviation should only stem from the randomness of the process of selecting a sample. So, on average, the parameter and its estimator should be the same. To evaluate this, the expectation of the estimator has to be compared with the true value of the parameter in the population. The expectation of an estimator is the sum over all possible samples, weighted by the probabilities of selection of all possible samples.[4] If the expectation of a parameter and the true value of the parameter in the population are the same, then the estimator is called "unbiased" (Särndal et al. 2003, p. 40):

$$B_T(\hat{\theta}) = E(\hat{\theta}) - \theta = 0 \tag{2.5}$$

where $B_T(\hat{\theta})$ denotes the "bias". More specifically, the bias presented in (2.5) is the bias due to sampling which is specified by the suffix "T" in order to refer to the sample membership indicator in (2.3). It can also be written as the expectation of the sampling error:

$$B_T(\hat{\theta}) = E(\hat{\theta} - \theta) = E(\hat{\theta}) - \theta \tag{2.6}$$

The parameter of interest in this thesis is the population total of income from employment or old-age benefits of private households. A well-known estimator for the total in (2.1) is the so-called "Horvitz-Thompson estimator", which was introduced by Horvitz & Thompson (1952).

$$\hat{\theta} = \sum_{k \in s} \frac{y_k}{\pi_k} \tag{2.7}$$

This estimator is unbiased for the total, because

$$\begin{aligned} E(\hat{\theta}) &= E(\sum_{k \in s} \frac{y_k}{\pi_k}) = E(\sum_{k \in U} T_k \frac{y_k}{\pi_k}) = \\ &= \sum_{k \in U} \pi_k \frac{y_k}{\pi_k} = \sum_{k \in U} y_k = \\ &= \theta \end{aligned} \tag{2.8}$$

[4] Cf. Särndal et al. (2003), p. 40.

The factor $1/\pi_k$ is also called the "design weight" (Särndal & Lundström 2005, p. 30). It inflates the values of the variables in the sample to the size of the population. This approach to weighting data from probability samples is also called "design-based" or based on "randomization theory" (Särndal & Lundström 2005, p. 49). It applies perfectly for data from probability samples where no data are missing, i.e. there is full response.

The calculation of the Horvitz-Thompson estimator depends on the kind of sample selection given by the distribution of $p(s)$. The survey of interest in this thesis is EU-SILC in Austria, which uses a stratified simple random sample without replacement as a "sampling design" (Särndal et al. 2003, ch. 2.3). A stratified simple random sample with $D$ strata may also be viewed as a collection of $D$ independent simple random samples. Therefore it is important to know how the probability of inclusion in a specific sample for element $k$ is calculated for a simple random sample without replacement.[5] If $n$ denotes the fixed size of the sample and $N$ the size of the population, then the number of possible samples is a combination without replacement and given by the binomial coefficient of $\begin{pmatrix} N \\ n \end{pmatrix}$. The number of possible samples that include a specific element $k$ of the population is given by $\begin{pmatrix} N-1 \\ n-1 \end{pmatrix}$, because if $k$ is fixed, the size of the remaining population is *N-1* and the size of the remaining samples is *n-1*. Applying Laplace's definition of probability[6], i.e. (number of favorable events)/(number of possible events) it is now easy to see, that the inclusion probability $\pi_k$ simplifies to the following expression (Särndal et al. 2003):

$$\pi_k = \sum_{s \ni k} p(s) = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N} \tag{2.9}$$

The quantity $n/N$ is often called the "sampling fraction" (Cochran 1977, p, 21). So in simple random sampling the probabilities of selection are the same for all elements. As mentioned above, a stratified simple random sample $s$ consists of $D$ disjunct strata drawn from the disjunct partitions $U_d$ $(d = 1, ..., D)$ of $U$ with size $N_d$. For each stratum $d$ independent samples $s_d$ of size $n_d$ are drawn. The sample sizes of the strata sum up to the size of the sample and the sizes of the strata sum up to the size of the population, i.e.

[5] For a detailed description of simple random sampling and stratified sampling please refer to Särndal et al. (2003), ch. 3.3. and ch. 3.7. and to Cochran (1977), ch. 2 and ch. 5.

[6] Compare Pfanzagl (1991), p. 29ff.

$$n = \sum_{d=1}^{D} n_d \text{ and } N = \sum_{d=1}^{D} N_d \tag{2.10}$$

So within every stratum, the probability of selection is defined as $\pi_{kd} = n_d/N_d$. The Horvitz-Thompson estimator for the population total presented above (2.7) and applied to a stratified simple random sample can easily be calculated as the sum of the stratum-estimators over all strata. In other words, for each partition $U_d$ of $U$ the total $\theta_d$ has to be estimated separately and the total for the whole population $U$ is $\theta = \sum_{d=1}^{D} \theta_d$. Let $T_k^{(d)}$ be the sample membership indicator for each stratum with the selection probability within a specific stratum as its expectation, i.e. $E(T_k^{(d)}) = \pi_{kd} = n_d/N_d$ (cf. formula (2.4)). Now it follows, that the Horvitz-Thompson estimator for stratified simple random sampling is unbiased for the total of the variable $y_k$ in the population $U$:

$$\begin{aligned}
E(\hat{\theta}_{STSI}) &= E(\sum_{d=1}^{D} \sum_{k \in s_d} \frac{y_k}{\pi_k}) = \\
&= E(\sum_{d=1}^{D} \sum_{k \in s_d} N_d \frac{y_k}{n_d}) = \\
&= \sum_{d=1}^{D} E(\sum_{k \in U_d} T_k^{(d)} N_d \frac{y_k}{n_d}) = \\
&= \sum_{d=1}^{D} \sum_{k \in U_d} \frac{n_d}{N_d} N_d \frac{y_k}{n_d} = \\
&= \sum_{d=1}^{D} \theta_d = \\
&= \theta
\end{aligned} \tag{2.11}$$

It is important to note that $\sum_{d=1}^{D} \theta_d = \theta$ holds because the strata are disjunct. The suffix "STSI" denotes the usage of stratified simple random sampling.[7]

[7] Cf. Särndal et al. (2003), p. 103.

## 2.2 Unit Nonresponse - An Overview

In surveys with voluntary participation there will almost certainly be a part of the selected units who cannot or will not take part in the survey. This source of missing data in a sample survey is called "unit nonresponse" (Groves et al. 2004). The following descriptions of the response status of units selected in a sample all refer to cases which belong to the population of interest and can be found in the sampling frame. In other words, under- and overcoverage are assumed to be non-existent in order to fully concentrate on missing data due to unit nonresponse.
In the following text let $y$ denote the survey variable of interest that is known only for the "unit response set" (Särndal et al. 2003, p. 560) which shall be referred to by $r$. Furthermore, let $\boldsymbol{x}$ denote a vector of auxiliary variables which are known for the whole sample. The unit response set contains all elements of the sample which are responding. Let $R_k$ denote the "response indicator" which identifies responding elements $k$ of the sample:[8]

$$R_k = \begin{cases} 1 & \text{if } k \in r \\ 0 & \text{if } k \notin r \end{cases} \tag{2.12}$$

Missing data due to unit nonresponse does not have to be a problem for unbiased estimation intrinsically. Groves states in this context that "low response rate surveys are not necessarily "bad" per se, but may yield some statistics subject to large nonresponse bias" (Groves 2006, p. 649). Therefore what is of concern here is the question whether estimation of $y$ for the unit response set $r$ differs from that based on the unit nonresponse set $q$ and furthermore how this difference may be explained.

In the literature on survey sampling and related issues two approaches of modeling the response process can be distinguished. Cochran (1977) divides the elements of the sample into two disjunct strata, where one contains all respondents and the other contains all nonrespondents. In one stratum all elements respond with probability one, in the other stratum the response probability is zero for all elements. Because this model is overly simple[9] and will not adequately mirror reality[10], most of the more recent works

[8] Little (1988) defines a "binary nonresponse indicator $r$". However, in order to resemble the sample membership indicator presented in ch. 2.1 a binary response indicator is used in the present thesis.
[9] By the way, this is even admitted by Cochran (1977) on p. 360.
[10] Särndal & Lundström (2005) refer to this approach as "simplistic and unrealistic" (Särndal & Lundström 2005, p. 49).

about nonresponse like Bethlehem (1999), Longford (2005) or Särndal & Lundström (2005) model the response process with the so-called "Random Response Model" (cf. Bethlehem (1999), p. 128). The probability that an element $k$ which has been selected in the sample $s$ responds to the survey shall be denoted $P(R_k = 1) = \delta_k$ in the present thesis.[11] It is the expected value of $R_k$ conditional on the selected sample $s$ (Särndal & Lundström 2005, p. 50), so $E(R_k \mid k \in s) = \delta_k$. This expectation is calculated according to the probability distribution of all possible unit response sets $d(r \mid s)$. Särndal & Lundström (2005) call $d(r \mid s)$ the "response mechanism". Since the realization of a specific unit response set depends on a random selection, the response process can be viewed as a subsampling phase, given $s$, where $\delta_k$ is the probability that a specific element $k$ is part of one of all possible unit response sets containing $k$:

$$\delta_k = P(k \in r \wedge k \in s) = \sum_{r \ni k} d(r \mid s) \tag{2.13}$$

Selecting $s$ and $r$ with probabilities $\pi_k$ and $\delta_k$ may be viewed as "two-phase sampling" (Särndal et al. 2003, ch. 9). Although, auxiliary information from a vector[12] of auxiliary variables $\boldsymbol{x}$ known for all the elements of the first sampling phase is not used for selecting the subsample, i.e. the unit response set, such auxiliary information may be related to the response probabilities $\delta_k$. Also the survey variable of interest $y$ may be related to the response probability. If the response probabilities were known for all elements in the sample prior to the event of response, then it should be easy to perform estimation of a population parameter just as in (2.8). Since the $\delta_k$ are unknown some assumptions about the response mechanism have to be made. Based on descriptions by Little & Rubin (1987) which are also mentioned in Bethlehem (1999) and Longford (2005), three different mechanisms for missing data due to nonresponse can be distinguished (for the sake of simpler presentation, the condition that $k \in s$ is omitted in the formulas below, because it should be obvious that an element $k$ can only respond if it belongs to a selected sample $s$):[13]

1. The response probability does not depend on the unknown survey variable $y$ or the known auxiliary data $\boldsymbol{x}$. The $\delta_k$ are the same for all $k$ and the response

[11] This is the same Greek letter that Bethlehem (1999) uses to denote the response probability.
[12] Vector-valued variables are written in bold letters in this thesis.
[13] The terminology for classifying different nonresponse processes introduced by Little and Rubin refers to item- and unit nonresponse and is of a more general kind than the one presented in this thesis. For details see Little & Rubin (1987), ch. 1.6.

mechanism resembles simple random sampling:

$$\delta_k = P(k \in r) \tag{2.14}$$

2. The response probability does not depend directly on the unknown survey variable $y$, but on the known auxiliary data $\boldsymbol{x}$. The survey variable $y$ is correlated with the response probability only via the auxiliary data which are also predictors for the survey variable. The response probabilities differ depending on the value of $\boldsymbol{x}$, which is known for the unit response set $r$ as well as for the unit nonresponse set $q$. In other words, the $\delta_k$ are constant within classes of $\boldsymbol{x}$, so this response mechanism can be described as stratified simple random sampling:

$$\delta_k = P(k \in r \mid \boldsymbol{x}) \tag{2.15}$$

3. The response probability depends directly on the unknown survey variable $y$. Since the $\delta_k$ depend on values of $y$ which are not known for all elements in the sample, their value cannot be assessed.[14]

$$\delta_k = P(k \in r \mid \boldsymbol{y}) \tag{2.16}$$

The first case mentioned above is also called "missing completely at random (MCAR)" (Little & Rubin 1987, p. 14). The probability of response is completely unrelated to the survey variable and to characteristics of the whole sample which are known prior to the survey. Therefore, the unit response set may be regarded as a random subsample of the selected sample. In this case estimations are the same for the respondents and nonrespondents and higher nonresponse rates only yield a higher variance for estimates but no bias. The second description of the response process refers to the case where the response probability depends on the auxiliary information known for the entire sample, but not directly on the survey variable of interest. So between different values or classes of $\boldsymbol{x}$ the probability of taking part in the survey is higher or lower, but for the same values or within the same classes of $\boldsymbol{x}$ the response probability is the same. Therefore the second case is called "missing at random (MAR)" (Little & Rubin 1987, p. 14). A unit response set which is constituted this way may also be regarded as a random

[14] Of course, if $y$ was known for the whole sample before the event of response, i.e. from a register, then $\delta_k$ could be estimated, but in this case a survey would not be necessary in order to gather information on $y$.

subsample of the sample set within classes of $\boldsymbol{x}$. Because of the fact that $\boldsymbol{x}$ is known for respondents and nonrespondent it is possible to adjust for bias due to unit nonresponse. Since in the first two cases bias caused by unit nonresponse does not occur or can be adjusted these mechanism are also called "ignorable" (Little & Rubin 1987, p. 15). In the last case the response probability depends on the variable that has to be surveyed and therefore is unknown for the nonrespondents. Here no adjustments to guarantee unbiased estimation based on the unit response set are possible and so this case is called "nonignorable" (Little & Rubin 1987, p. 15).

If it is assumed, that all elements in the sample are also units in the population, i.e. there are no coverage errors and all elements in the sample are "eligible" (Cobben 2009, p. 12), then four sources of unit nonresponse can be distinguished: unprocessed cases, non-contact, not-able and refusal (Cobben 2009, p. 13). The first one stems from lack of resources during the fieldwork. Facilitating a well-trained fieldwork unit or fieldwork institute the number of unprocessed cases may be held low. Therefore in this thesis nonresponse due to unprocessed cases is assumed to be negligible. Secondly, in some cases it may not be possible to establish contact, because the selected element (e.g. a household) cannot be reached due to physical obstacles (e.g. floods) or the selected unit cannot be reached because of absence of potential respondents (e.g. all the members of the household are temporarily away). If contact can be established it is then possible that a selected unit is not able to deliver information about the variables of interest. In a household survey, it may happen, that all household members are unable to respond due to illness or lack of language skills. The last of the four sources of unit nonresponse mentioned above refers to the case where the selected unit can be reached, but explicitly refuses to take part in the survey.

Despite unit nonresponse another source of missing data is "item nonresponse" (Groves et al. 2004, p. 169). Item nonresponse occurs, if a selected unit takes part in the survey, but does not answer all relevant questions and therefore some variables of interest are missing for that respondent. In this thesis it is always assumed that a "full rectangular data matrix" (Särndal & Lundström 2005, p. 158) is available. This means that all missing values which are a consequence of item nonresponse have been imputed. "Imputation" refers to various methods of estimating missing values in the respondents' dataset. Examples of these methods are mean imputation, regression imputation (with or without added stochastic error terms) or hot-deck imputation.[15]

[15]For a detailed presentation of these methods please refer to Little & Rubin (1987).

## 2.3 Bias due to Unit Nonresponse

As was described in section 2.1 unbiased estimation of a parameter based on a probability sample can easily be achieved with applying design weights, since these weights are the inverse of the known inclusion probabilities $\pi_k$. So one part of the error due to sampling, namely the sampling bias, can be avoided by using the unbiased Horvitz-Thompson estimator. Besides the sampling error, unit nonresponse can be another source of error. It belongs to the class of "non-sampling errors" (Cobben 2009, p. 125f.). The biases which correspond to these two sorts of error, namely the sample bias and the unit "nonresponse bias" (Särndal & Lundström 2005, p. 46) can be written in combined form. In order to demonstrate this, the unit nonresponse error, which originates from the response process, and the sampling error presented in (2.6) have to be formalized together. Referring to section 2.1, let $\hat{\theta}$ denote the estimator based on the whole sample. For the purpose of distinguishing this estimator for data with full response, $\hat{\theta}_{UNR}$ shall define the estimator under unit nonresponse. Then the error of using data from a probability sample that also encountered some amount of unit nonresponse can be written as follows:[16]

$$\hat{\theta}_{UNR} - \theta = (\hat{\theta} - \theta) + (\hat{\theta}_{UNR} - \hat{\theta}) \tag{2.17}$$

$\hat{\theta}_{UNR}$ denotes the parameter estimator based on the unit response set. The expectation of the sampling error $E(\hat{\theta}-\theta)$ is zero if an unbiased estimator like the Horvitz-Thompson estimator presented in (2.7) is used. The unit nonresponse bias can be written as the expected difference of the estimator based on the whole sample and the estimator based only on the unit response set.

$$B_{TR}(\hat{\theta}) = E_{TR}(\hat{\theta}_{UNR} - \hat{\theta}) = E_{TR}(\hat{\theta}_{UNR}) - E_T(\hat{\theta}) \tag{2.18}$$

The expectation of the estimator based on the sample refers to the sample membership indicator $T_k$ and the expectation of the estimator based on the unit response set refers to the indicator $R_k$ which identifies responding elements. In order to distinguish these two expectations the suffixes "$T$" and "$R$" are used, where "$TR$" denotes the condi-

[16] The derivations are based on Särndal & Lundström (2005), ch. 5.2.

tional expectation of the response indicator $R$, given the sample membership indicator $T$. Since the Horvitz-Thompson estimator is unbiased for probability samples, the unit nonresponse bias in (2.18) simplifies to

$$B_{TR}(\hat{\theta}) = E_{TR}(\hat{\theta}_{UNR}) - \theta \tag{2.19}$$

An alternative expression of the Bias in (2.19) is the so-called "relative bias" (Särndal & Lundström 2005, p. 91), which can be interpreted as the percentage of deviation of the parameter estimator from the the true parameter value (if it was multiplied by 100):

$$RB_{TR}(\hat{\theta}) = \frac{B_{TR}(\hat{\theta})}{\theta} = \frac{E_{TR}(\hat{\theta}_{UNR}) - \theta}{\theta} = \frac{E_{TR}(\hat{\theta}_{UNR})}{\theta} - 1 \tag{2.20}$$

The overall bias of the estimator $B_{TR}(\hat{\theta}_{UNR})$ is the sum of the sampling bias and the unit nonresponse bias which is easy to see by combining (2.6) and (2.18):

$$\begin{aligned} B_{TR}(\hat{\theta}_{UNR}) &= E_{TR}(\hat{\theta}_{UNR} - \theta) = \\ &= E_{TR}(\hat{\theta}_{UNR}) - E_{TR}(\theta) - E_T(\hat{\theta}) + E_T(\hat{\theta}) = \\ &= E_{TR}(\hat{\theta}_{UNR}) - E_T(\hat{\theta}) + E_T(\hat{\theta}) - \theta = \\ &= B_{TR} + B_T \end{aligned} \tag{2.21}$$

The part of the bias which relates to the sampling error can easily be avoided by using the Horvitz-Thompson estimator presented in formula (2.7) which expands the value of the survey variable in the sample by the inverse of the selection probability. However, if this estimator was applied to the response set, then $\hat{\theta}_{UNR}$, which is an estimator of a total, would certainly be biased if unit nonresponse occurred. If a total had to be estimated, the resulting estimator would be negatively biased, because inference would have to be made on a set that was smaller than the sample. The estimates of the total would be systematically too small, but if certain assumptions about the response probabilities $\delta_k$ hold, the exact values of the response probabilities do not have to be known. As was discussed in the previous section, the unit response set $r$ may be regarded as a simple random subsample of the selected sample $s$ if the $\delta_k$ are constant for all elements of the sample (compare (2.14)). So if the assumption holds, that the response mechanism is MCAR the response probability reduces to a constant factor equal to (size of unit response set)/(size of sample). Let $m$ denote the size of the

response set $r$ and $n$ denote the size of the sample $s$, then

$$\delta_k = \frac{m}{n} \tag{2.22}$$

By the definition of the MCAR case the response probability does not depend on characteristics of the sample, so the probability of being selected and being a respondent is the product of $\pi_k$ and $\delta_k$ for the MCAR case:

$$P(k \in s \wedge k \in r) = \pi_k \cdot \delta_k \tag{2.23}$$

Let $d_k = 1/\pi_k$ denote the design weight and $\delta_k = m/n$ is a constant. Since the product in (2.23) is known for all elements in the sample and also for all elements in the population, it can be used to construct an unbiased estimator just as in (2.7):

$$\begin{aligned} \hat{\theta}_{EXP} = \sum_{k \in s \cap r} \frac{d_k y_k}{\delta_k} = \frac{n}{m} \sum_{k \in s \cap r} d_k y_k = \\ = \frac{n}{m} \sum_{k \in s \cap r} d_k y_k \end{aligned} \tag{2.24}$$

The estimator shown above, which compensates the loss of data due to nonresponse simply by inflating the response set $r$ to the size of $s$ and subsequently $s$ to the size of $U$ is also called the the "(straight) expansion estimator" (Särndal & Lundström 2005, p. 68). The estimator shown in (2.24) resembles the Horvitz-Thompson estimator for two-phase sampling (Särndal et al. 2003, p. 348). It follows from (2.11) and (2.22) that

$\hat{\theta}_{UNR}$ is unbiased for stratified simple random sampling if MCAR is assumed.

$$\begin{aligned}
E(\hat{\theta}_{EXP.STSI}) &= E(\sum_{d=1}^{D} \sum_{k \in s_d \cap r} \frac{d_k y_k}{\delta_k}) = \\
&= E(\sum_{d=1}^{D} \sum_{k \in s_d \cap r} \frac{N_d}{n_d} \frac{y_k}{m_d/n_d}) = \\
&= E(\sum_{d=1}^{D} \sum_{k \in U_d} T_{kd} R_k N_d \frac{y_k}{m_d}) = \\
&= \sum_{d=1}^{D} E(\sum_{k \in U_d} \frac{n_d}{N_d} \frac{m_d}{n_d} N_d \frac{y_k}{m_d}) = \\
&= \sum_{d=1}^{D} \sum_{k \in U_d} y_k = \sum_{d=1}^{D} \theta_d = \\
&= \theta
\end{aligned} \tag{2.25}$$

As appealing the case of MCAR may be for estimation it unfortunately is rarely realistic and overly simple (cf. p. 12). Unit nonresponse will almost certainly have some distorting effect on estimators in sample surveys, therefore "nonresponse invariably causes some bias" (Särndal & Lundström 2005, p. 45).

So it is not very reasonable to assume that the response probability is not influenced at all by some characteristics of the elements in the sample. For example, taking part in a survey usually demands some time from the potential respondent. It is obvious that the amount of spare time is different between people who have an occupation or are jobless or people who have to look after their kids. Even if it is not known whether the response probability is different for certain demographics, it is reasonable to assume a response mechanism that differs for certain known groups. Therefore it is a good choice to take a conservative approach and assume the response mechanism to be MAR. If it turns out to be MCAR, nothing is lost, because the MAR assumption incorporates MCAR as a special case where there is only one stratum within which the response probability is the same. The primary task now is to find auxiliary information $\boldsymbol{x}$ with which it is possible to construct groups with constant response probabilities. This approach is called the "response homogeneity group (RHG) model" (Särndal et al. 2003, p. 578). The difficult task now is to find good auxiliary variables and methods with which adjustments and ameliorations of the estimators already shown can be achieved.

# 3 Weighting

## 3.1 Weighting and unit nonresponse

In the previous chapter some adjustments to counter unit nonresponse have already been introduced. The basic assumptions are that response probabilities are constant within certain groups (MAR). If these groups can be defined correctly, adjustment weights can be computed for each element $k$ in the response set. One weighted estimator has already been described in the previous chapter, namely the Horvitz-Thompson estimator which adjusts for (unequal) selection probabilities by multiplying each element by the inverse of its selection probability, also called the design weight (cf. section 2.1). However, these weights are only one form of adjustment weights used in sample surveys. Kish (1990) describes seven different forms of weighting:

**(i)** disproportional allocations, i.e. unequal selection probabilities

**(ii)** unequal sampling frames

**(iii)** nonresponse

**(iv)** statistical adjustments

**(v)** combination of samples

**(vi)** control statistics

**(vii)** nonprobability samples

Point (i) refers to the Horvitz-Thompson estimator and is an application of the design weights, (ii) is not considered here, because a sampling frame from which all elements are selected in one stage is assumed. (v) does not apply because only one sample is used and (vii) also does not hold, because a probability sample is utilized. The main reason why weighting is necessary in this thesis is (iii), i.e. unit nonresponse, more specifically,

weighting is a strategy to counter unit nonresponse bias. The response process is assumed to be random with equal selection probabilities within certain constructed classes according to specific auxiliary variables. Statistical adjustments (iv) may also help to reduce bias due to unit nonresponse as we shall see later with the method of calibration (cf. section 3.4). Adjustments to ensure comparability with other statistics (vi) refer to methods of ex-post adjustments in order to guarantee that inference on the same frame population is possible from different samples. For example, these ex-post adjustments are applied to the final weights of EU-SILC in Austria to ensure that the marginal counts of some common variables like age, sex, household size etc. are the same as in the Austrian Microcensus on a yearly average (Glaser & Till 2010, p. 571). Although the same methods as in (iv) may be used (post-stratification, calibration), the purpose of (iv) is not to adjust for errors due to sampling or nonresponse.

It is important to note that the weights considered in this thesis are all "individual case weights" (Kish 1990, p. 125). In every weighting scheme considered here, each element $k$ in the response set gets assigned one weight that is used for calculation of parameter estimates. Although bias is evaluated with the parameter "total income from employment and old-age benefits" the weights can be used to calculate all different kinds of statistics. This strategy is opposed to the one using different weights for different statistics.

In the previous section it became apparent, that some form of auxiliary information will be needed to find an ordering with which the response process may be explained. The probability of response, given an auxiliary information $\boldsymbol{x}$ is defined as the "response propensity" (Little 1986, p. 146):

$$p(x) = P(R = 1 \mid x) \tag{3.1}$$

Obviously, finding a good auxiliary vector $\boldsymbol{x}$ is a crucial part in using adjustments of weights in order to counter unit nonresponse. Särndal & Lundström (2005) distinguish three main properties that these auxiliary informations should have:

1. Explanation of the response propensity
2. Explanation of the most important study variable(s)
3. Explanation of the most relevant domains

The first property is evidently necessary in the RHG model. Otherwise the grouping according to $\boldsymbol{x}$ would not help in adjusting for unit nonresponse. Furthermore it gen-

erally leads to a reduction of the unit nonresponse bias for all variables.[1] The second point mentioned above does not only contribute to reducing the bias for the main survey variable(s) of interest, but also reduces the variance of these estimates. This closely resembles the concept of "optimum allocation" (Kish 1995, p. 92). It also refers to what Groves (2006) calls the "common cause model" (Groves 2006, p. 651). In this causal model the survey variable(s) and the response propensity are both caused by the same variable(s). Any covariance between the main survey variable(s) and the response propensity can therefore be explained by common factors. The third desirable property of an auxiliary vector ensures a correct representation of important domains. For example, even if the sampling design uses stratification for important categories (e.g. region, age, sex, ...), unit nonresponse may lead to a distorted representation of these groups and therefore has to be corrected if the totals of these domains shall be represented correctly.

This chapter will give an overview of different weighting methods and their power to counter unit nonresponse. In the RHG model already mentioned in the previous chapter response is explained as happening at random with equal probabilities within groups, resembling response as a second-phase stratified simple random sample. This traditional approach is also called the "two-phase approach to weighting for nonresponse" (Särndal & Lundström 2005, p. 50) and can also be used for the MCAR case by assuming just one response homogeneity group.

## 3.2 Two-phase weighting

Weighting for unit nonresponse was already described in the previous chapter for the case of having the same response probabilities for all elements in the sample (MCAR). Often the sampling frame and the whole sample contain some variables that are known for respondents and nonrespondents. Usually these variables are those used for drawing the sample (e.g. region, age) or characteristics of the sampled elements that become known after the sample is drawn, but before the survey takes place (e.g. type of building). Now there are different methods for incorporating these auxiliary informations and using them to model the response probabilities.[2] All approaches presented in this section have one thing in common: the response process is modeled as a second sam-

[1] Compare Särndal & Lundström (2005), ch. 10.2.

[2] A discussion of different adjustments methods can be found in Little (1986) and Potter et al. (2006).

pling phase after the sample was drawn. It is assumed that, besides the known inclusion probabilities $\pi_k$, there also exists an unknown probability of response $\delta_k$ for each element in the sample. This approach, which is also used throughout this section, was described by Oh & Scheuren (1983) and is called "quai-randomization theory".

### 3.2.1 Straight expansion estimator

The simplest form of adjusting an estimator for unit nonresponse is to divide the Horvitz-Thompson estimator by the response rate and therefore implicitly assuming a constant response probability for elements $k$ in the sample $s$ under the presumption of a random response mechanism.The straight expansion estimator has already been presented in the previous chapter as

$$\hat{\theta}_{EXP} = \frac{n}{m} \sum_{k \in r} d_k y_k \tag{3.2}$$

For the sake of convenience the subscript "$k \in s_d \cap r$" is simplified to "$k \in r$" from now on, because, obviously, element $k$ can only be part of the response set if it was drawn in the sample.

In the special case were the selection probabilities and the response probabilities are constant, which is the case for simple random sampling and MCAR, (2.24) simplifies to the mean value $\overline{y}_r$ of $y$ over the response set $r$ multiplied by the size of the population, $N$:[3]

$$\begin{aligned} \hat{\theta}_{EXP} &= \sum_{k \in r} \frac{d_k y_k}{\delta_k} = \sum_{k \in r} \frac{N}{n} \frac{y_k}{m/n} = \\ &= \frac{N}{m} \sum_{k \in r} y_k = N \overline{y}_r \end{aligned} \tag{3.3}$$

Presumably, the straight expansion estimator will not be suitable for countering unit nonresponse because of the unrealistic assumption that unit nonresponse happens completely at random. In the simulations presented in section 4 it will be a point of reference for other methods that are assumed to be better in the sense of producing less biased estimates than the straight expansion estimator. Therefore, the response process will

[3] Cf. Särndal et al. (2003), p. 68.

be assumed to be MAR from now on, meaning that response is happening at random only within certain subgroups of the population.

### 3.2.2 Response homogeneity groups (RHG) and weight adjustment cells

Let there be $H$ disjunct groups in the sample $s$. The formation of these groups can be achieved according to one variable or to a cross-tabulation of different variables. For each group $h = 1, ..., H$ there exists a response probability $\delta_h$ which is the fraction of the number of respondents $m_h$ in group $h$ divided by the number of selected elements in $n_h$, which shall be called the "response fraction". For each element $k$ in $h$ the response probability therefore is constant:[4]

$$\delta_h = P(k \in h \wedge k \in r) = \frac{m_h}{n_h} \tag{3.4}$$

In each of the $h$ so-called "adjustment cells" (Little 1988, p. 289) a constant weight equal to $n_h/m_h$ is applied to counter unit nonresponse. Note that in (3.4) it is assumed that the $h$ groups correctly assess the true response distribution. The corresponding adjustment weight is the inverse of the product of the selection probability and the response probability, i.e. $1/(\pi_k \delta_h)$. Since the response process is treated as a second sampling phase the Horvitz-Thompson estimator can be applied. Let $\hat{\theta}_{RHG}$ denote the unbiased estimator in the case of a stratified simple random sample and a response process referring to RHG and $\nu_k^{(RHG)} = 1/\delta_h$ the "unit nonresponse weight" where the superscript "$(RHG)$" shall distinguish this nonresponse weight from the ones used in subsequent sections:

$$\hat{\theta}_{RHG} = \sum_{k \in r} \frac{d_k}{\delta_h} y_k = \sum_{k \in r} d_k \nu_k^{(RHG)} y_k \tag{3.5}$$

The estimator shown above resembles the Horvitz-Thompson estimator for two-phase sampling with stratification in both phases and is therefore unbiased if the response probability really only differs between the groups $h$. The values of $y_k$ receive one weighting factor which is the product of the design weight and the inverse of the estimated response probability.

[4] The nomenclature in this section is mostly based on Särndal et al. (2003), ch. 15.6.

One major problem with the weight adjustment cells approach is the size of the groups. If they became too many and too diverse this may result in extreme values of the inverse sampling fraction or in groups which contain no respondents, because they are too small. Another way of using the two-phase approach is to estimate the response probability for each element $k$ in the sample with a regression model.[5]

### 3.2.3 Estimation of response probabilities with logistic regression

If a model can be found with which the true response probability can be estimated, then many variables may be used to find predictors without having to rely on adjustment factors that would have to be based on adjustment cells that might be quite small. In this subsection the response probability is described as an individual feature of each element $k$. In order to make this clear, the response probability is written in terms of the response indicator $R$:

$$\delta_k = P(R_k = 1) \tag{3.6}$$

An appropriate way for estimating a dichotomous variable like the response indicator $R$ is logistic regression.[6] Let $\boldsymbol{x} = (x_1, ..., x_I)'$ be a vector of $I$ auxiliary variables which shall be the predictors in the model. The dependent variable is the natural logarithm of the odds,[7] i.e. $P(R = 1)/P(R = 0)$, where $R$ denotes the response indicator defined in secction 2.2. Note that the probability of response and nonresponse always sum up to 1 for each element $k$, i.e. $P(R_k = 0) = 1 - P(R_k = 1)$. The logistic regression model for estimating the response probability $\delta_k$ is defined as follows:

$$ln\Big(\frac{P(R = 1 \mid \boldsymbol{x})}{P(R = 0 \mid \boldsymbol{x})}\Big) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + ... + \hat{\beta}_I x_I \tag{3.7}$$

With a few simple algebraic transformations, the model desricption shown above can be rewritten in terms of the probability of response:

[5] A proof of the unbiasedness of this estimator can be found in Särndal et al. (2003), ch. 9.4. where the estimator is denoted $\hat{t}_{\pi^*}$.

[6] A basic introduction to logistic regression models can be found in Backhaus (2006), ch. 7. A detailed description of this method is laid out in Hosmer & Lemeshow (1989).

[7] Cf. Hosmer & Lemeshow (1989), ch. 3.

$$\frac{P(R=1 \mid \boldsymbol{x})}{1-P(R=1 \mid \boldsymbol{x})} = \exp(\hat{\beta}_0 + \hat{\beta}_1 x_1 + ... + \hat{\beta}_I x_I)$$

$$P(R=1 \mid \boldsymbol{x}) = \exp(\hat{\beta}_0 + \hat{\beta}_1 x_1 + ... + \hat{\beta}_I x_I) -$$
$$- P(R=1 \mid \boldsymbol{x}) \cdot \exp(\hat{\beta}_0 + \hat{\beta}_1 x_1 + ... + \hat{\beta}_I x_I)$$

$$P(R=1 \mid \boldsymbol{x}) = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x_1 + ... + \hat{\beta}_I x_I)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x_1 + ... + \hat{\beta}_I x_I)} \tag{3.8}$$

After the parameters in the regression model have been estimated it is possible to calculate an estimate $\hat{\delta}_k$ of the response probability for each element $k$ in the sample by inserting the values of the auxiliary variables, where $\boldsymbol{x}_k$ denotes the vector of auxiliary information for element $k$ in the model, i.e. $\boldsymbol{x}_k$=$(x_{1k}, ..., x_{Jk})$. This results in an estimate of the response probability for each element $k$ in the sample. Let $\hat{\delta}_k^{(LOG)}$ denote the response probabilities estimated by logistic regression. The estimator is then similar to (3.5):

$$\hat{\theta}_{LOG} = \sum_{k \in r} \frac{d_k}{\hat{\delta}_k^{(LOG)}} y_k = \sum_{k \in r} d_k \nu_k^{(LOG)} y_k \tag{3.9}$$

Evidently the values of the estimated response probabilities remain constant for groups with the same values of the variables in $\boldsymbol{x}$. The bias of the estimator in (3.9) depends on the quality and the scope of the auxiliary information at hand. However, it has to be assumed that all the relevant factors describing the individual response probabilities will never be known, some amount of bias will inevitably occur. The major challenge is, that the esimation of the response probability has to be done with information known before the actual survey takes place. This information has to come from the sampling frame or characteristics of the drawn sample that can be collected without contacting the elements in the sample.

One possible consequence of using the inverse of the response probability as a weight to compensate for unit nonresponse are extremely large unit nonresponse weights. Since these weights are calculated as the inverse of the estimated response probabilities, a $\hat{\delta}_k$ which is close to zero will result in an extremely large weight. One strategy to avoid this problem, which is also applied in the EU-SILC survey, is trimming the unit nonresponse weights.[8] For each weight a measure of the relative change of the design weight due to

[8] Cf. Osier et al. (2006), ch. 2.

multiplication by the nonresponse weight is calculated. Let $\bar{d}_R$ denote the mean value of the design weights for the respondents and $\bar{\nu}$ the mean of the unit nonresponse weights:

$$\frac{1}{C} \leq \frac{\nu_k/\bar{\nu}}{d_k/\bar{d}_R} \leq C \tag{3.10}$$

The fraction in (3.10) shows the value of the unit nonresponse weights divided by their mean value in relation to the design weights devided by their mean value. If this value is below $1/C$ or above $C$ then the unit nonresponse weights are changed in order to fit the boundaries.[9] If $\nu_k$ in (3.10) is below $1/C$ then $\nu_k$ becomes

$$\nu_k^{(low)} = \frac{1}{C} d_k \frac{\bar{\nu}}{\bar{d}_R} \tag{3.11}$$

and if $\nu_k$ in (3.10) is above $C$ then $\nu_k$ is changed to

$$\nu_k^{(high)} = C d_k \frac{\bar{\nu}}{\bar{d}_R} \tag{3.12}$$

Although trimming inhibites extreme weights, it also may introduce an new source of bias, because the adjusted unit nonresponse weights $\nu_k^{(low)}$ and $\nu_k^{(high)}$ also change the mean $\bar{\nu}$. The reduction of variance is due to the elimination of values that are categorized as too high or too low by (3.10).

### 3.2.4 Neural Networks

The method for estimating response probabilities presented in the previous subsection does rely on the assumption, that with the appropriate predictor variables, the response probability can be estimated with a regression model. A logistic regression is best suited here, because of the binary dependent variable, i.e. the response indicator. However, only if the logisitic regression model is very well specified and a strong relation between the predictor variables and the response indicator are available, then the model can deliver a good estimate of the probability of response.

A method that circumvents this limitation is the neural network (NNW).[10] Neural networks do not rely on a specific model formulation and are allowed to iteratively

[9] EU-SILC in Austria uses $C = 2$ for the definition of the boundaries.
[10] A general introduction to artificial neural networks can be found in Ripley (2004).

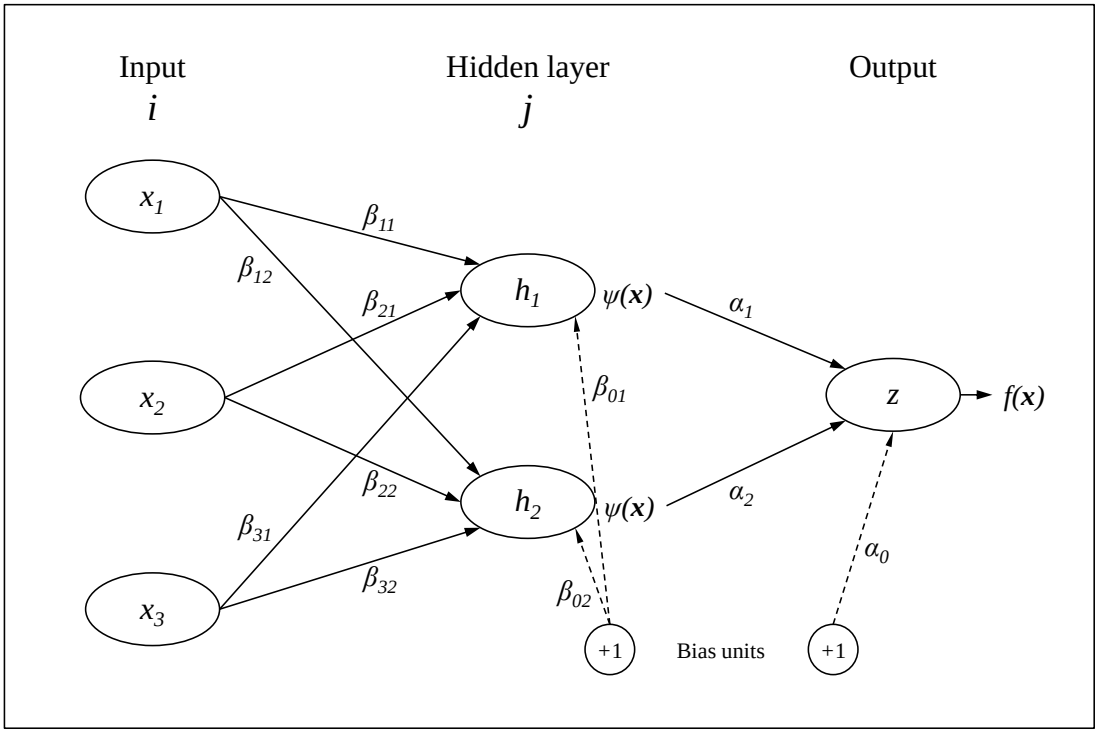
establish connections between predictor variables and the dependent variable by themselves. This may be particularly useful for estimating unit response probabilities when it is unknown how the auxiliary variables and the response process are exactly related. Furthermore, the response process incorporates several steps like establishing contact, first time asking to take part in the survey and addtitional contact attempts if the first one did not succeed. These steps are not easy to model explicitly and hence methods with their own heuristics may turn out to be particularly helpful.[11] Neural networks are a method of pattern recognition that were inspired by the way a network of neurons in the human brain processes information and does prediction based on sensory inputs.[12]

A common form of neural networks are "feed-forward neural networks" (cf. Ripley 2004, p. 141) with one hidden layer. In their simplest version they consist of a set of input units, the so-called "hidden layer" (cf. Ripley 2004, p. 144) and one output unit. The input units are known auxiliary variables which are used to predict the known output variable. The output variable is categorial and refers to a known correct classification of the elements in the model. The units in the hidden layer, which receive signals from the input and pass signals on to the output in a forward direction, can be trained to learn their own relations between the input and the output. So instead of explicitly formulating a model of how the dependent variable (i.e. the output) and the predictor variables (i.e. the input) are connected, the neural network establishes its own connections between the input and the output iteratively by learning from a given data set. Therefore, neural networks can cover a broad set of functional relations and, moreover, "neural networks with linear output units and a single hidden layer can approximate any continous function $f$ uniformly on compacta" (cf. Ripley 2004, p. 147). A proof of this statement can be found in Ripley (2004), p. 174.

The importance of each connection in the neural network is described by weights connecting the input and the output via the hidden layer. A network trained with the predictor variables in the input and the known correct classification in the output can be used to classify elements only using the predictor variables from the input. Figure 3.1 shows a feed-forward neural network with three input variables, a single hidden layer with two hidden units and one output unit.[13]

[11] Cf. Liang & Yung Cheung Kuk (2004), p. 220f.
[12] Cf. Ripley (2004), ch. 5.1.
[13] The figure is based on Ripley (2004), figure 5.1.

**Figure 3.1:** *Feed-forward neural network with a single hidden layer with two units and one output unit*




Generally a feed-forward neural network with one output unit consists of the vector $\boldsymbol{x} = (x_1, ..., x_I)'$ of $I$ input units $x_i$ referring to $I$ auxiliary variables, $J$ hidden units $h_j$ $(j = 1, ..., J)$ and the output, which is a linear function of the units in the hidden layer:

$$f(\boldsymbol{x}) = \alpha_0 + \sum_{j=1}^{J} \alpha_j h_j \tag{3.13}$$

The importance of the connections is determined by the weights $\beta_{ij}$ for the connections between the input and the hidden layer and the weights $\alpha_j$ for the connection between the hidden layer and the output. The value of the variables $h_j$ in the hidden layer is determined by the so-called "activation function" (Liang & Yung Cheung Kuk 2004, p. 221), denoted here by $\psi$:

$$h_j = \psi(\beta_{0j} + \sum_{i=1}^{I} \beta_{ij} x_i) \tag{3.14}$$

The activation function establishes a link between the input and the output through the hidden layer. It can take various forms like a function of binary outcome values, a linear function, a nonlinear function, and so forth. In this thesis the logistic function is used, because the application of neural networks is motivated by estimating response probabilities. Furthermore, feed-forward neural networks can be interpreted as an expansion of logistic regressions in a nonlinear way.[14] Therefore, the activation function $\psi(t)$ used in this thesis is defined as follows:

$$\psi(t) = \frac{1}{1 + exp(-t)} \tag{3.15}$$

The so-called "bias units" (cf. Ripley 2004, p. 144) are always connected to the hidden layer and the output. Note that the term "bias" here refers to the concept of "debiasing" (cf. Ripley 2004, p. 55) density estimators, which refers to adjustments of density estimators in order to make them theoretically unbiased. The bias units are always active with a constant value of 1. This is comparable to having established a connection where the outcome of the activation function is always equal to 1. The weights of the bias units refer to the constant terms in (3.13) and (3.14). The usage of bias units can be referred to adding a column with constants equal to 1 in the matrix of explanatory variables of a regression model and thus allowing a constant to be included in the regression.

Putting (3.13) and (3.14) together the feed-forward neural network with a single hidden layer and one linear output unit can be written as the following function:

$$\hat{z} = f(\boldsymbol{x}) = \alpha_0 + \sum_{j=1}^{J} \alpha_j \psi(\beta_{0j} + \sum_{i=1}^{I} \beta_{ij} x_i) \tag{3.16}$$

The fitting of the neural network is achieved by minimizing the loss referring to wrong classification by minimizing a suitable "loss function" (Ripley 2004, p. 18ff.).

Let $z$ denote the correct known classification for element $k$. The classification by $\hat{z} = f(\boldsymbol{x})$ based on the matrix of input vectors should be as close as possible to the correct

[14] Cf. Ripley (2004), p. 145.

classification $z$. A very often used loss function is quadratic loss:[15]

$$L(\boldsymbol{x}) = (f(\boldsymbol{x}) - z)^2 \tag{3.17}$$

The parameters of the neural network are estimated by minimizing (3.17) according to the weights $\alpha_0, \alpha_j, \beta_{0j}$ and $\beta_{ij}$:

$$\min_{\alpha_0, \alpha_j, \beta_{0j}, \beta_{ij}} L(\boldsymbol{x}) = E\big((f(\boldsymbol{x}) - z)^2\big) =: \Delta \tag{3.18}$$

The solution of this minimization problem is then found by taking the partial derivates with respect to the weights and so iteratively updating the weights. This learning procedure trains the neural network in order find the minimum. Since the calculation of the partial derivatives can be reformulated in terms of the desired correct output $z$, the referring fit algorithm is called "back-propagation" (cf. Ripley 2004, p. 149). A detailed presentation of back-propagation and the corresponding improvements of the algorithm can be found in Ripley (2004), ch. 5.3.

Although neural networks are primarily used in applications of pattern analysis (i.e. classifying plants according to their leafs, speech recognition)[16] they can also be used to estimate probabilities, more specificially Bayesian a posteriori probabilities.[17] In the following subsection it will be shown how response probabilities can be estimated by using the capability of neural networks to estimate a posteriori probabilities according to Bayes' theorem.

#### 3.2.4.1 Bayesian a posteriori probabilties

According to Bayes' theorem the conditonal probability $P(A_i|B)$ of event $A_i$, given event $B$ can be calculated as follows:

$$P(A_i \mid B) = \frac{P(A_i)P(B \mid A_i)}{P(B)} \tag{3.19}$$

In the formula above $P(A_i)$ is called the *a priori* probability of event $A_i$ and $P(A_i \mid B)$ is called the *a posteriori* probability of $A_i$, given that $B$ is known. In Bayesian statistics

[15] Cf. Richard & Lippmann (1991), ch. 2.2.
[16] Cf. Ripley (2004).
[17] Cf. Richard & Lippmann (1991), p. 466.

probability distributions are updated by starting with a prior distribution and further updating the distribution and gaining a posterior probability distribution. Since the unconditional probability $P(B)$ represents a constant factor in the case of finding the posterior distribution it is often omitted and the posterior distribution is formulated proportional to the numerator of the right hand side of formula (3.19):

$$P(A_i \mid B) \propto P(A_i)P(B \mid A_i) \tag{3.20}$$

Formula (3.19) can also be written in terms of the response indicator $R_k$ presented in section 2.2, where $P(R_k) = P(k \in r)$ is the unconditional probability of element $k$ belonging to set $r$, i.e. being a respondent, and $P(k \in r \mid \boldsymbol{x})$ is the probability of element $k$ being a respondent conditional on the vector of auxiliary variables $\boldsymbol{x}$. Let $P(R_k)$ denote the a priori probability of response and $P(R_k \mid \boldsymbol{x})$ the a posteriori probability of response, (3.19) then becomes:

$$P(R_k \mid \boldsymbol{x}) = \frac{P(R_k)P(\boldsymbol{x} \mid R_k)}{P(\boldsymbol{x})} \tag{3.21}$$

Since the unconditional probability of the auxiliary variables $P(\boldsymbol{x})$ is the same regardless of the value of the response indicator $P(R_k)P(\boldsymbol{x} \mid R_k)$ can be used for finding the posterior probabilities. In a traditional approch the joint probability distribution $P(\boldsymbol{x}, R_k) = P(R_k)P(\boldsymbol{x} \mid R_k)$ would have to be modeled in order to find the posterior probability distribution. The following section shows that this can also be achieved by using neural networks.

#### 3.2.4.2 Estimation of the a posteriori probability of response by neural networks

The following formulation of the conditional probability of response by using neural networks is based on Richard & Lippmann (1991) and was adapted for the notation used in this thesis. The expectation of the squared error in the minimization problem for the neural network described in (3.18) can be reformulated in terms of the probability of response, i.e. the response indicator $R_k$. In terms of a classification problem there are two classes, respondents and nonrespondents which shall be denoted by the values $r_k = 0$ and $r_k = 1$ which are realized values of the random variable $R_k$ and the associated probability of realization is $P(R_k = r_k)$.

$$\Delta = E\big((f(\boldsymbol{x}) - r_k)^2\big) \tag{3.22}$$

The expectation in (3.22) is calculated according to the joint probability distribution $P(\boldsymbol{x}, R_k)$. Richard & Lippmann (1991) show that for a quadratic error function $\Delta$ can be rewritten to[18]

$$\Delta = E\big((f(\boldsymbol{x}) - E(r_k \mid \boldsymbol{x})^2\big) + E\big(V(r_k \mid \boldsymbol{x})\big) \tag{3.23}$$

In the equation above $V(r_k \mid \boldsymbol{x})$ refers to the conditional variance of $r_k$ which does not depend on the output $f(\boldsymbol{x})$ of the neural network. The first term in the sum in (3.23) is the mean squared error between the output of the neural network and the expectation of the desired classification of $r_k$ conditional on the input values $\boldsymbol{x}$. In other words, the output of the neural network estimates the conditional expectation of classification according to the response indicator which can be formulated as the conditional probability of response, given the vector of auxiliary variables $\boldsymbol{x}$:

$$\begin{aligned} E(r_k \mid \boldsymbol{x}) &= 0 \cdot P(R_k = 0 \mid \boldsymbol{x}) + 1 \cdot P(R_k = 1 \mid \boldsymbol{x}) = \\ &= P(R_k = 1 \mid \boldsymbol{x}) = \delta_k \end{aligned} \tag{3.24}$$

The equation above leads to the conclusion that "when network parameters are chosen to minimize a squared-error cost function, the outputs estimate the Bayesian probabilities so as to minimize the mean-quared estimation error" (Richard & Lippmann 1991, p. 466). In the case of this thesis the a posteriori probabilities preseneted in (3.24) represent the response probabilities. Therefore, when a feed forward neural network with one binary output variable as the response indicator $R_k$, auxiliary input variables $\boldsymbol{x}_k$, one single hidden layer and a logistic activation function is fitted, the estimated values $\hat{z}$ in (3.16) can be used as estimates of the response probability. Let $\hat{\delta}_k^{(NNW)}$ denote the response probabilities estimated by neural networks. Once $\hat{\delta}_k^{(NNW)}$ have been calculated, they can be used as adjustment factors $\nu_k^{(NNW)}$ of the design weights just as in (3.9):

[18] The derivation is omitted here, because it is beyond the scope of the thesis. For detailed derivations please refer to Richard & Lippmann (1991), p. 464ff.

$$\hat{\theta}_{NNW} = \sum_{k \in r} \frac{d_k}{\hat{\delta}_k^{(NNW)}} y_k = \sum_{k \in r} d_k \nu_k^{(NNW)} y_k \tag{3.25}$$

## 3.3 Regression Estimation

In this and the following section methods are going to be explored which are all characterized by adjustments of the design weights that do not rely on estimated response probabilities, but by adjusting the design weights to auxiliary information that are known both for respondents and nonrespondents. An important way of incorporating auxiliary information in the estimation process with survey data regression estimation. The "generalized regression estimatior (GREG estimator)" (Särndal & Lundström 2005, p. 33) is a method of estimating the total of a survey variable in a nearly unbiased way.[19]

In order to allow for a simple illustration, full response is assumed. An important requirement is, that the totals of the auxiliary information $\sum_{k \in U} \boldsymbol{x}_k$ in the population are known. Starting point in the estimation process is again the Horvitz-Thompson estimator. Additionaly to the Horvitz-Thompson estimator a regression adjustment term will be added which uses the coefficients of the regression $y_k$ on $\boldsymbol{x}_k$. The following descriptions of the GREG estimator $\hat{\theta}_{GREG}$ are all based on Särndal et al. (2003), ch. 6.4, Särndal & Lundström (2005), ch. 4.3. and Särndal (2007), ch. 3.2.

$$\hat{\theta}_{GREG} = \sum_{k \in s} d_k y_k + \Big( \sum_{k \in U} \boldsymbol{x}_k - \sum_{k \in s} d_k \boldsymbol{x}_k \Big)' \Big( \sum_{k \in s} d_k \boldsymbol{x}_k \boldsymbol{x}_k' \Big)^{-1} \Big( \sum_{k \in s} d_k \boldsymbol{x}_k y_k \Big) \tag{3.26}$$

The two brackets on the right of the expression above contain regression coefficients weighted by the design weights. It is important to note that the matrix given by the product $\boldsymbol{x}_k \boldsymbol{x}_k'$ has to be positive definite in order to be invertible. The first bracket in (3.26) shows the difference of the true value of each auxiliary variable in the population minus the Horvitz-Thompson estimator of the auxiliary variables. The GREG estimator can also be written in terms of the values $\hat{y}_k$ predicted by the regression model:

[19] A basic introduction of the GREG estimator can be found in Särndal & Lundström (2005), ch. 4.3. Särndal et al. (2003) give a detailed description of regression estimation in ch. 6 and 7.

$$\hat{\theta}_{GREG} = \sum_{k \in U} \hat{y}_k + \sum_{k \in s} d_k (y_k - \hat{y}_k) \tag{3.27}$$

where

$$\hat{y}_k = \Big( \sum_{k \in s} d_k \boldsymbol{x}_k \boldsymbol{x}_k' \Big)^{-1} \Big( \sum_{k \in s} d_k \boldsymbol{x}_k y_k \Big) \tag{3.28}$$

Note that only the linear form of the GREG estimator is presented here, because it is widely known. For nonlinear GREG estimators please refer to Särndal (2007), ch. 3.3. As Särndal (2007) also points out, the GREG estimator is nearly unbiased, but has a considerably smaller variance than the Horvitz-Thompson estimator if the regression has a good model fit.

Yet another way of formulating the GREG estimator is in terms of an adjustment weight $g_k$ applied to the design weight. This adjustment weight is called the "g-weight" (Särndal et al. 2003, p.233). It enables a more handy formulation of the GREG estimator than in (3.26):

$$\hat{\theta}_{GREG} = \sum_{k \in s} d_k g_k y_k \tag{3.29}$$

The introduction of the g-weights is also necessery, because they are going to play an important role in the next section. For the GREG estimator the g-weights represent the regression adjustment of the Horvitz-Thompson estimator, as utilized in (3.29). As is shown in Särndal & Lundström (2005), p. 35, the g-weights can be written as

$$g_k = 1 + \boldsymbol{\lambda}_s' \boldsymbol{x}_k \tag{3.30}$$

Note that the vector $\boldsymbol{\lambda}_s'$ has dimension $I$ which is the same dimension as the vector $\boldsymbol{x}_k$ containing the values of the $I$ auxiliary variables of element $k$.

With the help of the vector $\boldsymbol{\lambda}_s'$ the coefficients from the regression of $y$ on $\boldsymbol{x}_k$ are added into (3.30), i.e.,

$$\boldsymbol{\lambda}_s' = \Big( \sum_{k \in U} \boldsymbol{x}_k - \sum_{k \in s} d_k \boldsymbol{x}_k \Big)' \Big( \sum_{k \in s} d_k \boldsymbol{x}_k \boldsymbol{x}_k' \Big)^{-1} \tag{3.31}$$

If the product of the weights $w_k = d_k g_k$ is applied to the vector of auxiliary variables in the sample, an estimate for the auxiliary information in the sample is obtained:

$$\sum_{k \in s} d_k g_k \boldsymbol{x}_k = \sum_{k \in s} w_k \boldsymbol{x}_k = \sum_{k \in U} \boldsymbol{x}_k \tag{3.32}$$

Therefore these weights "happen to be "calibrated" to (consistent with) the known population x-total" (Särndal 2007, p. 103).

## 3.4 Calibration estimators

Calibration is the term for a variety of estimators that use weights $w_k$ for which the "calibration property" (Särndal & Lundström 2005, p.35), which was already shown at the end of the previous section, holds:

$$\sum_{k \in s} w_k \boldsymbol{x}_k = \sum_{k \in U} \boldsymbol{x}_k \tag{3.33}$$

Consequently all calibration estimators used for estimating the total of $y_k$ from a sample can be written in the following form:

$$\hat{\theta}_C = \sum_{k \in s} w_k y_k \tag{3.34}$$

As in (3.26) the calibration estimator can also be written in cojunction with the Horvitz-Thompson estimator $\hat{\theta}_{HT} = \sum_{k \in s} d_k y_k$:

$$\begin{aligned} \hat{\theta}_C &= \sum_{k \in s} w_k y_k + \sum_{k \in s} d_k y_k - \sum_{k \in s} d_k y_k = \\ &= \sum_{k \in s} d_k y_k + \sum_{k \in s} (w_k - d_k) y_k \end{aligned} \tag{3.35}$$

With the help of (3.35) the bias of the calibration estimator can be expressed in a convenient fashion:

$$
\begin{aligned}
B_C &= E_T(\hat{\theta}_C - \theta) = \\
&= E_T\Big(\sum_{k\in s} d_k y_k + \sum_{k\in s}(w_k - d_k)y_k\Big) - \theta = \\
&= E_T\Big(\sum_{k\in s}(w_k - d_k)y_k)\Big) \qquad (3.36)
\end{aligned}
$$

From (3.36) it is obvious that the difference $(w_k - d_k)$ should be as small as possible in order to receive an estimator with minimum bias. The weights $w_k$ should be very close to the design weights $d_k$. This similarity is measured by the distance function $G(w_k, d_k)$ which has the following properties:[20]

- defined for all $w_k > 0$
- $G(w_k, d_k) \geq 0$ with $G(d_k, d_k) = 0$
- strictly convex
- differentiable with contious derivative $g(w_k, d_k) = \partial G(w_k, d_k)/\partial w_k$ with $g(d_k, d_k) = 0$
- $g(d_k, d_k)$ is invertible

In order to obtain calibrated weights the distance function has to be minimized with respect to $w_k$ ($d_k$ is fixed) and at the same time the calibration property presented in (3.33) has to be fulfilled. Hence the following optimization problem has to be solved:[21]

$$
\min_{w_k, k\in s} \sum_{k\in s} G(w_k, d_k) \text{ with constraints } \sum_{k\in s} w_k \boldsymbol{x}_k = \sum_{k\in U} \boldsymbol{x}_k \qquad (3.37)
$$

The optimization problem can be solved by finding the minimum of the following Lagrange function $\Lambda$ with the vector of Lagrange multipliers $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_I)'$:

$$
\Lambda(w, d, \boldsymbol{\lambda}) = \sum_{k\in s} G(w_k, d_k) - \boldsymbol{\lambda}'\Big(\sum_{k\in s} w_k \boldsymbol{x}_k - \sum_{k\in U} \boldsymbol{x}_k\Big) \qquad (3.38)
$$

[20] Cf. Särndal (2007), ch. 4.2.

[21] A thorough presentation of distance functions used in calibration weighting can be found in Särndal & Deville (1992), section 1 and section 2.

Derivation with respect to $w_k$ and setting (3.38) equal to zero yields the following Lagrange equations which are solved for all $k$:[22]

$$\begin{aligned}\frac{\partial \Lambda}{\partial w_k} &= \frac{\partial (G(w_k, d_k))}{(\partial w_k)} - \frac{\partial (\boldsymbol{\lambda}' w_k \boldsymbol{x}_k)}{(\partial w_k)} + \frac{\partial (\boldsymbol{\lambda}' \boldsymbol{x}_k)}{(\partial w_k)} \\ \frac{\partial \Lambda}{\partial w_k} &= g(w_k, d_k) - \boldsymbol{\lambda}' \boldsymbol{x}_k = 0 \end{aligned} \tag{3.39}$$

In order to solve (3.39) it is practical to transform the the arguments $w_k$ and $d_k$ of the function $g$ into one singular argument $w_k/d_k$. Särndal & Deville (1992) point out that "In most of our applications $g_k(w, d) = g(w/d)$" (Särndal & Deville (1992), p. 377) a fact that will also become apparent later in this section. So it becomes easy to derive the calibrated weights $w_k$ from (3.39). Let $F(u) = g^{-1}(u)$ denote the inverse of the derivative $g$:

$$\begin{aligned} g(w_k/d_k) &= \boldsymbol{\lambda}' \boldsymbol{x}_k \\ \frac{w_k}{d_k} &= g^{-1}(\boldsymbol{\lambda}' \boldsymbol{x}_k) \\ w_k &= d_k F(\boldsymbol{\lambda}' \boldsymbol{x}_k) \end{aligned} \tag{3.40}$$

Note that $F(\boldsymbol{\lambda}' \boldsymbol{x}_k)$ are g-weights similar to those shown in the case of the GREG estimator in (3.29), i.e. $g_k = F(\boldsymbol{\lambda}' \boldsymbol{x}_k)$. $F$ is also normalized to gain $F(0) = 1$ and $F'(0) = 1$.[23] Now inserting (3.40) into the calibration property (3.33) yields $I$ "calibration equations" (Särndal & Deville 1992, p. 377) which are used to solve for the $I$ variables in $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_I)'$:

$$\sum_{k \in s} w_k \boldsymbol{x}_k = \sum_{k \in s} d_k F(\boldsymbol{\lambda}' \boldsymbol{x}_k) \boldsymbol{x}_k = \sum_{k \in U} \boldsymbol{x}_k \tag{3.41}$$

After (3.41) is solved each element $k$ in the sample receives a calibrated weight $w_k$. Now the calibration estimator presented in (3.34) can be written in terms of a weight adjustment of the design weights:

[22] Since the partial derivatives with respect to $w_k$ are calculated, $\partial(\boldsymbol{\lambda}' w_k \boldsymbol{x}_k)/(\partial w_k) = \boldsymbol{\lambda}' \boldsymbol{x}_k$ and $\partial(\boldsymbol{\lambda}' \boldsymbol{x}_k)/(\partial w_k) = 0$.

[23] Cf. Särndal & Deville (1992), p. 378.

$$\hat{\theta}_C = \sum_{k \in s} d_k F(\boldsymbol{\lambda}' \boldsymbol{x}_k) y_k = \sum_{k \in s} d_k g_k y_k = \sum_{k \in s} w_k y_k \tag{3.42}$$

All the abovementioned descriptions of the calibration estimator were presented for application on the whole sample $s$, i.e. in the case of full response. If used in this fashion, calibration can reduce the variance of the estimator and also guarantee consistency with known population totals. However, the main concern in this thesis is unit nonresponse. Calibration can principally be applied also to the response set $r$, but apparently in this case the calibrated weights $w_k$ will not be "as close as possible" to the design weights, because they also have to compensate for the loss due to unit nonresponse. This means the $w_k$ have to be, on average, larger than the design weights. One way of using calibration in the case of unit nonresponse is by breaking the vector of auxiliary information $\boldsymbol{x}_k$ up in two parts.[24] One part contains information for every element $k$ in the sample $s$ and therefore is known for respondents and nonrespondents, the second part contains information on the population level. Calibrated weights using the response set are computed just as in the case of using the full sample as described above, where $F(\boldsymbol{\lambda}' \boldsymbol{x}_k)$ plays the role of an estimator for the unknown response probabilities $\hat{\delta}_k$ (cf. Särndal 2007, p. 115). So point 1 of the three desired porperties of auxiliary information used in weighting to fight potential nonresponse bias mentioned in section 3.1 is fulfilled here. The calibration procedure can be done in one step or in two steps. In the latter case first the response set $r$ is calibrated to the sample $s$ in order "to obtain *intermediate weights*" (Särndal & Lundström 2005, p. 82). In a scond step these intermediate weights are the input of the calibration from the response $r$ set to the population $U$. The resulting calibrated weights $w_k$ are very similar in both approaches as Särndal & Lundström (2005) point out (cf. Särndal & Lundström (2005), ch. 8.1) because they all use the same auxiliary information.

### 3.4.1 Distance functions for calibration estimators

As was shown in the previous section the function $F(\boldsymbol{\lambda}' \boldsymbol{x}_k)$ determines the adjustment of the design weights to gain calibrated weights $w_k$. The choice of the distance function $G(w_k, d_k)$ gives different possibilities of the values of the calibrated weights. One must distinguish if the g-weights shall all be positive, if too extreme weights have to be

[24] Cf. Särndal (2007), section 9.2.

avoided or trimmed, etc. The following methods describe different choices of distance functions and their properties.[25]

#### 3.4.1.1 Linear method:

This is the most basic of possible methods. It resembles a "chi-square distance" (Särndal & Deville 1992, p. 377):

$$G(w_k, d_k) = \frac{(w_k - d_k)^2}{d_k} \tag{3.43}$$

Derivation with respect to $w_k$ delivers a linear function of the fraction $w_k/d_k$:

$$\begin{aligned} \frac{\partial G(w_k, d_k)}{\partial w_k} = g(w_k, d_k) &= \frac{w_k}{d_k} - 1 \\ \Rightarrow F(\boldsymbol{\lambda}'\boldsymbol{x}_k) &= 1 + \boldsymbol{\lambda}'\boldsymbol{x}_k \end{aligned} \tag{3.44}$$

Evidently, negative g-weights are a possible result of the linear method. Consequently, this would also lead to negative calibrated weights, because the design weights are always positive. If these weight-adjustments of the design weights are modelled to compensate for unit nonresponse, then negative g-weights are impossible. This lies in the fact that the g-weight resembles $\delta_k$ which is strictly positive, because it is defined as the inverse of the unit response probability (which is also strictly positive).

The calibrated weights $w_k = d_k F(\boldsymbol{\lambda}'\boldsymbol{x}_k)$ can be calculated directly by inserting (3.44) into the calibration equations (3.41):

$$\begin{aligned} d_k F(\boldsymbol{\lambda}'\boldsymbol{x}_k) &= d_k(1 + \boldsymbol{\lambda}'\boldsymbol{x}_k) \\ \sum_{k\in s} d_k x_k(1 + \boldsymbol{\lambda}'\boldsymbol{x}_k) &= \sum_{k\in U} x_k \\ \sum_{k\in s} d_k x_k + \boldsymbol{\lambda}'\Big(\sum_{k\in s} d_k \boldsymbol{x}_k \boldsymbol{x}_k'\Big) &= \sum_{k\in U} x_k \\ \boldsymbol{\lambda}' &= \Big(\sum_{k\in U} x_k - \sum_{k\in s} d_k x_k\Big)' \Big(\sum_{k\in s} d_k \boldsymbol{x}_k \boldsymbol{x}_k'\Big)^{-1} \end{aligned} \tag{3.45}$$

[25] Cf. Särndal & Deville (1992), section 2.

The formula for calculating the values in $\boldsymbol{\lambda}'$ is the same as for the GREG estimator shown in (3.31) and thus the linear method in calibration weighting yields the GREG estimator.[26]

#### 3.4.1.2 Exponential method:

The distance function for calculating the g-weights is defined as follows (where "ln" denotes the natural logarithm):[27]

$$G(w_k, d_k) = w_k \cdot \ln\Big(\frac{w_k}{d_k}\Big) - w_k + d_k \tag{3.46}$$

Using the exponential method the undesired property of negative g-weights can be avoided, because the calculation of $F(\boldsymbol{\lambda}'\boldsymbol{x}_k)$ yields the exponential function:

$$\begin{aligned} \frac{\partial G(w_k, d_k)}{\partial w_k} &= g(w_k, d_k) = \ln(\frac{w_k}{d_k}) \\ &\Rightarrow F(\boldsymbol{\lambda}'\boldsymbol{x}_k) = \exp(\boldsymbol{\lambda}'\boldsymbol{x}_k) \end{aligned} \tag{3.47}$$

Clearly the resulting g-weights from (3.47) are all positive, so there will not be any negative adjusting weights as is possible with the linear method. However, one downfall of the exponential method is that it may lead to extremely high values of the g-weights and thus to unpractically high calibrated weights $w_k$ because of the exponential function. This can be a major challenge in estimation for comparatively small domains, a fact that also Särndal & Deville (1992) point out.[28]

#### 3.4.1.3 Logit Method

In order to avoid extremely large weights, which is especially possible with the exponential method, it may be helpful to formulate boundaries for the argument of the distance function, i.e. $L < w_k/d_k < U$, where $L$ designates the lower boundary and $U$ designates the upper boundary. Besides the "truncated linear method" (Sautory 2003, p. 10), which just introduces $L$ and $U$ as cut-off points for the linear methods shown

[26] This fact is also shown in Särndal & Lundström (2005), ch. 7.4.
[27] Cf. Särndal & Deville (1992), p. 378.
[28] Cf. Särndal & Deville (1992), section 2.

in subsection 3.4.1.1, a bounded calibration method can also be introduced for the exponential method presented in section 3.4.1.2. This bounded method is also called the "logit method" (Sautory 2003, p. 10), which is implemented in the macro CALMAR of the statistical software SAS.[29] The logit method is also used by Statistics Austria in the weighting process of EU-SILC in Austria.[30] For the logit method the choice of $L$ and $U$ should satisfy $L < 1 < U$. The distance function of the logit method is given by[31]

$$\begin{aligned} G(w_k, d_k) = (w_k/d_k - L) \cdot \ln\Big(\frac{w_k/d_k - L}{1 - L}\Big) + \\ + (U - w_k/d_k) \cdot \ln\Big(\frac{U - w_k/d_k}{U - 1}\Big) \end{aligned} \tag{3.48}$$

yielding g-weights in the form of

$$F(\boldsymbol{\lambda}'\boldsymbol{x}_k) = \frac{L(U-1) + U(1-L) \cdot \exp(A \cdot \boldsymbol{\lambda}'\boldsymbol{x}_k)}{(U-1) + (1-L) \cdot \exp(A \cdot \boldsymbol{\lambda}'\boldsymbol{x}_k)}. \tag{3.49}$$

$A$ is defined as the fraction $A = (U - L)/(1 - L)(U - 1)$. The modified g-weights of (3.49) are now bounded with $F(-\infty) = L$ and $F(\infty) = U$.

### 3.4.2 Generalized Calibration

As the title already proclaimes, generalized calibration uses a more general form of formulating the calibration equations.[32] Osier (2010) shows that the weights $w_k$ derived in generalized calibration do not rely on auxiliary information $\boldsymbol{x}_k$ although they fulfill the calibration property, hence

[29] Cf. Sautory (2003).
[30] Cf. Glaser & Till (2010).
[31] Cf. Särndal & Deville (1992), p. 378.
[32] Cf. Deville (2002), p. 6f.

$$\sum_{k \in s} w_k \boldsymbol{x}_k = \sum_{k \in U} \boldsymbol{x}_k$$

with

$$w_k = \frac{d_k}{m/n} F(\boldsymbol{\lambda}' \boldsymbol{z}_k) \tag{3.50}$$

where $m/n$ is the response rate and $\boldsymbol{z}_k$ denotes a vector of variables that explain unit nonresponse. The remarkable thing here is that $\boldsymbol{z}_k$ is only required to be known for the response set $r$, so a vast amount of survey variables may be used. However, the number of variables in $\boldsymbol{z}_k$ has to be the same as the number of calibration variables in $\boldsymbol{x}_k$.[33] In other words, if the dimensions of the vectors $\boldsymbol{x}_k$ and $\boldsymbol{z}_k$ were not equal, the corresponding calibration equations could not be solved.

Based on Deville (2002), Osier (2010) points out that "the generalized calibration is equivalent to the following two-step re-weighting process" (Osier (2010), p. 2):

1. Weighting adjustment of the design weights by the inverse of the true response probability $\delta_k$
2. Calibration according to the $I$ auxiliary variables known for the population $U$

Therefore generalized calibration promises to adjust for unit nonresponse and establish coherence with auxiliary population totals in one step with the possibility of using factors that explain unit nonresponse very well.

[33] Cf. Sautory (2003), p. 11.

# 4 Application and evaluation of different weighting procedures

In the previous chapter various methods for countering potential bias due to unit nonresponse have been presented. The important question now is which of these techniques are best suited for countering unit nonresponse bias in estimation of the parameter of interest described in section 2.1. Evidence from literature suggests that calibration has a high potential of reducing unit nonresponse bias considerably.[1] However, this is only true if the auxiliary information at hand is well chosen and can explain the response propensity as well as the most important study variables and domains (cf. section 3.1). In order to evaluate the practical implications of different weighting schemes on a real survey, the effects of unit nonresponse bias and methods to counter it are going to be evaluated using the example of the EU-SILC survey in Austria that took place in the year 2010.

## 4.1 EU-SILC in Austria

The EU-SILC survey has already been briefly mentioned throughout this thesis. This section is going to give a more thorough description of the survey, because it is going to play an important role in this chapter as the basis for simulations. An in-depth report of EU-SILC 2010 in Austria can be found in Till-Tentschert et al. (2011) and in Statistics Austria (2010). For a detailed description of EU-SILC for all participating countries please refer to Atkinson & Marlier (2010), Wolff et al. (2010) and Verma & Betti (2010). The following description of EU-SILC is based on the sources mentioned above.

[1] Cf. Särndal & Lundström (2005), Ch. 6.

"EU-SILC" is an acronym for "European Union Statistics on Income and Living Conditions"[2] which is a yearly statistic covering income and living conditions of private households in Europe. The main topics of EU-SILC are income, eomployment and living conditions as well as questions regarding subjective health, well-being and the financial situation. EU-SILC is one of the most important sources of social statistics in Europe and plays a central role in delivering indicators of poverty and social exclusion in the so-called "Europe 2020" growth strategy of the European Union.[3] The most important indicators of EU-SILC are the at-risk-of-poverty-rate and the rate of persons at-risk-of-poverty or social exclusion. The legal basis of EU-SILC are national regulations and EU-regulations.

The at-risk-of-poverty-rate is defined as the rate of persons living in households where the so-called "equivalised household income" (Till-Tentschert et al. 2011, p. 32) is below 60% of the national median of the equivalised household income. The equivalised household income is defined as the total disposible household income (i.e. net income) standardized by the consumption needs of all persons living in the household. These needs are defined by specific factors, which are equal to 1 for a grown-up person and 0.5 for each additional grown-up (i.e. a person aged at least 14 years) in a household. Children (up to 13 years of age) receive a factor of 0.3.[4] For example, the sum of the consumption equivalents of a household with two grown-ups and two children is 2.1.[5]

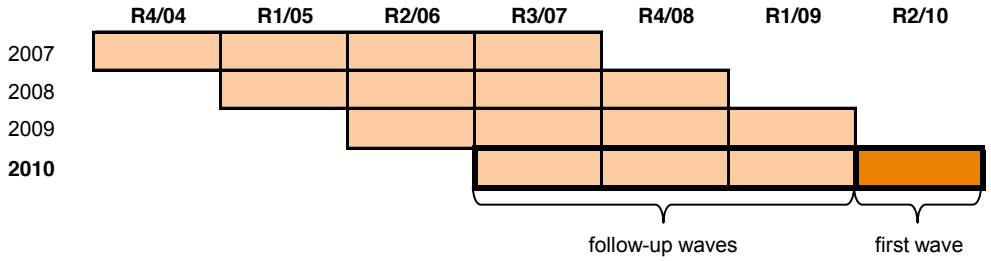
The rate of persons at-risk-of-poverty or social exclusion, which is also the main indicator of the Europe 2020 growth strategy, is defined as persons who are at-risk-of-poverty[6] and/or are materially deprived and/or live in "jobless" households.[7]

EU-SILC in Austria started in 2003 as a cross-sectional survey conducted by Statistics Austria. From 2004 onwards EU-SILC in Austria has been carried out as an integrated cross-sectional and longitudinal survey with a rotational sample structure consisting of four rotational subsamples. This means that about three quarters of the households surveyed in one year are being followed up in the next year. Each new survey year one quarter of the sample is renewed and hence comprises a newly selected subsample.

[2]Cf. http://epp.eurostat.ec.europa.eu/portal/page/portal/microdata/eu_silc (retrieved September 14, 2012).
[3]Cf. http://ec.europa.eu/europe2020/index_en.htm (retrieved September 14, 2012).
[4]Cf. Förster & D'Ercole (2009).
[5]Cf. Till-Tentschert et al. (2011), p. 32.
[6]The at-risk-of-poverty-rate indicator is one of three parts of the indicator of persons at-risk-of-poverty or social exclusion.
[7]For a more detailed description of this indicator please refer to Atkinson & Marlier (2010).

Therefore, the structure of the sample of EU-SILC is called a "rotating panel design" (Groves et al. 2004, p. 12). Each rotational subsample has a duration of four years. Figure 4.1 gives an impression of the structure of the EU-SILC rotational design.

**Figure 4.1:** *Rotating panel design of EU-SILC*




Source: Statistics Austria, EU-SILC 2010, unpublished results.

Each rotational subsample in figure 4.1 is labeled with the rotation number and the first year of the rotation. For instance R3/07 is rotation number 3 which started in 2007 and R2/06 is the rotation which commenced in 2006 and ended after four years in 2009. If regarded from the viewpoint of the survey year 2010, the EU-SILC sample comprises one subsample which has been followed up for the fourth wave (R3/07), one subsample that has been followed up for the third wave (R4/08), one subsample that has been followed up for the second wave (R1/09) and one newly selected first wave subsample (R2/10) in 2010.

The sampling frame of the new first wave sample selected each year is the central residence register (*Zentrales Melderegister* - ZMR) restricted to addresses of private households with at least on person 16 years or older with her/his main residence at the household. The sampling design of the first wave sample of 2010 was a single stage stratified simple random sample with disproportional allocation for each of the nine Austrian provinces. The disproportional allocation considered different expected unit response rates for each province, where the actual unit response rates of the first wave sample 2009 were taken as estimates for the unit response rates of the first wave sample 2010. Compared to the sampling fraction of a simple random sample of the whole country, provinces with a unit response rate expected to be lower than the average expected response rate received a slight oversample and provinces with a unit response rate expected to be higher than the average response rate received a slight undersample.

This strategy was used in order to gain efficiency of estimates based on the surveyed households (i.e. the net-sample).[8]

The gross sample $s$, i.e. the sample consisting of all selected elements, i.e. addresses of private households, of the first wave of EU-SILC 2010 consists of 3,430 addresses selected from the central residence register. 2,005 of these households successfully took part in the survey and therefore form the net sample.

In this thesis the first wave sample is of primary interest, because in the first year of the survey unit nonresponse is usually highest and potential bias due to unit nonresponse is hard to counter because of limited auxiliary information. As was already described in chapter 2, if the response process is not MCAR ("missing completely at random") all methods of countering bias due to unit nonresponse need some auxiliary information that is known for the whole population or at least for the gross sample. For the samples of the follow-up waves of EU-SILC, i.e. the waves after the first wave, information of the previous year's wave can be used for estimating the response probability which is needed for calculating the unit nonresponse weight $\nu_k$ in the two-phase approach to weighting (cf. section 3.2). However, for the first wave mainly information from the sampling frame and information about the type of building from the gross sample were availabe at the time of data editing of EU-SILC 2010.

With the calibration approach to weighting (cf. section 3.4) information from the survey can be used to fight bias in estimates as long as external marginal distributions of the variables that are believed to control for nonresponse exist. One potential drawback of this method is, that the external marginal distributions have to be known for the whole population (with the exception of generalized calibration).

## 4.2 Possibilities of bias evaluation

Based on the first wave sample of EU-SILC 2010, different methods of weighting in order to avoid or diminish bias due to unit nonresponse will be compared. This comparison could be carried out by using estimates and indicators of EU-SILC calculated with weights from different weighting procedures. However, definitely unbiased estimation is only possible with the gross sample and the design weights (cf. Horvitz-Thompson estimator, section 2.1). As soon as unit nonresponse occurs, some sort of bias will be

[8] For more information about the sampling design of the first wave sample of EU-SILC 2010 please refer to Statistics Austria (2010), ch. 2.1.

introduced when using unit nonresponse adapted weights and estimates based on the net sample. In order to compare estimates based on the gross sample (i.e. from respondents and nonrespondents) and the net sample (i.e. only from respondents) a parameter that is known for the whole gross sample is needed. As it was laid out in the previous section, the most relevant indicators of EU-SILC are the at-risk-of-poverty-rate and the rate of persons at-risk-of-poverty or social exclusion. Since these indicators can only be produced from data based on the interviews of the EU-SILC net sample, different, but related auxiliary variables have to be used for a comparison of the effictiveness of different weighting strategies in countering unit nonresponse bias.

For the survey year 2010 register data of some income survey variables of EU-SILC are available for the selected sample.[9] Among these income data available from registers, the register of wage taxes delivers reliable income data that cover the largest part of the overall household income.[10] From this register the sum of the net income from employment and old-age benefits aggregated on household level can be computed for the gross sample. These two income components combined are the most relevant part in the overall household income.[11] Among the availabe register data this quantity shows the strongest correlation with the overall household income based on data from interviews and the at-risk-of-poverty-rate as well as the rate of persons at-risk-of-poverty or social exclusion. In EU-SILC 2010 income variables from both interviews and register data are availabe. Income values from registers were not published, but Statistics Austria is still in the process of comparing results based on interviews and on registers. Therefore no detailed analysis of income from register data based on the gross sample of the first wave of EU-SILC 2010 will be presented in this thesis. Only the design-weighted sum of $y_k$ (i.e. total sum of income from employment and old-age benefits on household level) across all households in the gross and the net sample, i.e. the response set $r$, will be used as a benchmark for comparing estimators of $y_k$ based on different weighting strategies. The major advantage of register data is the fact that it is available for respondents as well as for nonrespondents. Hence the estimate of the parameter of interest $\theta$ (i.e. income from employment and old-age benefits on household level) based on the gross

[9] The usage of these data in a cryptographically secured way is covered by a national regulation. Cf. BGBl II 2010/277 *Einkommens- und Lebensbedingungen-Statistikverordnung* (ELStV). Accessable online: http://www.statistik.at/dynamic/wcmsprod/groups/b/documents/webobj/055277.pdf (retrieved September 1, 2012).

[10] For a detailed description of the household income in EU-SILC please refer to Till-Tentschert et al. (2011).

[11] Reliable register data for income from self-employment are not available for EU-SILC 2010.

sample can be compared with estimates facilitating different weighting strategies based on the net sample. Note that the estimate based on the gross sample is a realized value of the unbiased Horvitz-Thompson estimator.

In this thesis only the bias due to unit nonresponse is of interest, therefore only estimates based on the net sample and the gross sample will be compared. It is assumed that the unbiased estimate based on the gross sample using the design weights (Horvitz-Thompson estimator) is equivalent to the true value in the population (which is guaranteed on average over repeated samples).

However, just comparing results from one net sample with the corresponding gross sample will not give an indication of unit nonresponse bias. The bias is defined as the expected value of an estimator compared to the true value of the parameter that has to be estimated in the population (cf. section 2.1) and cannot be evaluated with one estimate. It is important to note that the bias of an estimator must not be confused with values of estimates which may be "off the mark" (Särndal et al. 2003, p.41). Recalling formula (2.5) in section 2.1 helps making this distinction clearer. Unbiasedness is evaluated by taking the expectation of an estimator and comparing it with the true parameter. For the Horvitz-Thompson estimator this expectation is calculated with respect to the random variable $T_k$, that is the sample membership indicator defined in (2.3). So the exptectation is taken over all samples that can possibly be drawn from the population. Where the estimator has an expectation based on a probability distribution, the estimate is a constant quantity that has no expectation. To be precise, it is incorrect to associate an estimate with the concept of bias. As written above, the value of the estimate can be more or less different compared to the true parameter, but this can be due to the distribution of realizations of the estimator or the fact, that the estimator indeed deviates on average from the true parameter and thus is biased.

In order to evaluate the unit nonresponse bias of estimators applying different weighting procedures, it would be necessary to calculate the expected value of all possible response sets for each estimator. Apparently this is not possible, because only one realized response set of the first wave of EU-SILC 2010 exists in the form of the net sample. If the true response probabilities $\delta_k$ were known exactly for all households $k$ at addresses selected in the gross sample, a repeated response process could be simulated by repeatedly selecting subsamples with selection probabilities equal to $\delta_k$. The next section will present a way of estimating the response distribution and using the resulting estimates of the response probability for simulating the response process.

## 4.3 Outline of a Monte Carlo simulation study based on the Austrian EU-SILC 2010 survey

The following Monte Carlo simulations are principally based on the concept presented by Särndal & Lundström (2005) who also used simulations based on the net sample of a Swedish survey of 965 clerical municipalities. The major conceptional difference between the simulations carried out by Särndal and Lundström and the simulations of the present thesis is, that in this thesis the parameter of interest is known for the gross sample. Starting with the actual gross sample of the first wave of EU-SILC 2010, the response process can be simulated based on the actual number of selected elements and subsampling starting with the net sample as in Särndal & Lundström (2005) is not necessary. Besides a larger sample size in the simulation another very important advantage here is that estimation bias for a repeating survey can be analyzed and current weighting procedures can be evaluated in a simulation that closely resembles the real survey situation.

In order to make repeated sampling from the 3,430 selected addresses possible, the unknown response probabilities $\delta_k$ had to be estimated. Since 165 addresses turned out to be ineligible to the population at the time of the interview (i.e. building is not existent, not a private household, empty dwelling or no person with main residence was present), the response probabilities of these elements were set to zero. Eleven households where the address could not be found or there was no access to the address were included in the set of possible respondents. So the response rate defined by 2,005 responding households and 3,265 households that where eligible to respond is 61.4%. This adjusted gross sample consisting of eligible households shall be referred to as $s^*$ with size $n^* = 3,265$.

For $s^*$ a logistic regression model with the dichotomous indicator of response $R_k$ (cf. ch. 2.2) as dependent variable was created. Assuming a "logit response distribution" (Särndal & Lundström 2005, p. 76) the response probabilities were estimated using the adjusted gross sample and the information of which household is a respondent or a nonrespondent based on the net sample. The response probabilities were then estimated with a logistic regression model where the dichotomous variable of response with the values 0 - 'nonrespondent' and 1 - 'respondent' was the dependent variable. Any variables that were known for the gross sample could be used as predictors as long as they played a significant role in explaining the response behaviour. These possible

predictor variables were available from the sampling frame (ZMR), the gross sample and register data merged with the gross sample using an anonymized key.[12]

All metric variables were converted into categorial variables and each category was recoded into a (0;1) dummy variable. The statistical software SAS automatically exludes redundant categories due to dummy coding in logistic regressions. This strategy makes it possible to account for effects of single categories. The 68 potential predictor variables (including redundant categories) were put into a logistic regression model with a stepwise backwards elimination algorithm using the statistical software SAS. The rule of remaining in the model for each predictor was defined as a significant value of Wald's Chi-Square statistic with $\alpha < 0.1$.

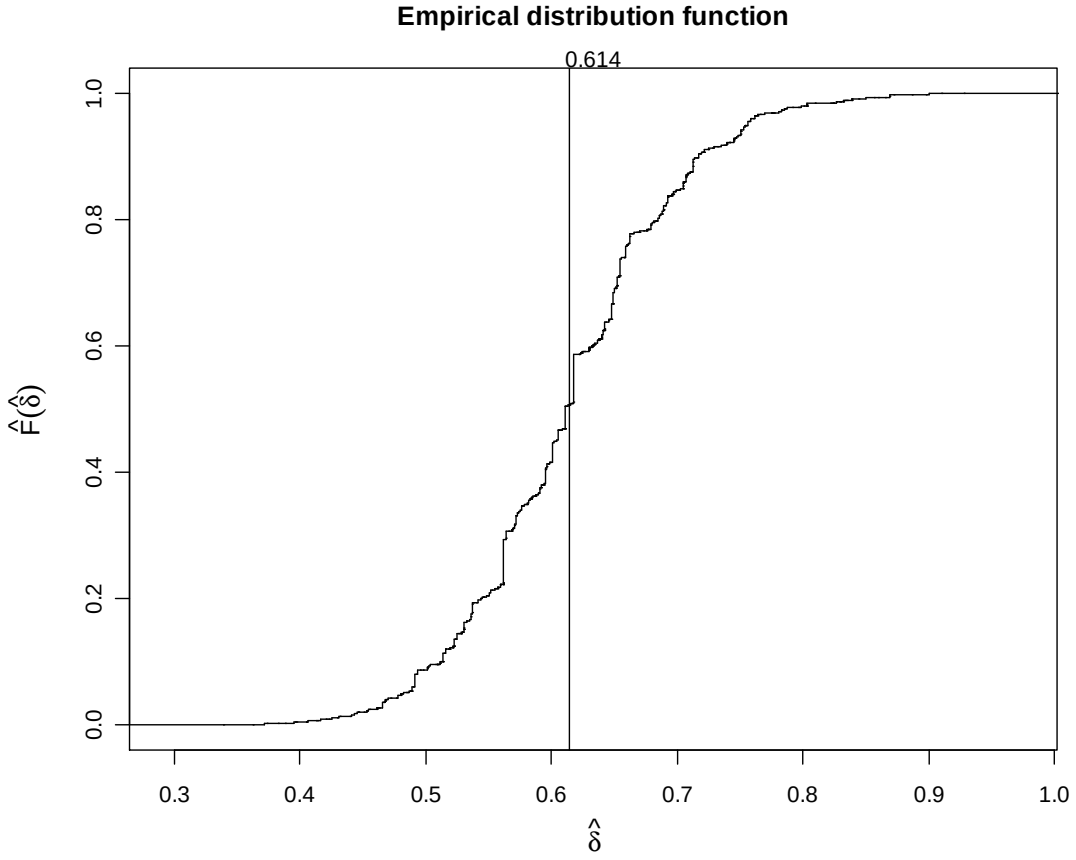
The resulting model reduced the predictor variables to 19 (binary) variables (i.e. $\boldsymbol{x} = (x_1, ..., x_{19})$. The most significant ones ($\alpha < 0.01$) referred to categories of the variables: Austrian provinces, interactions between degree of urbanisation and Austrian provinces, type of building and social status of the main earner. The corresponding likelihood ratio test produced a chi-square value of 99.0 with 19 degrees of freedom, meaning that the null hypothesis that the true coefficients in the population are all equal to zero can be rejected with high significance, i.e. $\alpha < 0.001$.

It is important to note here that the aim of the logistic regression model described above is not to find the best predictor variables for the response probability. Since the categorial predictor variables contain single categories in the form of dummy variables, interpretation with regard to the content of relationships between the predictors and the response rate are rather difficult. The backwise elimination of dummy variables is a technical procedure in order to reduce the big number of potential predictors. However, for simulating the response process, some form of response distribution has to be used that differs from a constant response probability, because the comparison of different weighting schemes for a missing at random (MAR) response process is the main topic of this thesis. The most important result of the logisitic regression are the estimated response probabilities $\hat{\delta}_k$. The mean value $\overline{\hat{\delta}_k}$ of $\hat{\delta}_k$ from the model is exactly 0.614 which is equal to the response rate defined by 2,005 responding households and the 3,265 households in the adjusted gross sample. The estimated values of the response probability are distributed rather symmetrically around the mean value and show only few very small or very large values. Figure 4.2 presents the empirical distribution

[12] The bPK (*bereichsspezifies Personenkennzeichen*) allows for a cryptographically secured data merging. Cf. http://www.stammzahlenregister.gv.at/site/5970/default.aspx (retrieved September 10, 2012).

function of the estimated response probabilities for the adjusted gross sample $s^*$.

**Figure 4.2:** *Empirical distribution function of the estimated response probability $\hat{\delta}$ for $s^*$*




The shape of the distribution presented in figure 4.2 shows a response distribution that refers to a response process that is not missing completely at random (MCAR). For the following simulations the values of the estimated response probabilities were fixed and treated as the real response probabilities $\delta_k$. The response probabilities for the 165 cases that are known to be ineligible where set to zero.

In order to be able to calculate the expected value of the parameter estimators weighted by each weighting procedure presented in chapter 3 the response process had to be simulated for the 3,430 households in the gross sample with size $n$. The event of response was treated as an additional sampling phase according to "Bernoulli sampling" (Särndal et al. 2003, p. 26) with $\delta_k$ as probabilities of selection. For each weighting strategy simulations were carried out by iterating the following procedure 10,000 times:

1. Generate a vector $\boldsymbol{\xi}$ of random numbers distributed uniformely on the [0, 1] interval by using the pseudo-random number generator of the statistical software SAS.
2. Select household $k$ belonging to the gross sample into the simulated response set, if its associated random number $\xi_k$ from step 1 is smaller than the response probability, i.e. $\xi_k < \delta_k$.
3. Apply a weighting procedure $\omega$ to the simulated response set selected in step 2.
4. Rescale the sum of weights from step 3 to the sum of design weights in the gross sample.
5. Calculate the weighted value of the parameter estimate of interest, i.e. $\hat{\theta}$.
6. Add the estimate $\hat{\theta}$ from step 5 as well as the sample size of the estimated response set to the vector $\hat{\boldsymbol{\theta}}$ of estimates.

After each iteration the resutling estimate of the iteration is stored in vector $\hat{\boldsymbol{\theta}}$. This means that only the sum of the variable of interest $y$, i.e. income from employment and old-age benefits from all households in the simulated response, (as well as the sample size) remains. The values $y_k$ of each household in the simulated response set are deleted after each iteration. The rescaling in step 3 is necessary in order to make the expected values of the parameter estimates comparable. Since the benchmark for comparison is the design-weighted sum of the variable of interest, i.e. sum of income from employment and old-age benefits, the sum of weights resulting from the weighting procedure $\omega$ should be the same in order to refer to the same population size. For example, during the testing of simulation methods it turned out that the expected values of the calibration estimators were quite high. This was due to the fact that the marginal distributions used in calibration stemmed from a population of households that is a bit larger than the one gained from the design-weighted estimate from the gross sample that is used as a point of reference in this thesis. This estimate can be "off the mark" (cf. section 4.2), because it is the result of the random event of sample selection. Another reason lies in the fact that the sampling frame is inevitably a bit smaller than the true size of the population of private households. Since EU-SILC applies a rotational design, households that have been in the sample for the last four years or were in the sample five years ago are exluded from the sampling frame. However, at this point it is important to remember that the main topic of this thesis is bias due to unit nonresponse. Bias that refers to sampling errors or frame errors is not interest, only bias that occurs because of the event of unit nonresponse is of importance.

After the simulation has finished the vector $\hat{\boldsymbol{\theta}}$ consists of 10,000 entries with a column of sample sizes and a column with estimates $\hat{\theta}_i$ of the parameter of interest. Since the simulated response sets were created by Bernoulli sampling the sample sizes follow a binomial distribution with parameters $n$ and $\delta_k$. The expected value of the estimator $\hat{\theta}$ is the mean value of the 10,000 estimates in $\hat{\boldsymbol{\theta}}$.[13]

$$E_{sim}(\hat{\theta}) = \sum_{i=1}^{10,000} \frac{\hat{\theta}_i}{10,000} \tag{4.1}$$

Referring to the definition of the Bias presented in (2.18) the unit nonresponse bias of a specific weighting strategy $\omega$ can be written in the following way (where $E_T$ denotes the expectation referring to the gross sample and $E_{sim}$ denotes the expectation of the estimator of a specific weighting strategy based on the simulated response sets):

$$B_{TRsim}(\hat{\theta}_\omega) = E_{sim}(\hat{\theta}_\omega) - E_T(\hat{\theta}_{HT}) = E_{sim}(\hat{\theta}_\omega) - \theta \tag{4.2}$$

In the equation above $\hat{\theta}_{HT}$ denotes the Horvitz-Thompson estimator which is unbiased for the gross sample $s$. For the sake of a more clear depiction of the size of the bias, the relative bias, which was already described in section 2.3, will be used for the comparison of the estimators in the simulations (see also formula (2.20)):

$$RB_{TRsim}(\hat{\theta}_\omega) = \frac{B_{TRsim}(\hat{\theta})}{\theta} = \frac{E_{sim}(\hat{\theta}_\omega) - \theta}{\theta} \tag{4.3}$$

As a measure of variation for the estimators from the simulations the "simulation variance" (Särndal & Lundström 2005, p. 79) is being used:

$$V_{sim}(\hat{\theta}_\omega) = \frac{1}{9,999} \sum_{i=1}^{10,000} \left(\hat{\theta}_{\omega(i)} - E_{sim}(\hat{\theta}_\omega)\right)^2 \tag{4.4}$$

An estimator that has a small bias does not necessarily have to incorporate a small variance. In order to compare the estimators used in the simulation study regarding their bias and their simulation variance at the same time, the simulation mean-squared error (MSE) will be used. The MSE can be calculated as the sum of the the variance of the estimator and the squared bias of the estimator:

[13] Cf. Särndal & Lundström (2005), p. 78.

$$MSE_{sim}(\hat{\theta}_\omega) = V_{sim}(\hat{\theta}_\omega) + \left(B_{TRsim}(\hat{\theta}_\omega)\right)^2 \tag{4.5}$$

## 4.4 Results of the Monte Carlo simulation study

The simulations were carried out in accordance with the weighting procedures presented in chapter 3. The expected values of the estimators of each simulation were compared with the value of the design-weighted estimate.

The results show that the relative bias is largest, if only the straight expansion estimator is used. Obviously, assuming a response process with equal probabilities of response (MCAR) does not seem to be valid, because this method produces estimates that are on average about 4.1% off the design-weighted value of the parameter of interest.

If weighting cells according to a grouping of the nine Austrian provinces is introduced, the relative bias drops slightly to 3.8%. This is not surprising since the provinces were also used as predictors in the logistic regression that estimated the response probabilities which are the basis of the simulation. The weighting cells method, also denoted here by "RHG" (response homogeneity groups) facilitates unit nonresponse weights $\nu_h$ that are simply the inverse of the response rates of each province. These unit nonresponse weights are multiplied by the design weights to produce weights which are used to estimate the sum of income from employment and old-age benefits with data from the simulated net samples.

If the inverse of the response probabilities estimated by a logistic regression model are used instead as unit nonresponse weights then the relative bias becomes noticeable smaller in comparison to the straight expansion estimator. This time also the simulation variance gets notably smaller. This is not surprising since only Austrian provinces were used as a response homogeneity groups and the logistic regression uses far more variables than the weighting cells. In the case of weighting by the inverse of the response probability estimated by a logistic regression two models were formulated which delivered similar results, i.e. a relative bias of about 3.1% and 3.0%.

Neural networks (NNW) were never before used for estimating the response probabilities for the Austrian EU-SILC survey and were introduced as a method for estimating response probabilities in the two-phase approach to weighting. As in the case with the two-phase approach that uses logistic regression models in order to estimate response

probabilities, the inverse of these estimated response probabilities are multiplied by the design weights in order to get weights that are adjusted for unit nonresponse. The application of neural networks for unit nonresponse weights resulted in a rather big reduction of bias as compared to the logistic regression models mentioned above. The relative bias drops to roughly 2% if all relevant auxiliary information is used as predictor variables and there is an appropriate number of hidden units in the hidden layer. However, simulations show that the variance is larger than the one resulting from the usage of logistic regressions.

The utilization of some sort of calibration for computing weights with the aim of countering unit nonresponse bias of estimators based on the response set results in a big reduction of unit nonresponse bias. The results of the simulations show that this methods delivers the lowest relative unit nonresponse bias and also shows the lowest variance. This advantage of the calibration method was indicated in related literature[14] and is also evident in the simulation results of the calibration methods used in this thesis. A comparison of the relative bias of the calibration estimators from simulations shows that the biases of the estimators only differ slightly (ranging from -0.76% to -0.74%). Also the corresponding variances are roughly of the same size. Stricter boundaries for the truncated methods could result in a smaller variances for the logit and the truncated linear method, but would also produce a larger amount of bias for the estimators that result from these weighting methods. Using the inverse of the estimated response probabilities (estimated by logistic regression models) multiplied by the design weights as input weights in the calibration only slightly changes the bias, as can be seen in the simulation results of these methods. The variances of these estimators are only a little bit larger. Comparing estimators with calibrated weights and estimators without calibration input weights from logistic regressions, adjusting the design weights with unit nonresponse weights before calibration only delivers a small enlargement of bias and variance. This may be due to the fact that mostly variables from the sampling frame could be used as predictors in the logistic regression models.

The following subsections give a more detailed insight in the results of the simulation study. For a theoretical presentation of the estimator please refer to ch. 3.

[14] Cf. Särndal & Lundström (2005), ch. 6.3.

### 4.4.1 Straight expansion estimator

If it is assumed that unit nonresponse is happening completely at random (cf. MCAR response process in section 2.2) then no sophisticated weighting scheme has to applied in order to adjust for the units lost because of unit nonresponse. The design weights of the elements remaining in the response set $r$ (of size $m$) just have to be rescaled to the selected sample $s$ (of size $n$) by multiplying the design weights by the inverse $n/m$ of the overall response rate in order to allow the weights compensate for the lost units. However the strong assumption of MCAR can presumably never be justified in practice and the results of the simulation show that it is the estimator weighted by the expansion weight which is the most biased of all (cf. table 4.1).

**Table 4.1:** *Simulation results - straight expansion estimator*

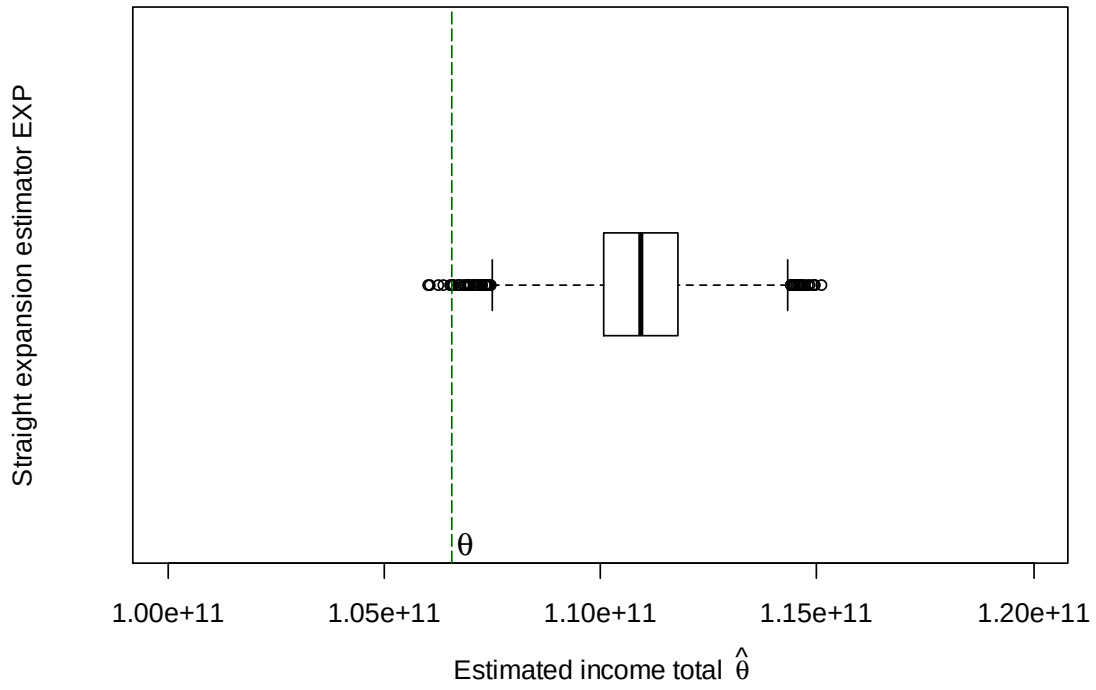
| **Weighting procedure** | **Estimator** | **Rel. Bias in %** | **Variance** | **MSE** |
|---|---|---|---|---|
| Inverse response rate $\nu = n/m$ | $\hat{\theta}_{EXP}$ | 4.07 | $1.61 \cdot 10^{18}$ | $2.04 \cdot 10^{19}$ |

Source: Statistics Austria, EU-SILC 2010, unpublished results.

The large relative bias and the rather large variance together with the large MSE show that simply inflating the design weights by the inverse of the overall response rate is insufficient in order to achieve design weight adjustments for EU-SILC that counter unit nonresponse.

The 10,000 estimates from the simulations are symmetrically distributed around the mean and the mean value of the estimates, i.e. the expectation of the estimator. More specifically, the estimated values of the straight expansion estimator approximately follow a normal distribution with $\mu = 1.109 \cdot 10^{11}$ and $\sigma = 1.268 \cdot 10^{9}$ (Kolmogorov–Smirnov test: p-value $> 0.15$). However, as figure 4.3 shows, the straight expansion estimator also heavily over-estimates the "true value" of the parameter $\theta$ computed from the design-weighted estimate from the gross sample which is indicated by the dashed vertical line.

**Figure 4.3:** *Simulation distribution of the straight expansion estimator*




Source: Statistics Austria, EU-SILC 2010, unpublished results.

### 4.4.2 Two-phase weighting: weighting cells estimator

A simple method of adjusting for unit nonresponse without a complicated model can be achieved by using the inverse of the response rates of certain groups. These unit nonresponse weights then only have to be multiplied by the design weights in order to gain an estimator with unit nonresponse adjusted weights. This method described in section 3.2.2 is the second of the broader set of weights based on the two-phase approach to weighting introduced in section 3.2. Under the assumption of homogenous response probabilities within certain groups, appropriate unit nonresponse weights can be found if the response process really only differs between those groups. For this simulation the groups were chosen to be the nine Austrian provinces, since it is known from EU-SILC that the response rates differ by province. However, the bivariate correlations (Pearson's correlation coefficient) between the dummy-coded variables for Austrian provinces and the response indicator $R_k$ is on average quite low with a value of about 0.05 for the

significant correlations ($\alpha = 0.05$) for the provinces Lower Austria, Upper Austria, Vorarlberg and Vienna. So the building of response homogeneity groups according to provinces will not achieve a big reduction of bias.

The simulations show that for EU-SILC the weighting cells estimator is less biased than the straight expansion estimator, but has a variance that is a little larger. Overall the MSE drops by about 13% as compared to the straight expansion estimator which shows that weighting cells adjustments deliver an improvement in the precision of the estimator even if only big response homogeneity groups are formed and the variable responsible for the grouping is not highly correlated with the study variable of interest.

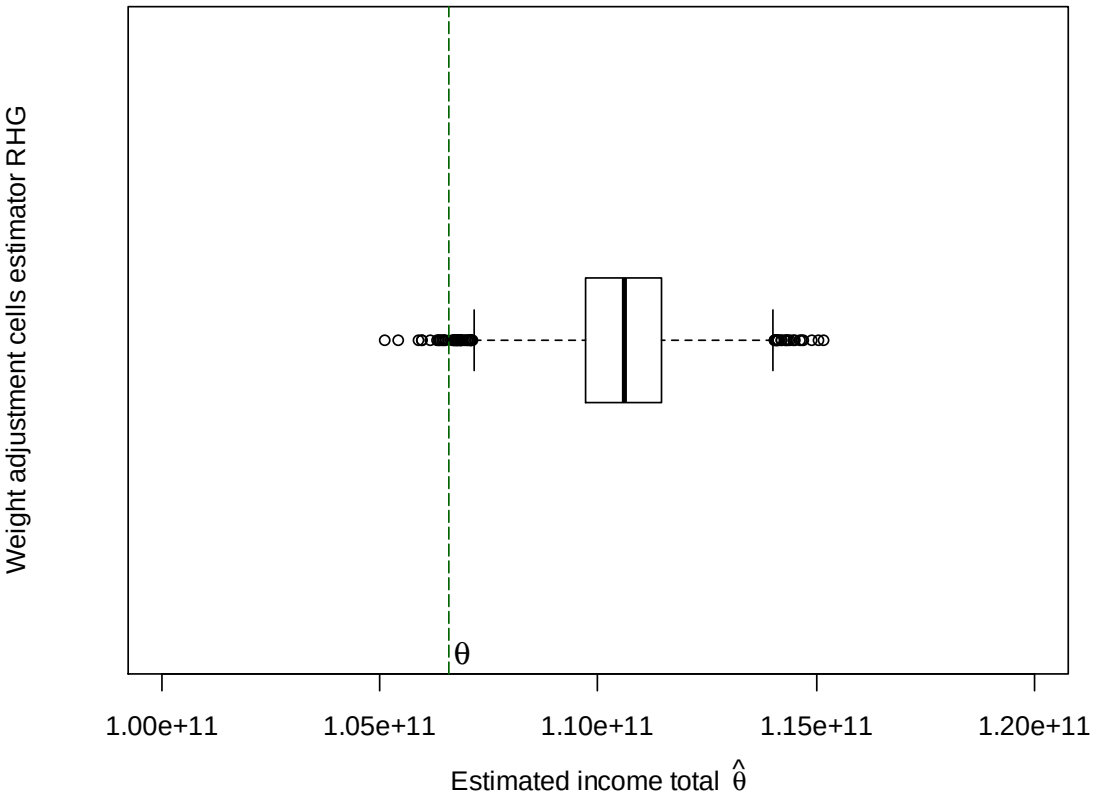
**Table 4.2:** *Simulation results - weighting cells estimator*

| **Weighting procedure** | **Estimator** | **Rel. Bias in %** | **Variance** | **MSE** |
|---|---|---|---|---|
| Inverse response rate per group h $\nu_h = n_h/m_h$ | $\hat{\theta}_{RHG}$ | 3.76 | $1.63 \cdot 10^{18}$ | $1.77 \cdot 10^{19}$ |

Source: Statistics Austria, EU-SILC 2010, unpublished results.

Figure 4.4 also shows the rather large discrepancy between the true parameter value and the mean value from the simulation.

**Figure 4.4:** *Simulation distribution of the weight adjustment cells estimator*




Source: Statistics Austria, EU-SILC 2010, unpublished results.

### 4.4.3 Two-phase weighting: Logistic regression

Instead of using weight adjustment cells for finding unit nonresponse adjustment weights, logistic regression models were used to estimate the response probabilities in each iteration of the simulation. Two logistic regression models with a slightly different set of predictor variables were applied in two separate simulations. Model 1 employed dummy-coded variables of the nine Austrian provinces, the degree of urbanization according to three levels, the type of building, the number of men, women, children and teens per household as well as the minimum and maximum age in five age categories per household and the information if there is at least one person with foreign citizenship registered at the household as predictor variables. All these variables were taken from the sampling frame and were therefore known for the whole gross sample. Logistic regression model 2 additionally took interaction effects between provinces and the

degree of urbanization into account. Table 4.3 shows a listing of the variables available. Dummy-coding was carried out for each category of the variables and categories with less than 50 households were aggregated in order to avoid dummy variables with too few cases.

**Table 4.3:** *Variables of logistic regression models in simulations*

| **Model 1** | **Model 2** | **Number of categories** |
|---|---|---|
| *Characteristics of the household address* | | |
| Austrian provinces | Austrian provinces | 9 |
| Degree of urbanisation | Degree of urbanisation | 3 |
| Type of building | Type of building | 4 |
| | Austrian provinces × degree of urbanisation | 36 |
| | Austrian provinces × degree of urbanisation | 27 |
| *Characteristics of persons aggregated on household level* | | |
| Number of females | Number of females | 4 |
| Number of males | Number of males | 4 |
| Number of children | Number of children | 3 |
| Number of teens | Number of teens | 2 |
| Number of foreign Citizens | Number of foreign Citizens | 2 |
| Minimum age | Minimum age | 5 |
| Maximum age | Maximum age | 5 |

The difference between model 1 and model 2 lies in the usage of interactions between characteristics of the household address. The combination of provinces with degree of urbanisation or type of building allows for a more detailed inclusion of regional predictors. The household size was omitted as a predictor because it clearly interacts with the variables "number of females", "number of males", "number of children" and "number of teens". Explicitly Including interaction effects between household size and these interaction effects would have also resulted in dummy variables with categories that were too small.

For the prediction of the response probability in each iteration a stepwise backwards elimination algorithm was used. The condition for remaining in the model was defined as having a significant effect in the model as measured by Wald's chi-square statistic ($\alpha < 0.1$). The bias is noticeable lower in the case of using logistic regressions compared to the estimators in the previous two subsections, but still considerably high. Also the variance drops slightly which, togehter with the bias, is recognizeable in the MSE, which also falls perceptibly. Table 4.4 compares the results for model 1, model 2 and model 2 including trimming.

**Table 4.4:** *Simulation results - logistic regression adjustments*

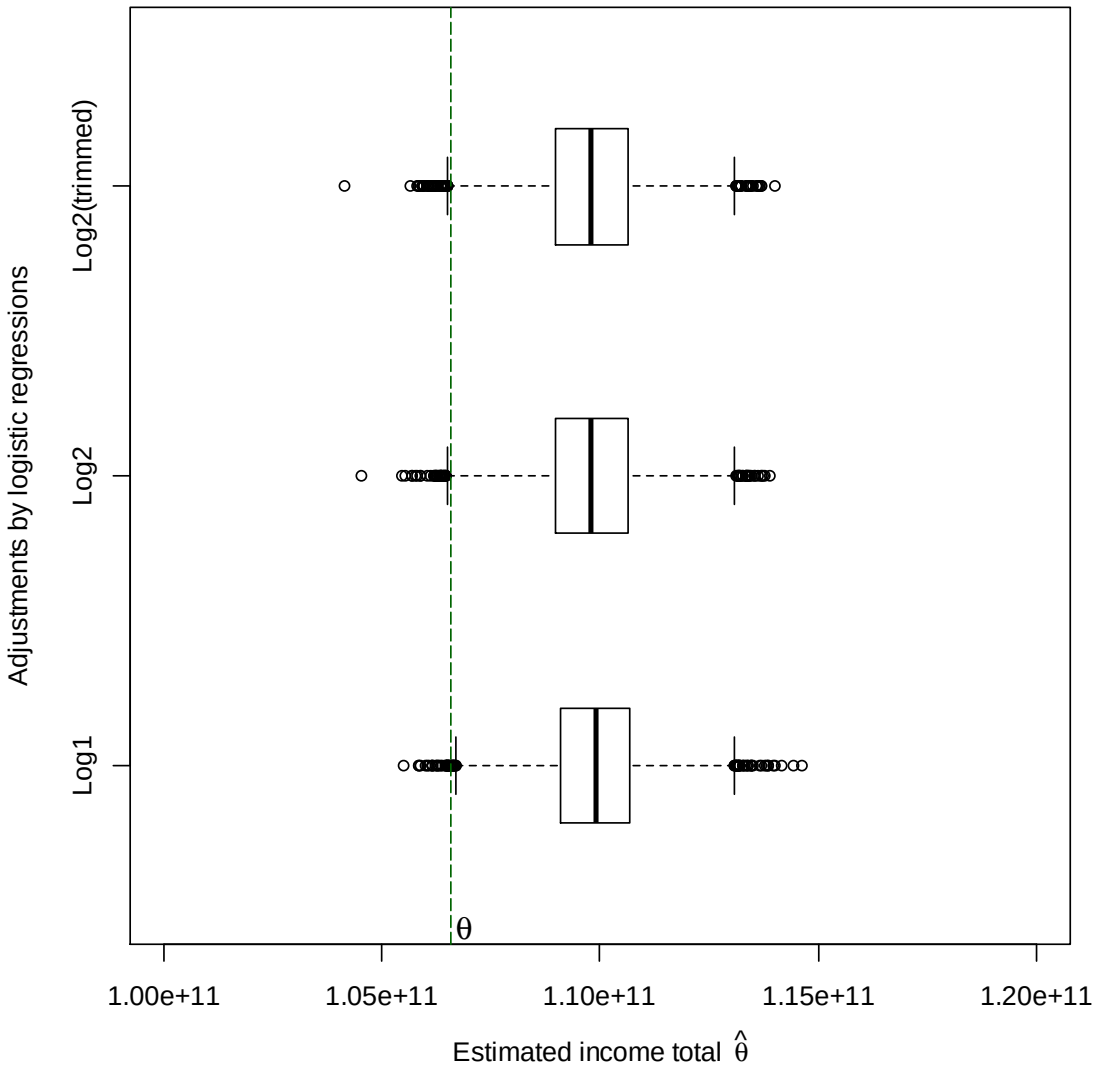
| **Weighting procedure** | **Estimator** | **Rel. Bias in %** | **Variance** | **MSE** |
|---|---|---|---|---|
| Estimated response probability from logistic regression Model 1 $\nu = 1/\hat{\delta}$ | $\hat{\theta}_{LOG_1}$ | 3.10 | $1.43 \cdot 10^{18}$ | $1.24 \cdot 10^{19}$ |
| Estimated response probability from logistic regression Model 2 $\nu = 1/\hat{\delta}$ | $\hat{\theta}_{LOG_2}$ | 3.01 | $1.49 \cdot 10^{18}$ | $1.18 \cdot 10^{19}$ |
| Estimated response probability from logistic regression Model 2 $\nu_k^{(low)}$, $\nu_k^{(high)}$ | $\hat{\theta}_{LOG_2}$(trimmed) | 3.02 | $1.49 \cdot 10^{18}$ | $1.18 \cdot 10^{19}$ |

Source: Statistics Austria, EU-SILC 2010, unpublished results.

Apart from outliers the two logistic regression models used in this version of two-phase weighting do not differ much. Using more predictor variables in the logistic regression

(as in model 2) only produces small changes in bias and variance. As was expected for the method of trimming the unit nonresponse weights (cf. ch. 3.2.3), trimming introduces a new source of bias. However, relative bias only grows by a very small amount (from 3.01% to 3.02%). Variance drops by a very small amount from $1.493 \cdot 10^{18}$ to $1.489 \cdot 10^{18}$, the MSE rises by a very small magnitude from $1.182 \cdot 10^{19}$ to $1.183 \cdot 10^{19}$.

**Figure 4.5:** *Simulation distribution of the two-phase estimator with nonresponse weights from logistic regressions*




Source: Statistics Austria, EU-SILC 2010, unpublished results.

### 4.4.4 Two-phase weighting: Neural networks

The usage of neural networks (NNW) in the two-phase approach to adjusting the design weights to unit nonresponse was introduced in section 3.2.4. In this thesis neural networks were tried out as an alternative to logistic regressions to find estimates of response probabilities, where the inverse of the response probabilities are used as unit nonresponse weights to adjust design weights.

Estimating Bayesian a posteriori probabilities with NNW has been in practice for about two decades. Liang & Yung Cheung Kuk (2004) applied neural networks for the estimation of the population mean (for finite populations). Their results show that neural networks have a smaller MSE than linear regression methods. Amer (2009) uses neural networks to incorporate characteristics of complex survey designs in imputation and thus adjusting estimators for item nonresponse. In this thesis neural networks are solely used for adjusting the design weight to counter unit nonresponse bias. Other characteristics of the survey, e.g. the sampling design, are not considered by neural networks here.

The results from the simulations show a considerably smaller bias for some versions of the NNW. It seems that the number of units in the single hidden layer of the feed-forward neural network is the key to a small bias. It seems that the larger the number of hidden units the smaller the relative bias due to unit nonresponse. This is not too surprising since more hidden units also make it possible for the network to find more diverse relations between the input units and the output.

The simulations applying neural networks was implemented in the statistical software R with the package "nnet".[15] The input variables were similar to the ones used for the logistic regressions. This time also the household size was taken into account, because for the neural networks no dummy variables were introduced and all variables that relate to the size of the household were included as metric variables in the neural networks. This is also true for the other variables of personal characteristics that were aggregated on household level, such as number of females and males as well as number of children, teens and foreign citizens. Since the neural networks allow for the approximation of interactions through the connections established by the hidden layer, no interaction variables were explicitly entered in the input. Altogether three sets of variables were used as inputs in the neural network. Input set 1 comprises the

[15] Cf. cran.r-project.org/web/packages/nnet/ (retrieved August 27, 2012).

variables of the logistic regressions plus the household size. The smaller input set 2 only contains characteristics of the household address and input set 3 adds the minmum and maximum age of household members to set 2. Table 4.7 presents the auxiliary variables used by the different input sets of the neural networks.

**Table 4.5:** *Variables of the neural network input*

| **Input set 1** | **Input set 2** | **Input set 3** |
|---|---|---|
| *Characteristics of the household address* | | |
| Austrian provinces | Austrian provinces | Austrian provinces |
| Degree of urbanisation | Degree of urbanisation | Degree of urbanisation |
| Type of building | Type of building | Type of building |
| *Characteristics of persons aggregated on household level* | | |
| Household size | | |
| Number of females | | |
| Number of males | | |
| Number of children | | |
| Number of teens | | |
| Number of foreign Citizens | | |
| Minimum age | | Minimum age |
| Maximum age | | Maximum age |

During the preparations of the simulations it turned out that sometimes the neural networks produce extremely large nonresponse adjustment weights. It was decided to truncate these weights. In the first simulation procedure applied a maximum value of

100 for the nonresponse adjustment weights was allowed. In the further simulations this maximum was always fixed at a value of 10, because in the response distribution used for the simulation there are no response probabilities greater than zero below this value. After the neural networks were fitted, the unit nonresponse weights $\nu$ calculated from the inverse of the estimated response probabilities where then multiplied by the design weights and the resulting weights were used for calculating the estimates $\hat{\theta}_{NNW_i}$ $(i = 1, ..., 8)$.

The results of the simulations presented in table 4.6 show a reduction of the bias for almost all estimators using design weight adjustments by neural networks as compared to logistic regressions. For the neural networks using input set 1 (cf. table 4.5) the following tendency is regonizable: the more hidden units are represent in the hidden layer, the smaller the bias gets. However, with a big number of hidden units, the risk of occasionally computing estimates with extreme values gets higher. This can be seen in the distribution of outliers for $\hat{\theta}_{NNW_5}$ and $\hat{\theta}_{NNW_6}$. One downfall of the method is a slightly increased variance compared to the results of the preceding section, but this does not affect the MSE as greatly as the smaller bias and thus the MSE drops almost by half in comparison with logistic regressions. The number and sort of input variables is also an important factor of the precision of the resulting estimator. The usage of input set 2 leads to an estimator ($\hat{\theta}_{NNW_7}$) that is less precise than $\hat{\theta}_{LOG_1}$ and $\hat{\theta}_{LOG_2}$. Using input set 3 results in the estimator $\hat{\theta}_{NNW_8}$, which is more precise than $\hat{\theta}_{NNW_7}$ because of more auxiliary variables in the input set which also have the capacity of more accurately explaining the response rate and the survey variable of interest.

**Table 4.6:** *Simulation results - adjustments by neural networks*

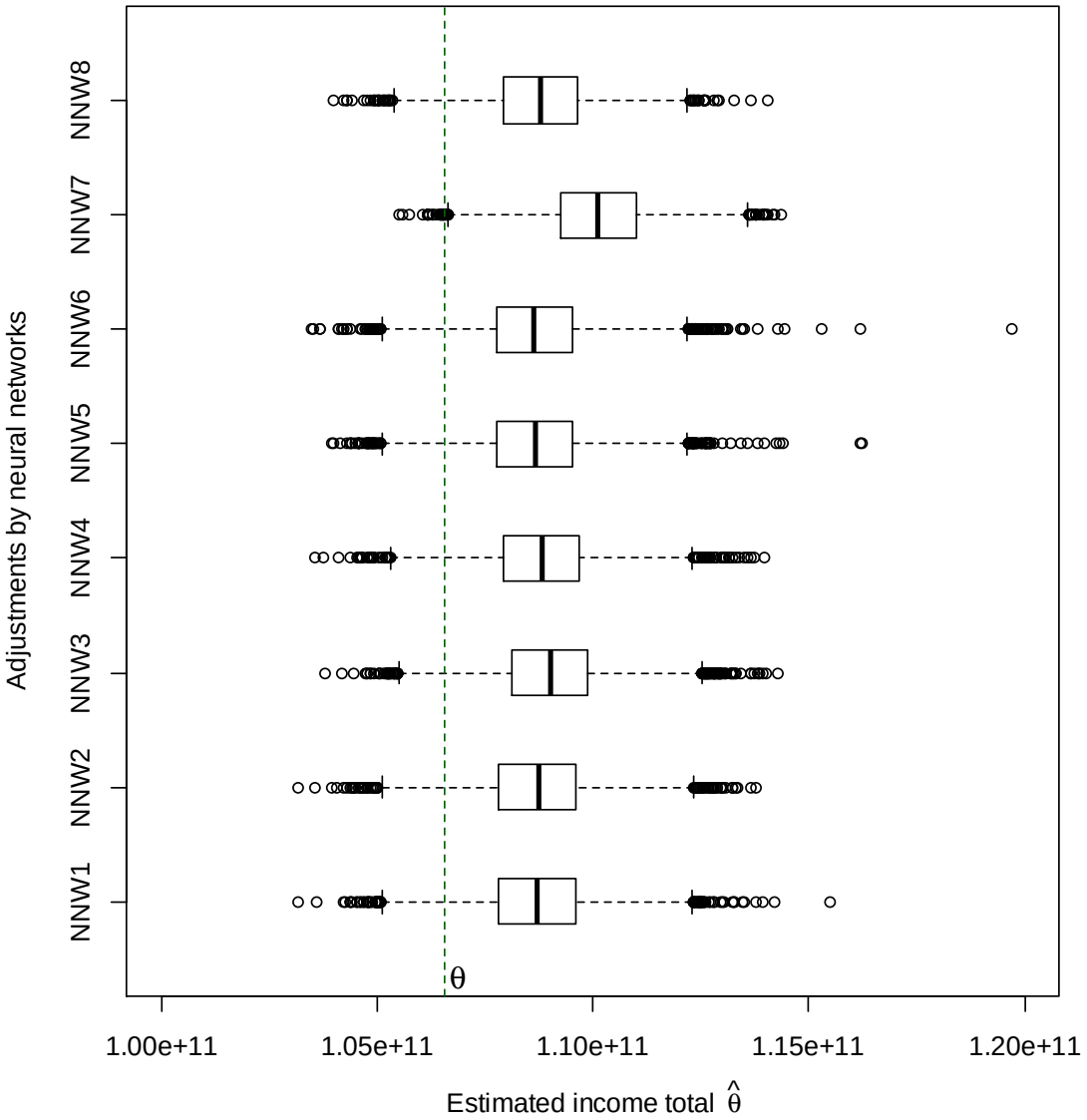
| Weighting procedure | Estimator | Rel. Bias in % | Variance | MSE |
|---|---|---|---|---|
| NNW1: Input set 1<br>9 hidden units<br>$max(\nu) = 100$ | $\hat{\theta}_{NNW_1}$ | 1.99 | $1.79 \cdot 10^8$ | $6.30 \cdot 10^{18}$ |
| NNW2: Input set 1<br>9 hidden units<br>$max(\nu) = 10$ | $\hat{\theta}_{NNW_2}$ | 2.01 | $1.78 \cdot 10^{18}$ | $6.36 \cdot 10^8$ |
| NNW3: Input set 1<br>3 hidden units<br>$max(\nu) = 10$ | $\hat{\theta}_{NNW_3}$ | 2.29 | $1.71 \cdot 10^{18}$ | $7.68 \cdot 10^{18}$ |
| NNW4: Input set 1<br>6 hidden units<br>$max(\nu) = 10$ | $\hat{\theta}_{NNW_4}$ | 2.10 | $1.70 \cdot 10^{18}$ | $6.72 \cdot 10^{18}$ |
| NNW5: Input set 1<br>10 hidden units<br>$max(\nu) = 10$ | $\hat{\theta}_{NNW_5}$ | 1.96 | $1.79 \cdot 10^{18}$ | $6.14 \cdot 10^{18}$ |
| NNW6: Input set 1<br>11 hidden units<br>$max(\nu) = 10$ | $\hat{\theta}_{NNW_6}$ | 1.94 | $1.81 \cdot 10^{18}$ | $6.08 \cdot 10^{18}$ |
| NNW7: Input set 2<br>9 hidden units<br>$max(\nu) = 10$ | $\hat{\theta}_{NNW_7}$ | 3.31 | $1.64 \cdot 10^{18}$ | $1.41 \cdot 10^{19}$ |
| NNW8: Input set 3<br>9 hidden units<br>$max(\nu) = 10$ | $\hat{\theta}_{NNW_8}$ | 2.07 | $1.53 \cdot 10^{18}$ | $6.42 \cdot 10^{18}$ |

Source: Statistics Austria, EU-SILC 2010, unpublished results.

In figure 4.6 the different distributions of the eight estimators using neural networks presented in table 4.6 show the main challenge of this method: although the estimates are symmetrically distributed around the mean value and the bias is quite low, eg. for

$\hat{\theta}_{NNW_1}$ or $\hat{\theta}_{NNW_2}$ the outliers are quite widely distributed. This depends especially on the number of hidden units. $\hat{\theta}_{NNW_5}$ or $\hat{\theta}_{NNW_6}$ have the biggest numbers of hidden layers and also produce the most extreme estimates.

**Figure 4.6:** *Simulation distribution of two-phase weighting using NNW*




Source: Statistics Austria, EU-SILC 2010, unpublished results.

It is obvious that neural networks have a big potential for usage in adjusting design

weights for unit nonresponse, but further inquiry is necessary to deal with the problem of seldomly gaining extreme values for weights by this method. However, using neural networks for calculating unit nonresponse adjusted weights has to be handled with care, because during the 10,000 iterations of the simulations, on a few occasions, extremely large or small weights were produced which can result in extreme values of the estimates. This can also be seen in the enlarged simulation variance of this method. Trimming the nonresponse weights can confront this problem, but still large variance and occasionally very large or small estimates come as a downfall of this otherwise promising method.

### 4.4.5 Calibration estimators

A different approach to weighting in order to counter unit nonresponse than the estimators shown in the previous four subsections are calibration estimators which were laid out in section 3.4. The GREG estimator of section 3.3 is a special case of calibration (with a linear distance function for the g-weights) and therefore evaluated together with the other calibration estimators.

The external marginal distributions to which the weights had to be calibrated were taken from the yearly average of the four quarters of the Austrian Microcensus[16] of 2010. The weights were calibrated to the following marginal distributions: Nine Austrian provinces, degree of urbanisation in three categories, household size in four categories, number of men and women per household according to four age categories and number of Austrian or foreign citizens. Table 4.7 presents the variables for which the marginal distributions were calibrated to the ones of the Austrian Microcensus. Categorial variables are described by the number of categories, metric variables are indicated by a hyphen, i.e. '-'.

[16]The Austrian Microcensus is a survey carried out quarterly by Statistics Austria and consists of about 22,500 private households selected from a probability sample. Cf. Kytir & Stadler (2004).

**Table 4.7:** *Marginal distributions used for calibration*

| Calibration | | Number of categories |
|---|---|---|
| *Characteristics of the household address* | | |
| Austrian provinces | | 9 |
| Degree of urbanisation | | 3 |
| *Characteristics of persons aggregated on household level* | | |
| household size | | 4 |
| Number of females | age<13 | - |
| | 14<age<34 | - |
| | 35<age<64 | - |
| | age>65 | - |
| Number of males | age<13 | - |
| | 14<age<34 | - |
| | 35<age<64 | - |
| | age>65 | - |
| number of Austrian citizens | | - |
| number of foreign citizens | | - |

One major difference of calibration as compared to the other weighting procedures presented in the preceding subsections is the fact that information from an external statistic is used to compute estimators which should compensate for unit nonresponse bias. These external marginal distributions are known to be representative for the population of private households in Austria. Therefore, after calibration, the marignal distributions of the weighted variables from the response set are euqal to the ones in the external source. "Type of building" cannot be used any more as auxiliary variable, because no appropriate auxiliary variable from external sources was available.

The solving of the calibration equations and the calculation of the g-weights was carried out with the SAS macro CALMAR which was developed by the French national

statistical institute INSEE. For the calibration methods that use truncation, i.e. the logit method and the truncated linear method, a lower boundary of 0.4 and an upper boundary of 2.5 were used. As a matter of fact, the weights that were adjusted by the application of calibration were not exactly the design weights from the gross sample. Since for the bias evaluation of estimators the net samples, i.e. the response sets $r$, were used, the distance $G(w_k, d_k)$ of the design weights $d_k$ to the calibrated weights $w_k$ would have been quite large on average, because the design weights alone do not compensate for the sum of design weights lost by unit nonresponse. Therefore, the design weights were rescaled to the sum of the population (i.e. total number of private households in Austria in 2010) before calibration in order to allow for a faster convergence of the calibration algorithm.

The calibration estimators all perform similarly well which is evident in table 4.8. Both variance and bias are much lower than for the estimators in the previous subsections. A comparison of the different distance functions shows that the bias is lowest for the estimators using the logit or the truncated linear method. In other words, the expected values for the simulated estimates for $\hat{\theta}_{C_{logit}}$ and $\hat{\theta}_{C_{trunc.lin}}$ show very similiar relative bias, variance and MSE. With an MSE of $1.7490 \cdot 10^{18}$ $\hat{\theta}_{C_{logit}}$ is a little bit more precise than $\hat{\theta}_{C_{trunc.lin}}$ (MSE $= 1.7496 \cdot 10^{18}$), but only by a very small amount. The MSE for $\hat{\theta}_{C_{exp}}$ has an MSE that is about 2.3% higher, because of a slightly bigger bias. The linear method delivered the most biased calibration estimator of table 4.8 which also has the highest variance and therefore the highest MSE.

**Table 4.8:** *Simulation results - calibration estimators*

| Weighting procedure | Estimator | Rel. Bias in % | Variance | MSE |
|---|---|---|---|---|
| Calibration Linear method | $\hat{\theta}_{GREG}$, $\hat{\theta}_{C_{linear}}$ | -0.76 | $1.17 \cdot 10^{18}$ | $1.82 \cdot 10^{18}$ |
| Calibration Exponential method | $\hat{\theta}_{C_{exp}}$ | -0.76 | $1.13 \cdot 10^{18}$ | $1.79 \cdot 10^{18}$ |
| Calibration Logit method | $\hat{\theta}_{C_{logit}}$ | -0.74 | $1.13 \cdot 10^{18}$ | $1.75 \cdot 10^{18}$ |
| Calibration Exponential method | $\hat{\theta}_{C_{trunc.linear}}$ | -0.74 | $1.13 \cdot 10^{18}$ | $1.75 \cdot 10^{18}$ |

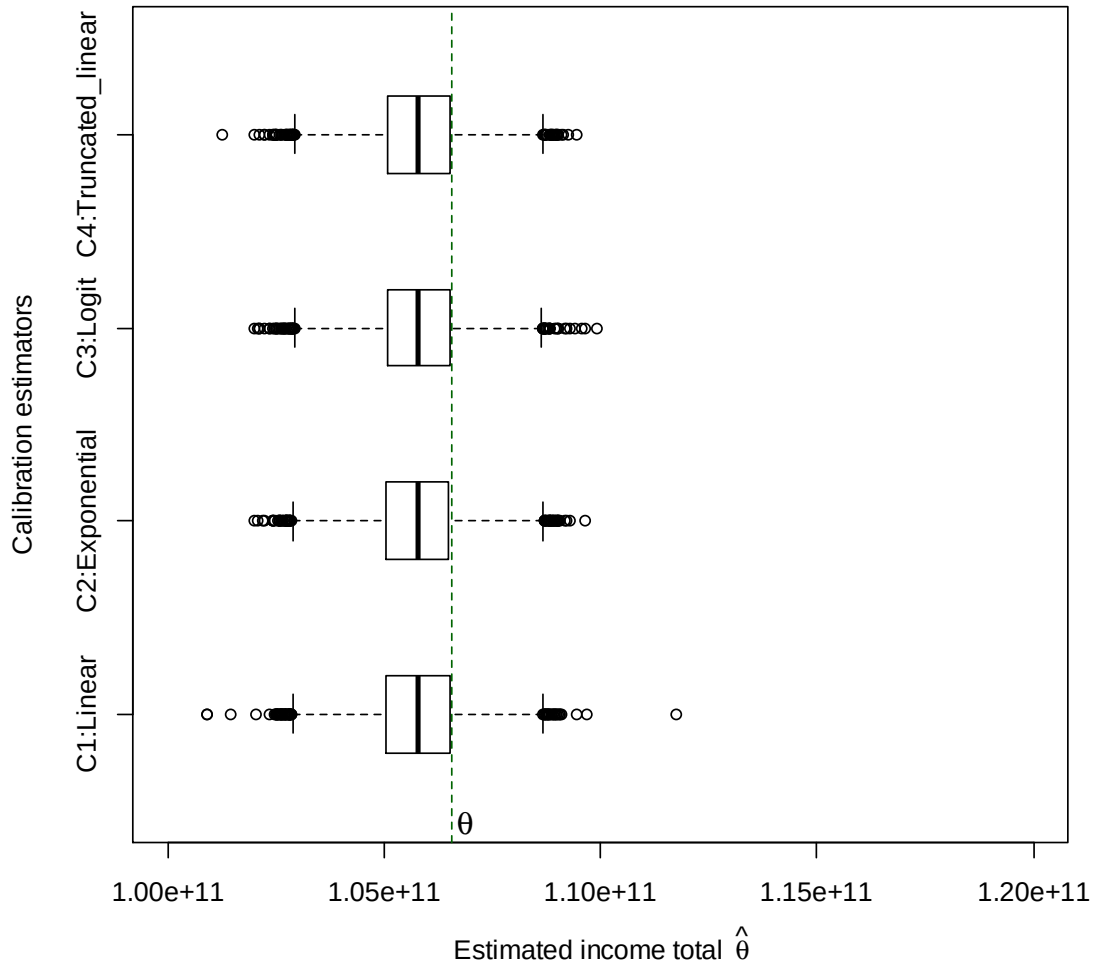
Source: Statistics Austria, EU-SILC 2010, unpublished results.

For generalized calibration[17] the simulation produced a very high variance which is due to some extreme values. Cutting off the top and bottom 1% or 5% from the simulation estimates did not help to make the results from generalized calibration comparable with the other calibration estimators. It reduced variance, but made the bias unproportionally large (up to 90%). The method is best suited if a survey variable is known to be highly associated with unit nonresponse. The choice of variables is crucial and more research is needed to assess appropriate variables, but this is beyond the scope of this thesis. Therefore generalized calibration is not presented in detail here.

Figure 4.7 compares the distribution of the calibration estimators (without generalized calibration). It can bee seen that these estimators have similiar simulation distributions and are also closer to the true parameter $\theta$ than the other estimators already presented. The distribution of $\hat{\theta}_{C_{linear}}$ has the most outliers since there are no limits for the values of the g-weights for computing the calibrated weights.

[17]Generalized calibration is implemented in CALMAR2, also developed by INSEE.

**Figure 4.7:** *Simulation distribution of the calibration estimators*




Source: Statistics Austria, EU-SILC 2010, unpublished results.

### 4.4.6 Logistic regression adjustment and calibration

The calibration estimators shown in the section above were calculated by solving calibration equations that used the design weights, rescaled to the population total, as input weights. A modified approach to calibration is using the nonresponse adjusted design weights as input weights. In the simulations presented in this thesis these adjustment were carried out by applying the logistic regression model 1 and 2 of section 4.4.3.

**Table 4.9:** *Simulation results - calibration estimators with preceding logistic regression adjustment (Model 1)*

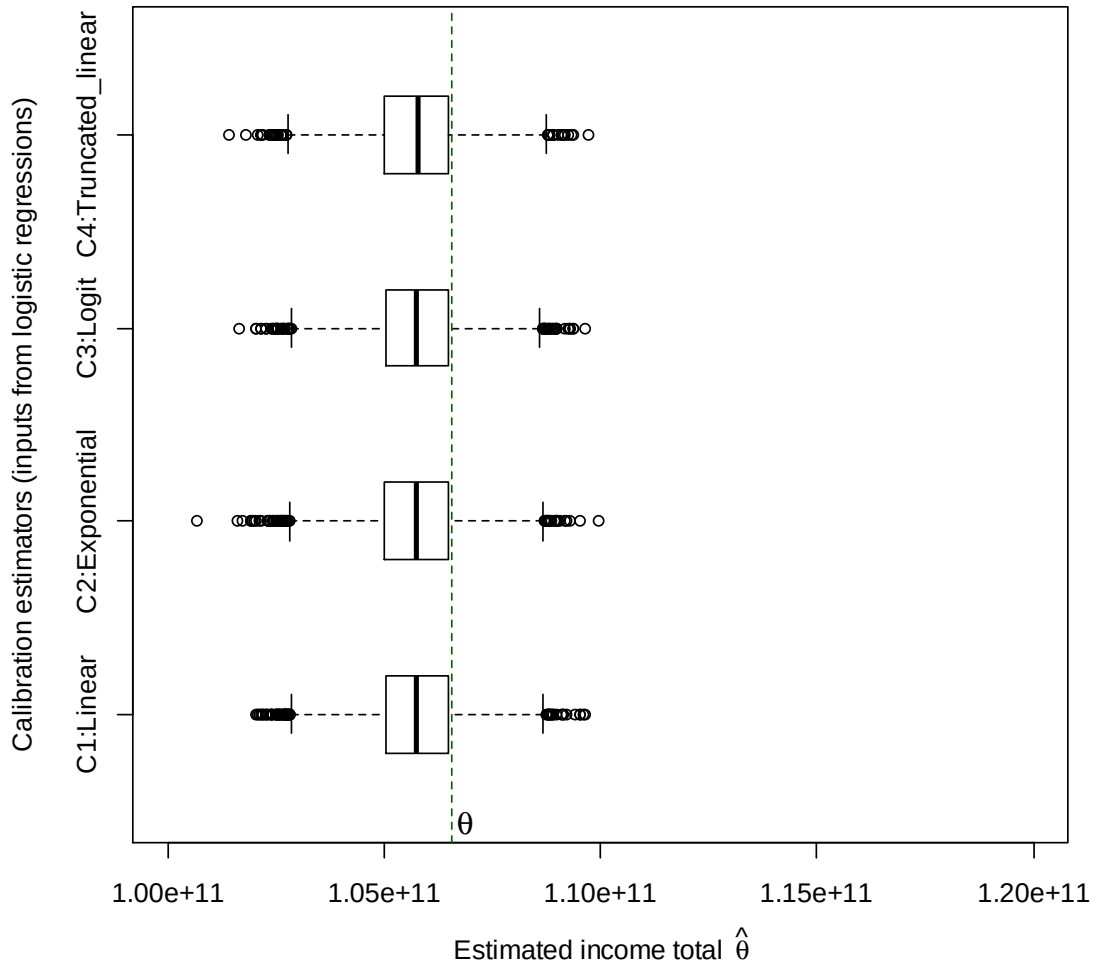
| Weighting procedure | Estimator | Rel. Bias in % | Variance | MSE |
|---|---|---|---|---|
| Log. Reg. Model 1 & Calibration: linear | $\hat{\theta}_{LOG_1}$, $\hat{\theta}_{GREG} = \hat{\theta}_{C_{linear}}$ | -0.78 | $1.18 \cdot 10^{18}$ | $1.87 \cdot 10^{18}$ |
| Log. Reg. Model 1 Calibration: exponential | $\hat{\theta}_{LOG_1}$, $\hat{\theta}_{C_{exp}}$ | -0.79 | $1.17 \cdot 10^{18}$ | $1.88 \cdot 10^{18}$ |
| Log. Reg. Model 1 Calibration: logit | $\hat{\theta}_{LOG_1}$, $\hat{\theta}_{C_{logit}}$ | -0.79 | $1.14 \cdot 10^{18}$ | $1.75 \cdot 10^{18}$ |
| Log. Reg. Model 1 Calibration: truncated linear | $\hat{\theta}_{LOG_1}$, $\hat{\theta}_{C_{trunc.lin}}$ | -0.77 | $1.16 \cdot 10^{18}$ | $1.84 \cdot 10^{18}$ |

Source: Statistics Austria, EU-SILC 2010, unpublished results.

As table 4.9 shows, the bias of using a logistic regression model to calculate input weights is similar to omitting this weighting step. The relative bias is slightly larger than for the estimators that use only calibration as weight adjustment for the design weights. This may be due to the fact that only a very limited amount of information is available from the sampling frame and the gross sample as predictors in the logistic regressions. Table 4.10 states that using a logistic regression model that uses more predictors by including interactions before calibration delivers estimators that are a little more biased. Trimming the unit nonresponse weights has shown a small reduction of the bias and the MSE. Altogether the findings of this subsection show rather similar results where no big improvement or degratation of estimators becomes apparent. Perhaps errors in calculating weight adjustments accumulate by using more than one adjustment step and therefore introduce a new source of bias. Further research that is beyond of the scope of this thesis has to be done in order to assess these effects in detail.

Once again, also the distributions of the different calibration estimators behave likewise. Little differences lie in a slightly bigger amount of outliers in $\hat{\theta}_{C_{exp}}$ and fewer outliers for $\hat{\theta}_{C_{linear}}$ which can be seen in figure 4.8 and figure 4.9.

**Figure 4.8:** *Simulation results - calibration estimators with preceding logistic regression adjustment (Model1)*



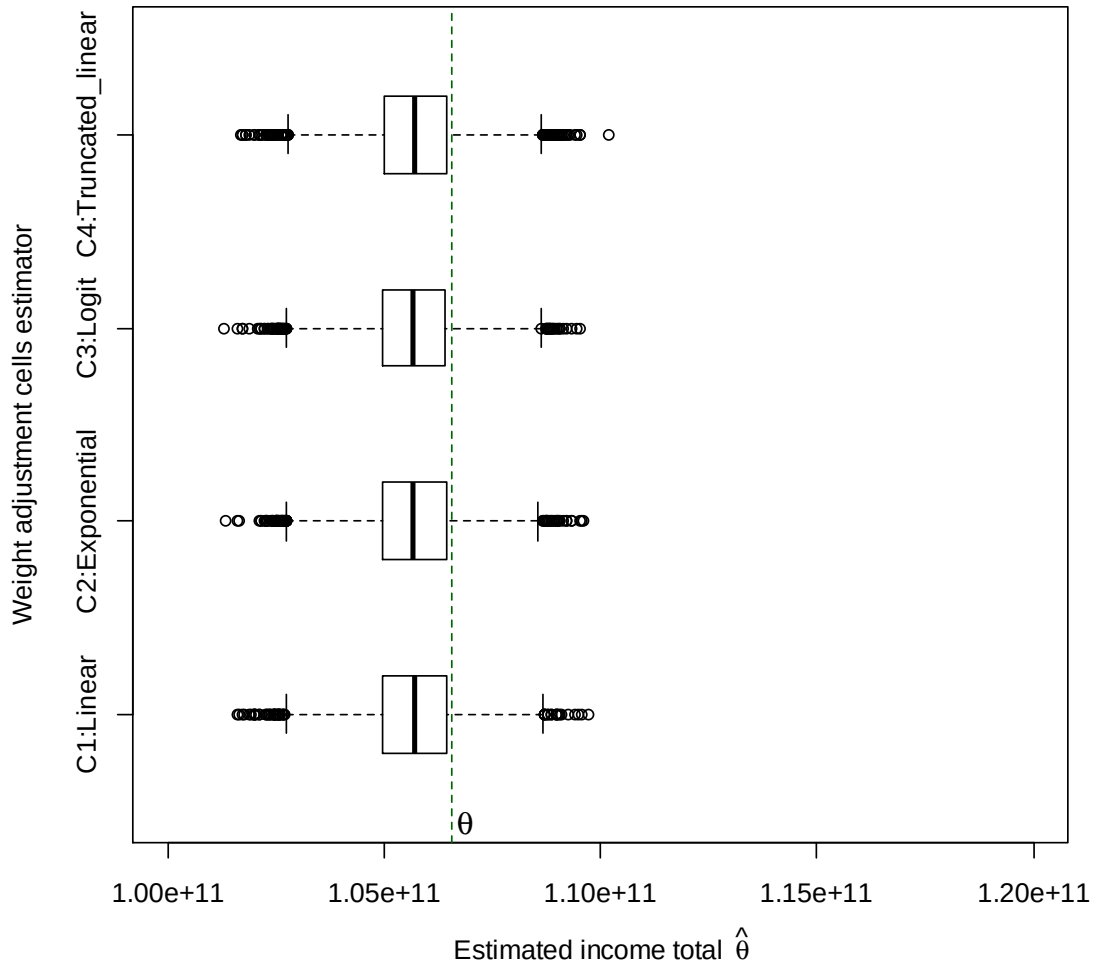

Source: Statistics Austria, EU-SILC 2010, unpublished results.

**Table 4.10:** *Simulation results - calibration estimators with preceding logistic regression adjustment (Model 2)*

| Weighting procedure | Estimator | Rel. Bias in % | Variance | MSE |
|---|---|---|---|---|
| Log. Reg. Model 2 & Calibration: linear | $\hat{\theta}_{LOG_2}$, $\hat{\theta}_{GREG} = \hat{\theta}_{C_{linear}}$ | -0.83 | $1.21 \cdot 10^{18}$ | $2.00 \cdot 10^{18}$ |
| Log. Reg. Model 2 Calibration: exponential | $\hat{\theta}_{LOG_2}$, $\hat{\theta}_{C_{exp}}$ | -0.85 | $1.18 \cdot 10^{18}$ | $2.00 \cdot 10^{18}$ |
| Log. Reg. Model 2 Calibration: logit | $\hat{\theta}_{LOG_2}$, $\hat{\theta}_{C_{logit}}$ | -0.85 | $1.18 \cdot 10^{18}$ | $2.00 \cdot 10^{18}$ |
| Log. Reg. Model 2 Calibration: truncated linear | $\hat{\theta}_{LOG_2}$, $\hat{\theta}_{C_{trunc.lin}}$ | -0.82 | $1.19 \cdot 10^{18}$ | $1.96 \cdot 10^{18}$ |

Source: Statistics Austria, EU-SILC 2010, unpublished results.

**Figure 4.9:** *Simulation results - calibration estimators with preceding logistic regression adjustment (Model 2)*




Source: Statistics Austria, EU-SILC 2010, unpublished results.

# 5 Concluding remarks and outlook

In this thesis procedures for estimating population parameters by using data from sample surveys that suffer some sort of unit nonresponse have been presented. Unit nonresponse was assumed to occur at random depending on an unknown response probability. Different weighting procedures resulted in different estimators that are modified versions of the unbiased Horvitz-Thompson estimator. Finally, simulations showed results that speak strongly in favour for calibration, as was indicated in the referring statistical literature discussed in this thesis.

At this point the three main research questions which were introduced at the beginning of the thesis (cf. section 1.2) shall be reviewed.

*1. Is there a considerable amount of bias due to unit nonresponse that has to be dealt with?*

The simulations presented in ch. 4.4 showed that the assumption of a response process that is missing completely at random leads to the most biased estimator which also shows the largest MSE (cf. ch. 4.4.1). It can be concluded that the response process and the according event of unit nonresponse is related to the survey variable of interest $y_k$, i.e. income from employment and old-age benefits in some way. Therefore, weighting methods that counter this unit nonresponse bias of the estimator $\hat{\theta}$ of the population parameter $\theta$ (i.e. total sum of income from employment and old-age benefits) have to be employed. Appropriate auxiliary variables that explain the response process as well as the survey variable can be used to adjust the design weights in order to counter the bias.

*2. Which methods to adjust estimators in order to counter unit nonresponse exist and how can they be improved?*

Two big approaches to adjust the design weights, i.e. the inverse of the Horvitz-Thompson estimator to account for unit nonresponse bias have been presented: Two-phase weighting and calibration. The first one delivers adjustment weights that resemble

the design weights used for the Horvitz-Thompson estimator, i.e. they are the inverse of the (estimated) response probability. By treating the event of response as a second sampling phase, the unit nonresponse weights are factors that are multiplied by the design weights in order to adjust for unit nonresponse. Using estimated response probabilities from neural networks has shown an improvement compared to logistic regression models as long as appropriate auxiliary variables are put into the neural network and enough hidden units in the hidden layer are used. The calibration approach also adjusts for unit nonresponse, but at the same time reduces variance too. Calibrated weights are computed by solving the calibration equations which minimize the distance of the design weights to the calibrated weights. Additionally this minimization problem is constrained by the condition that totals of the weighted auxiliary variables in the survey fit the corresponding totals from external marginal distributions. Calibration estimators can be improved by introducing boundaries for the fraction $w_k/d_k$ of calibrated weights and design weights for the distance function. For example, the unbounded linear method for the distance function results in an estimator that is a little less precise than the one which applies the bounded logit method for the distance function. Using design weights that are adjusted by the inverse of the estimated response probabilities and are subsequently calibrated adds a small amount of bias to the estimators. Since calibration estimators have a small negative bias and estimators using adjustments by estimated response probabilities from logistic regressions have a larger, positive bias, these two weighting procedures may interact in a way that affects the estimator that combines adjustments for unit nonresponse by logistic regression and calibration negatively. Perhaps this method could be improved if there was a larger set of variables that explain the response process available to use as predictors in the logisitc regression models. In this thesis only a limited amount of variables from the sampling frame and one variable from the gross sample were used in the simulations in order to resemble the EU-SILC survey of 2010 were no register data were yet available for the computation of indicators. Perhaps the future use of many income data from registers may improve the models and then deliver input weights for logistic regressions that introduce a further reduction of bias already facilitated by calibration. The weights adjusted by the inverse of the response probabilities in this way are then calibrated in the final step of the weighting procedure. EU-SILC also applies logistic regression adjustments for attrition weights for the follow-up waves, i.e. nonresponse weights to adjust for attrition after the first year wave of a rotational subsample. In this case survey variables from the preceding year can be used as predictors in logistic regression models. These are better

suited to estimate response probabilities which becomes evident in an improvement of the pseudo r-square for logistic regressions used in the follow-up waves of EU-SILC.[1]

*3. May a reduction of bias have undesirable consequences, for example a considerable enlargement of variance?*

The only set of methods that always results in a reduction of bias and variance are the calibration estimators. For the estimators employing the two-phase approach to weighting a reduction in bias may be achieved, but variance may rise at the same time. For example this becomes evident in the comparison of the straight expansion estimator and the weighting cells estimator. However, the reduction in bias here has a stronger effect on the MSE than the enlargement of variance and hence the MSE is smaller for the weighting cells approach. Another good illustration of the tradeoff between a reduction of bias and an increasing of variance is given in table 4.6. $\hat{\theta}_{NNW_6}$ is the estimator resulting from the weighting procedure that applys the neural network with the most hidden units and the largest set of availabe input variables. Simulations showed that it has the smallest relative bias of all neural network methods, but has the largest variance. In the end the reduction of bias has a greater impact on the MSE than the variance and thus $\hat{\theta}_{NNW_6}$ has the smallest MSE of all estimators using unit nonresponse adjustments by multiplying the design weights by the inverse of the response probabilities estimated by neural networks.

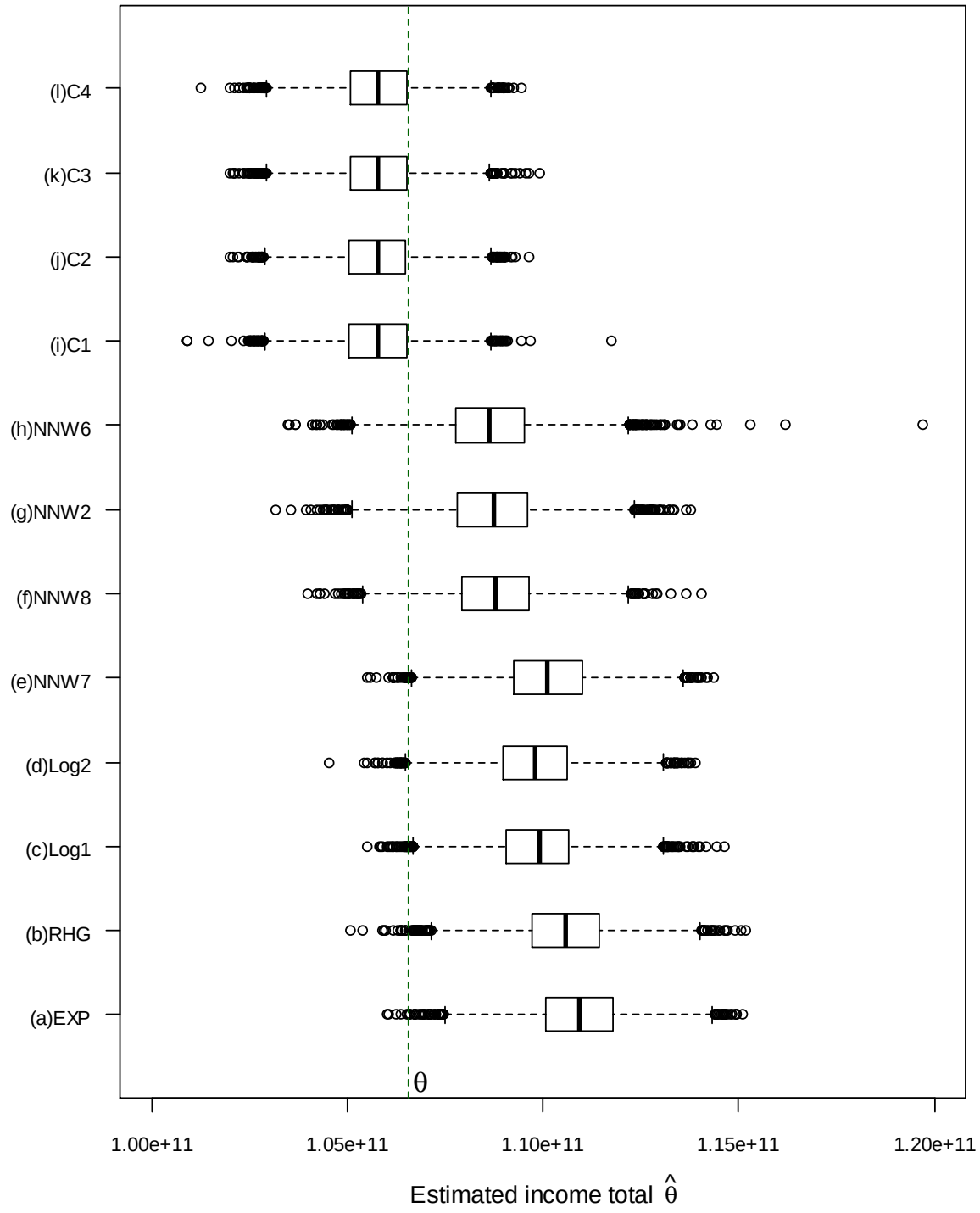
Figure 5.1 shows a comparison of the esitmators which turned out to be most relevant to illustrate strengths and weaknesses of the used weighting procedures. They are labeled by the letters (a)-(l), where methods (a)-(h) refer to the two-phase approach to weighting and (i)-(l) show the calibration estimators using different distance function (C1: linear, C2: exponential, C3: logit, C4: truncated linear).

The findings from the simulation results in figure 5.1 makes the power of the calibration methods evident. The corresponding boxplots for (i)-(l) show that the distribution of the simulated estimates applying calibration weighting are nearest to the true parameter $\theta$ (indicated by the dashed vertical line). Another advantage that speaks strongly in favour of calibration is the fact, that the totals of the survey variables that are calibrated are exactly the same as in the sources of these distributions. This is an important fact for the presentation of official statistics where important domains according to age, sex and region should be comparable in size across different statistics.

The neural networks ((e)-(h) in figure 5.1) show the best results for the two-phase

[1] Cf. Glaser & Till (2010).

**Figure 5.1:** *Simulation results - comparison of most relevant weighting methods*



approach to weighting. The neural network NNW6 uses 11 hidden units, which is the highest number of hidden units used in the simulations. NNW6 has the smallest MSE, but also shows few very extreme outliers. Therefore, one has to be careful with using neural networks and should not employ too many hidden units, because on a few occasions (occuring with a rate less than 0.1% in the simulated estimates) the resulting estimates are greatly off-the mark. $\hat{\theta}_{NNW_2}$ resulting from NNW2 (9 hidden units) may be better suited. Although its MSE is a little larger it has no extreme outliers.

Summing up all the results from simulations, all the weighting schemes presented in this thesis result in some sort of bias of the estimator of the total income from employment and old-age benefits. The methods tested in the simulations had to be limited to the variables that were available for all possible response sets. Therefore, only variables from the sampling frame and the gross sample were used as auxiliary information. For example, EU-SILC uses also variables that can only be determined by a questionnaire and for which marginal distributions from external sources are available, for example the number of people who receive unemployment benefits. If more and appropriate register information were available for the gross sample and external marginal distributions for the correct population, then more auxiliary information could be used for simulations. This may be a topic for future research.

For the simulations the "true parameter" was defined as the design-weighted estimate. The aim of this thesis was to show the impact of unit nonresponse on the design weighted estimate. So, concerning the population total, the design-weighted estimate was used as a benchmark, i.e. as the "true parameter", because of its theoretical unbiasedness. As a matter of fact it may be a bit lower than the true value in the real population. For drawing the sample of EU-SILC it is necessary to exclude some households from the sampling frame, because they are already in one of the rotational subsamples. This is unavoidable, because of the rotational design. This frame error due to a smaller frame usually is corrected by calibration to the known population total. The considerations in this thesis solely refer to the bias that happens due to unit nonresponse. Another issue with the simulation method used in this thesis are sampling frame errors. Sampling frame errors were assumed to be negligible, because for the simulated response set there is no time lag between the reference date and the date of "interview". In practice this is of course not true. Usually a few month pass between the drawing of the sample and the actual date of the interview. However, the logistic regression model to construct the response distribution by using the estimates from the logistic regression as probabilities

in the simulations only has a foundation in reality, if the vector of response is the one from the actual response set of EU-SILC 2010. A solution to this challenge could be to use a theoretical logistic distribution for the entire gross sample with distribution parameters taken from the empirical distribution of the estimated response probabilities.

For the practice of the EU-SILC survey the finding that calibration is the best method of the ones tested in this thesis is reassuring, because it is also used as the final step in the weighting process of EU-SILC in Austria. Further inquiry is needed for the follow-up waves, i.e. from the second survey year onwards, and for the combined cross-sectional sample. From EU-SILC 2012 onwards more income register data sources will be available and income based indicators will be calculated mainly with register data. As soon as the according data analysis procedure is implemented simulations for more components of the overall household income can be carried out in order to give a more precise evaluation of unit nonresponse bias and the effictiveness of the weighting methods applied. Among the methods at hand the usage of neural networks instead of logistic regressions, especially for the follow-up waves of EU-SILC, has to be evaluated in more detail. The method shows some promising results in the simulations carried out in this thesis, but further investigations concerning the handling of extreme weights that result from this method on a few occasions has to be given special attention.

# Bibliography


Amer, S. R. (2009), 'Neural Network Imputation In Complex Survey Design', *International Journal of Electrical, Computer, and Systems Engineering* **3**, 52–57.

Atkinson, A. B. & Marlier, E. (2010), Living conditions in Europe and the Europe 2020 agenda, *in* A. B. Atkinson & E. Marlier, eds, 'Income and living conditions in Europe', Eurostat Statistical Books, European Union, pp. 21–35.

Backhaus, K. (2006), *Multivariate Analysemethoden*, Springer-Verlag, Berlin-Heidelberg.

Bethlehem, J. G. (1999), Cross-sectional research, *in* H. J. Adèr & G. J. Mellenbergh, eds, 'Research methodology in the life, behavioural and social sciences', Sage, London, pp. 110–142.

Cobben, F. (2009), Nonresponse in Sample Surveys, PhD thesis, University of Amsterdam.
**URL:** *http://www.cbs.nl/NR/rdonlyres/2C300D9D-C65D-4B44-B7F3-377BB6CEA066/0/2009x11cobben.pdf (retrieved September 10, 2012)*

Cochran, W. (1977), *Sampling Techniques*, Wiley, California.

Deville, J.-C., ed. (2002), *La correction de la non-réponse par calage généralisé*, Journées de méthodologie statistique, École Nationale de la Statistique et de l'Analyse de l'Information, INSEE.
**URL:** *http://vserver-insee.nexen.net/jms/files/documents/2002/325_1-JMS2002_SESSION1_DEVILLE_CALAGE-GENERALISE_ACTES.PDF (retrieved September 10, 2012)*

Förster, M. & D'Ercole, M. M. (2009), The OECD approach to measuring income distribution and poverty: strengths, limits and statistical issues, *in* 'Measuring Poverty, Income Inequality, and Social Exclusion: Lessons from Europe', OECD.
**URL:** *http://umdcipe.org/conferences/oecdumd/conf_papers/ (retrieved September 10, 2012)*

Glaser, T. & Till, M. (2010), 'Gewichtungsverfahren zur Hochrechnung von EU-SILC Querschnittergebnissen', *Statistische Nachrichten* **65**, 566–576.

Groves, R. M. (2006), 'Nonresponse Rates and Nonresponse Bias in Household Surveys', *Public Opinion Quarterly* **70**(5), 646–675.

Groves, R. M., Fowler Jr., F. J., Couper, M. P., Lepkowski, J. M., E., S. & Tourangeau, R. (2004), *Survey Methodology*, John Wiley and Sons, California.

Horvitz, D. G. & Thompson, D. J. (1952), 'A Generalization of Sampling Without Replacement From a Finite Universe', *Journal of the American Statistical Association* **47**(260), 663–685.

Hosmer, D. W. & Lemeshow, S. (1989), *Applied Logistic Regression*, Wiley-Interscience.

Kish, L. (1990), 'Weighting: Why, When and How?'.
**URL:** *http://www.amstat.org/sections/srms/Proceedings/papers/1990_018.pdf (retrieved September 10, 2012)*

Kish, L. (1995), *Survey Sampling*, Wiley.

Kytir, J. & Stadler, B. (2004), 'Die kontinuierliche Arbeitskräfteerhebung im Rahmen des neuen Mikrozensus', *Statistische Nachrichten* **59**, 511–518.

Lehmann, E. & Caselly, G. (1998), *Theory of Point Estimation*, Springer.

Liang, F. & Yung Cheung Kuk, A. (2004), 'A finite population estimation study with Bayesian neural networks', *Survey Methodology* **30**, 219–234.

Little, R. J. A. (1986), 'Survey Nonresponse Adjustments for Estimates of Means', *International Statistical Review* **54**(2), 139–157.

Little, R. J. A. (1988), 'Missing-Data Adjustments in Large Surveys', *Journal of Business & Economic Statistics* **6**(3), 287–296.

Little, R. J. A. & Rubin, D. B. (1987), *Statistical Analysis with Missing Data*, Wiley, New York.

Longford, N. T. (2005), *Missing Data and Small-Area Estimation*, Springer, London.

Oh, H. L. & Scheuren, F. J. (1983), Weighting adjustment for unit nonresponse, *in* W. G. Madow, I. Olkin & D. B. Rubin, eds, 'Inomplete Data in Sample Surveys', Academic Press, New York, pp. 143–184.

Osier, G. (2010), Dealing with household non-response using generalized calibration, 21st International Workshop on Household Survey Nonresponse, Nuremberg. 21st

International Workshop on Household Survey Nonresponse.
**URL:** *http://www.nonresponse.org/uploadi/editor/1282915577S6_3Osier.doc (retrieved September 10, 2012)*

Osier, G., Museux, J., Seoane, P. & Verma, V. (2006), Cross-sectional and longitudinal weighting for the EU-SILC rotational design, *in* 'MOLS 2006: Methodology of Longitudinal Surveys', Eurostat, Luxembourg.
**URL:** *http://www.iser.essex.ac.uk/files/survey/ulsc/methodological-research/mols-2006/scientific-social-programme/papers/Osier.pdf (retrieved September 10, 2012)*

Pfanzagl, J. (1991), *Elementare Wahrscheinlichkeitsrechnung*, Walter der Gruyter, Berlin.

Potter, F., Grau, E., Williams, S., Diaz-Tena, N. & Lepidus-Carlson, B. (2006), An Application of Propensity Modeling: Comparing Unweighted and Weighted Logistic Regression Models for Nonresponse Adjustments, *in* 'Proceedings of the Survey Research Methods Section', American Statistical Association.
**URL:** *http://www.amstat.org/sections/srms/proceedings/y2006/Files/JSM2006-000801.pdf (retrieved September 10, 2012)*

Richard, M. & Lippmann, R. (1991), 'Neural Network Classifiers Estimate Bayesian a posteriori Probabilities', *Neural Computation* **3**, 461–483.

Ripley, B. D. (2004), *Pattern Recognition and Neural Networks*, 7 edn, Cambridge University Press, Cambridge.

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

Särndal, C.-E. & Deville, J.-C. (1992), 'Calibration Estimators in Survey Sampling', *Journal of the American Statistical Association* **87**(418), 376–382.

Särndal, C.-E. & Lundström, S. (2005), *Estimation in Surveys with Nonresponse*, John Wiley and Sons, West Sussex, England.

Särndal, C.-E., Swensson, B. & Wretman, J. (2003), *Model assisted survey sampling*, Springer, New York.

Sautory, O. (1993), La macro CALMAR - Redressement d'un échantillon par calage sur marges, Technical report, Institut National de la Statistique et des Études Économiques (INSEE), Paris.
**URL:** *http://www.insee.fr/fr/methodes/outils/calmar/doccalmar.pdf (retrieved September 10, 2012)*

Sautory, O., ed. (2003), *Calmar 2: A new version of the Calmar calibration adjustment program*, Statistics Canada. Statistics Canada International Symposium Series.
**URL:** *http://www.statcan.gc.ca/pub/11-522-x/2003001/session13/7713-eng.pdf (retrieved September 10, 2012)*

Statistics Austria (2010), 'Intermediate Quality Report relating to the EU-SILC 2010 Operation'.
**URL:** *http://circa.europa.eu/Public/irc/dsis/eusilc/library?l=/quality_assessment/quality_reports/at/2010_intermediate/_EN_1.0_&a=d (retrieved August 10, 2012)*

Till-Tentschert, U., Glaser, T., Heuberger, R., Kafka, E., Lamei, N. & Skina-Tabue, M. (2011), *Armut und Ausgrenzungsgefährdung in Österreich. Ergebnisse aus EU-SILC 2010*, Bundesministerium für Arbeit, Soziales und Konsumentenschutz, Vienna.

Verma, V. & Betti, G. (2010), Data accuracy in EU-SILC, *in* A. B. Atkinson & E. Marlier, eds, 'Income and living conditions in Europe', Eurostat Statistical Books, European Union, pp. 57–77.

Wolff, P., Montaigne, F. & González, G. R. (2010), Investigating in statistics: EU-SILC, *in* A. B. Atkinson & E. Marlier, eds, 'Income and living conditions in Europe', Eurostat Statistical Books, European Union, pp. 37–55.

# Lebenslauf:

**Persönliche Daten:**

| | |
|---|---|
| Name: | Mag. Thomas Glaser |
| Geburtsdatum: | 17. Jänner 1981 |
| Geburtsort: | Wien |
| Staatsbürgerschaft: | Österreich |

**Ausbildung und beruflicher Werdegang:**

| | |
|---|---|
| Seit Oktober 2007 | Statistik Austria, Guglgasse 13, 1110 Wien<br>▪ Angestellter im Bereich Soziales und Lebensbedingungen |
| Seit März 2006 | Universität Wien, Universitätsring 1, 1010 Wien<br>▪ Studium der Statistik |
| 2001–2007 | Universität Wien, Universitätsring 1, 1010 Wien<br>▪ Studium der Soziologie mit Spezialgebiet Politische Soziologie<br>▪ Abschluss im April 2007 (mit Auszeichnung) |
| 2000–2001 | Zivildienst, Österreichisches Rotes Kreuz |
| 1999–2005 | TU-Wien, Karlsplatz 13, 1040 Wien<br>▪ Studium der Technischen Mathematik, 1. Abschnitt abgeschlossen |
| 1991–1999 | BG, BRG und WiskuRG XI, Geringergasse 2, 1110 Wien<br>▪ Matura 1999 (mit Auszeichnung) |
| 1987–1991 | Volksschule, Herderplatz 1, 1110 Wien |

**Ausgewählte Publikationen:**


Geisberger, T., Glaser, T. (2010), Analyse der Lohn- und Gehaltsunterschiede von Frauen und Männern, *in* Frauenbericht 2010. Bundeskanzleramt Österreich. Wien. 197-199.

Glaser, T. (2010), *Zwischen Altruismus und Eigennutz?: Motivationen für ehrenamtliches Engagement in sozialen Nonprofit Organisationen.* VDM Verlag Dr. Müller. Saarbrücken.

Glaser, T. & Till, M. (2010), 'Gewichtungsverfahren zur Hochrechnung von EU-SILC Querschnittergebnissen', Statistische Nachrichten 65, 566–576.

Till-Tentschert, U., Glaser, T., Heuberger, R., Kafka, E., Lamei, N. & Skina-Tabue, M. (2011), *Armut und Ausgrenzungsgefährdung in Österreich. Ergebnisse aus EU-SILC 2010*, Bundesministerium für Arbeit, Soziales und Konsumentenschutz, Wien.