Master's Thesis

# Sensitivity of Binary Prediction Models to Boundary Shifts

Department of Statistics
Ludwig-Maximilians-Universität München

**Bavo Frederik Gerhard Sturm**

Munich, August 1st, 2023

Submitted in partial fulfillment of the requirements for the degree of M. Sc.
Supervised by Prof. Dr. Thomas Augustin

## Abstract

Poverty is a widespread and complicated social problem that affects individuals all over the world. To reduce the suffering of the poor, it is critical first to understand the causes of poverty and then identify the individuals affected by it. Boundary Shifts is a method that can address both challenges. It is a new technique in the field of fuzzy poverty measurement. Like other fuzzy poverty measurement methods, Boundary Shifts accounts for the fact that there is not only one dimension of poverty. It further overcomes the binary view of some poverty measurement methods where individuals are either poor or non-poor by assigning each individual a degree of being poor.

A central concept of Boundary Shifts is that repeatably a binary prediction model is fit to a data set, where progressively more individuals with higher incomes are classified as poor. This gives for each fit of the binary prediction model estimated parameters which reveal what distinguishes the poor from the non-poor when the boundary, which divides poor and non-poor, rises. A central task of this thesis is to evaluate the changes in the parameter estimates under different model assumptions. To overcome the problem of different parameter estimates despite only small boundary shifts, a new binary prediction model is introduced and analysed. Because Boundary Shift can assign a degree of poverty to each individual in the data set, it will be examined if this method assigns the same degrees of poverty to the same individuals as other fuzzy poverty measurement methods.

The sensitivity analysis showed that the logistic regression model as the binary prediction model caused the least fluctuating parameter estimations. Using Robust logistic regression or Bootstrap could not reduce the fluctuation of the parameter estimations. Further, it could be observed that count and categorical variables cause quasi-complete separation at some boundaries, which means that either the poor or non-poor can, to a certain degree, be perfectly predicted with a single variable. Comparing Boundary Shifts and other fuzzy poverty measurement methods showed that the household-assigned degree of poverty differs between the fuzzy poverty measurement methods.

It can be concluded that Boundary Shifts is a useful fuzzy poverty measurement method with advantages over other fuzzy poverty measurement approaches since the drivers of poverty at different boundaries and the poverty indices can be evaluated simultaneously.

# Contents

# 1 Introduction

Poverty is a challenging problem that affects individuals all around the world. Events with global impact, like COVID-19 and the war in Ukraine, are derailing the progress on ending extreme poverty. The world's poorest individuals bore the highest costs of the pandemic. Their income losses were twice as great as the world's richest, and global inequality climbed for the first time in decades. In addition, the poorest suffered major setbacks in health and education, leading to premature mortality and pronounced learning losses (World Bank, 2022b, p. XXI). Poverty reduction is more challenging than ever due to increasing food and energy prices caused in part by the Russian Federation's invasion of Ukraine and climate shocks in the world's largest food producers (World Bank, 2022b, p. 1).

Understanding poverty and its causes is essential for reducing its impacts and improving the lives of those in it. To quantify, monitor and compare the poverty of different populations or the individuals in it, poverty measurement is used. Depending on the method, poverty measurement can go beyond just analysing the income dimension and include other dimensions like health, education and housing. In this way, it can be determined to what extent the population's basic needs are met or to what extent individuals are particularly disadvantaged in certain dimensions compared to the population as a whole.

One method the World Bank uses to measure international extreme poverty is an absolute poverty measurement approach where the poverty line is placed at $ 2.15 per individual per day and anyone living on less than $ 2.15 a day is in extreme poverty (World Bank, 2022a). The At-risk-of-poverty rate (AROP) is a relative poverty measure used as a social indicator by the European Commission. Individuals who earn less than 60% of the median income are at risk of poverty, according to this measure (European Commission, n.d.). The Multidimensional Poverty Index (MPI) measures acute multidimensional poverty across over 100 developing countries. The MPI report was first launched in 2010 by the Oxford Poverty and Human Development Initiative at the University of Oxford and the Human Development Report Office of the United Nations Development Programme (UNDP). The MPI is calculated from 10 Indicators of the three dimensions Health, Education and Standard of Living. The indicator income is not used for poverty measurement in this index (UNDP and OPHI, 2022, p. 3). The Human Development Index (HDI) is a summary measure of average achievement in major dimensions of human development and is another poverty measurement method used by the UNDP. Like in the MPI, the used dimensions are Health, Education and Standard of Living, but the indicators differ. For example, income is used here as an indicator of the standard of living (UNDP, n.d.).

The methods mentioned have in common that they assess poverty using only income as an indicator, or arbitrary decisions are made in the calculation. Examples are the poverty line placed at $ 2.15 or 60% of the median income or that the HDI is calculated as the geometric mean and not as the arithmetic mean of the dimension-related indices (UNDP, 2022, p. 3). Additional poverty measurement methods can be derived from the mathematical theory of fuzzy sets. Some of these methods overcome the mentioned issues and have a way to deal with the property of poor being a vague predicate. Poor being a vague predicate means that there exist borderline cases where it is not possible to assign individuals to the set of the poor or non-poor since it can not be said where the poverty line lies that separates the poor from the non-poor or it does not even exist depending on the view of vagueness (Qizilbash, 2006, pp. 12-17).

There are a few fuzzy poverty measurement methods that will be discussed in the following and a new fuzzy poverty measurement method will be introduced as well. The name of it is **Boundary Shifts (BS)**. The structure, procedures and model modifications of this method are explained in this thesis. Also, it is explained why this method is a fuzzy poverty measurement method and it is shown how it compares to other methods.

To get started, the following chapter shows that poverty can be defined and understood differently. The resulting consequences are then explained. Afterwards, in Chapter 3, the fuzzy sets theory is introduced to understand the selection of poverty measurement approaches given in the same chapter. To go more in detail, four fuzzy poverty measurement approaches are looked at, and afterwards, these approaches are compared and discussed. In Chapter 4, logistic regression is briefly introduced to understand the process of estimating the regression coefficients. This knowledge is needed to understand logistic lasso regression and another modification of the logistic regression model that will be presented in a later chapter. The theoretical foundation for the new approach, BS, is laid in Chapter 5. In Chapter 6, the data set “Filipino Family Income and Expenditure” is introduced, and the structure, variables and other characteristics of the data set are summarised. Afterwards, the data set is prepared. The chapter also contains an exploratory data analysis to get a good data overview. Chapter 7, Sensitivity Analyses, covers various BS-related topics. An initial model is used to localise the first characteristics and to make the analyses clearer. Since different statistical data types affect the BS differently, they are considered in detail. In the remaining part of the chapter, model changes are analysed. A new binary prediction model to improve BS is introduced in Chapter 8. Sensitivity analyses are performed subsequently. In Chapter 9, it is demonstrated why BS may now be referred to as a fuzzy poverty measurement technique and how various methods for computing poverty indices from BS are established. Afterwards, the poverty indices derived from these methods are compared to those calculated from other non-BS fuzzy poverty measurement methods. In the last Chapter 10, the results are summarised and discussed. In addition, an outlook is given on where to conduct further research on this topic. The Appendix A contains further illustrations. In the electronic Appendix B, the thesis-related R-files are described.

The following chapter clarifies poverty and poverty measurement.

# 2 The Concept of Poverty

Poverty can be defined as follows. Poverty is the state of being poor, whereby being poor is defined as having very little money or not having enough money for basic needs (Hornby et al., 2015, pp. 1190, 1201). Another definition is that poverty is a denial of choices and opportunities, as well as a violation of human dignity; it is a lack of fundamental ability to participate effectively in society, which includes not having enough to feed and clothe a family, not having a school or clinic, and other things (ACC, 1998, p. 1). There is a fundamental difference between the two definitions since one depicts poverty as a lack of money, making it a one-dimensional problem, and the other as a lack of several needs, making it a multi-dimensional problem.

According to Vero (2006, p. 211), poverty is a difficult concept, and it may be defined in various ways which correspond to different philosophical approaches. Still, the basic idea is that poverty is a consequence of inequality, between individuals, in the control of certain things. Keeping this in mind, I would argue that deciding which one of the two definitions better represents poverty is impossible. Deciding which one to use is, moreover, a matter of personal taste. Still, it is clear that an individual who has a lot of money and is therefore non-poor according to the first definition does not necessarily have more quality of life when basic needs such as access to nearby hospitals, education, freedom, availability of healthy food and clean water are missing. Therefore poverty measurement will be seen as a multi-dimensional problem in this thesis.

In poverty measurement, vagueness is another topic that needs to be considered since “poor” is a **vague predicate**. Simply said, vague indicates that something is unclear. According to Qizilbash (2006, p. 10), a predicate is whatever is affirmed or denied of a subject by means of the copula. This explanation probably contains various unknown terms from the field of linguistics, so the examples “tall”, “bald” and “nice” might better show what vague predicates are. All of these predicates additional to “poor” have in common that (Qizilbash, 2006, p. 10)

1. they allow for situations in which it is unclear whether the predicate applies or not,
2. there is no sharp borderline between circumstances where the predicate applies and where it does not,
3. they are prone to a Sorites paradox.

The Sorites paradox was invented by the philosopher Eubulides. He starts with the assumption that one million grains of sand constitute a heap. Removing one grain of sand never turns a heap into a non-heap. When this principle is used repeatedly, it leads to the conclusion that one grain of sand constitutes a heap (Kim et al., 2009, p. 565). But according to Qizilbash (2006, p. 11), there is a contradiction in this situation since a single grain can not be considered a heap.

Checking if the three characteristics apply to the previously mentioned predicates reveals that they are all vague predicates. The following example shows that the three characteristics apply to the predicate poor, whereby in this example, an individual is considered poor in terms of the single indicator, income. That poor has the first characteristic can be seen in situations where individuals have a medium-high or medium-low income since here it can not be said that these individuals are definitely poor or non-poor. The second characteristic is present since there is no exact income value where it can be said that someone with a higher income is non-poor and someone with an income below is considered poor. Lastly, if a non-poor individual gives away one income unit, the individual is a bit poorer but still non-poor. After repeated removal of one income unit, the individual should still be non-poor according to this logic. But, the individual is poor after removing too many income units. This is a contradiction since the individual is poor and non-poor at the same time (Qizilbash, 2006, pp. 10-11). This is why poverty measurement needs special treatment.

A further relevant aspect in poverty measurement is **absolute** and **relative poverty**. One could argue that depending on the degree of poverty approaching the problem changes. Sachs (2005, p. 20) describes the two degrees of poverty as follows. Absolute poverty, also known as extreme poverty, means that the ones affected by it cannot meet their basic needs for survival because they are chronically hungry, have no access to health care, no clean drinking water and much more. This degree of poverty occurs only in developing countries. Relative poverty means, according to Eskelinen (2011, pp. 942-943), that the ones affected by it are poor in comparison to individuals of the analysed population. This means that someone can be relatively and not absolutely poor at the same time if basic needs are met but common goods are not possessed. Examples of lack of goods include not owning a car, mobile phone or not having access to entertainment (Sachs, 2005, p. 20).

One may consider the later introduced BS method as a method to analyse relative poverty since observations of the data set are compared among each other, and the poverty indices are calculated on a population basis.

One can conclude from this chapter that from here on, poverty measurement will be seen as a multidimensional problem, and there are different degrees of poverty. Further, accounting for it that “poor” is a vague predicate might help to understand poverty better. Fuzzy poverty measurement methods aim to account for the vagueness of poverty, and these methods can be applied to multidimensional problems. A few of these methods are explained next, but before, fuzzy set theory is introduced.

# 3 Fuzzy Poverty Measurement

## 3.1 Fuzzy Set Theory

In this chapter, fuzzy set theory is introduced. It starts with the definition and two examples. Afterwards, related properties are explained. This chapter is about teaching fuzzy set theory to the extent that different methods of fuzzy poverty measurement can be understood.

An "ordinary" set $A$, which will be referred to as **crisp set**, is defined by Zimmermann (2001, p. 11) as a collection of elements or objects $v \in V$ that can be finite, countable, or over countable and each element either belongs or does not belong to this set $A$ with $A \subseteq V$. A fuzzy set is defined according to Zimmermann (2001, pp. 11-12) as follows.

**Definition 3.1** (Fuzzy set)**.** If $V$ is a collection of objects denoted generically by $v$, then a **fuzzy set** $\tilde{A}$ in $V$ is a set of ordered pairs

$$\tilde{A} = \{(v, \mu_{\tilde{A}}(v)) \mid v \in V\}. \tag{1}$$

The **membership function** $\mu_{\tilde{A}}$ is the degree of membership. It maps $V$ to the membership space $M$. When $\mu_{\tilde{A}}(v) \in \{0, 1\}$ applies, then is $\tilde{A}$ a crisp set as well and could be denoted as usual. For clarification of Definition 1 and the analogy to crisp sets, Fustier (2006, p. 31) gives the upcoming example. Given is the set $V = \{a, b, c, d\}$, which represents a set of regions. Region $a, d$ are islands and $b, c$ are mainland regions. The crisp set $A = \{a, d\}$ that contains the "insular" regions can according to Definition 1 in fuzzy terms then be written as

$$\tilde{A} = \{(a, 1), (b, 0), (c, 0), (d, 1)\}.$$

So in this example, each element of $V$ belongs either to $A$ or does not.

The more interesting case of the membership degree $\mu_{\tilde{A}}(v)$ being not just equal to 0 or 1 requires the usage of fuzzy sets, which is of central importance in this thesis, as it is not the task to distinguish between belonging to a set $A$ or not but moreover to identify the degree of belonging to $\tilde{A}$, whereby $\mu_{\tilde{A}}(v)$ the degree of $v$ belonging to a fuzzy set $\tilde{A}$ is. To clarify what a fuzzy set is, Zimmermann (2001, p. 12) gives the following example. There is a set $V = \{1, \ldots, 10\}$ of available house types that reflects the number of bedrooms available. For example, a house of type three contains three bedrooms. A real estate agent wants to classify these houses now by the comfort of each house type for a four-person family. It is said that the indicator of comfort is the number of bedrooms in a house. The resulting fuzzy set $\tilde{A}$, which contains the comfortable types of houses for a four-person family then could be

$$\tilde{A} = \{(1, 0.2), (2, 0.5), (3, 0.8), (4, 1), (5, 0.7), (6, 0.3)\}.$$

House type four with four bedrooms is, in this example, the most comfortable house since the membership degree is one and for either increasing or decreasing house types, the degree of membership decreases. That means that house types could still suit the family, but less than a house of type 4.

The examples have now shown what fuzzy sets are. The main interest in the following will be to determine the membership degree of individuals to the set of poor. Therefore, membership functions and the approaches they are used in are introduced.

## 3.2 Poverty Measurement Approaches

It is recommended to involve multiple poverty indicators for poverty measurement (Betti, Cheli, Lemmi and Verma, 2006, p. 125; Panek, 2006, p. 234; Betti, D'Agostino and Neri, 2006, p. 257) . Miceli (2006, p. 196) adds that multivariate poverty measurement makes capturing the general living conditions possible instead of just the material situation. It is assumed that each individual $i$ out of $n$ individuals possesses a $k$-vector $(v_1, v_2, \ldots, v_k) = \boldsymbol{v} \in \mathbb{R}^k$ of indicators. Furthermore, $\mu_j$ is the membership function for indicator $V_j$ out of the $k$ indicators, so each indicator has its own membership function. It starts now with the non-fuzzy traditional poverty measurement approach, which assigns individuals to the set of the poor based on a single indicator. Afterwards, fuzzy poverty measurement approaches are introduced, where multiple poverty indicators are used.

### 3.2.1 Traditional

In the **traditional approach**, poverty is measured based on a single indicator, such as income. A **poverty line** $z \in \mathbb{R}$ is placed at a certain value in this approach. Every individual that is more disadvantaged according to the measured indicator is considered poor. As a result, there is a crisp set of poor $A$, that includes all the poor individuals. Note that in this thesis, the term “poverty line” is a poverty measurement-specific term, and it refers to the more general term “sharp borderline” used previously.

As seen in Chapter 3.1, can a crisp set $A$ be expressed as a fuzzy set $\tilde{A}$. The membership function of the traditional approach is then defined as

$$\mu_{\tilde{A}}(v) = \begin{cases} 1 & \text{if } v < z, \\ 0 & \text{if } v \geq z. \end{cases} \tag{2}$$

$v$ is the measurement of a single poverty indicator of an individual, and it is assumed that having less of some good that is measured by the indicator makes an individual more disadvantaged. One could thus argue that the traditional approach is theoretically a fuzzy poverty measurement approach.

In the traditional poverty measurement approach, the poverty line is drawn in reference to the mean or median income of a society (Berenger and Celestini, 2006, p. 139). The World Banks' extreme poverty measurement and the calculations of AROP mentioned in the introduction are two examples of the application of the traditional approach. These examples showed that absolute and relative poverty can be measured with this approach.

### 3.2.2 Totally Fuzzy and Absolute (TFA)

The second mentioned approach for poverty measurement, which is multivariate, is the **TFA approach** by Cerioli and Zani (1990, p. 274). For continuous variables, the membership functions

$$\mu_j(v_j) = \begin{cases} 1 & \text{if } v_j \leq v_j^{(L)}, \\ \dfrac{v_j^{(H)} - v_j}{v_j^{(H)} - v_j^{(L)}} & \text{if } v_j^{(L)} < v_j \leq v_j^{(H)}, \\ 0 & \text{if } v_j > v_j^{(H)}, \end{cases} \tag{3}$$

are used with the boundary values $v_j^{(L)}$ up to which individuals are definitely poor and $v_j^{(H)}$ above which individuals are definitely non-poor. Again, this means an individual who owns less of a good is more disadvantaged. The membership function has to be defined in this method for the indicators individually.

To highlight here, the difference between this and the traditional approach is, when it is concentrated on a single indicator, that between the two boundary values, the membership function linearly declines from 1 to 0. Thus, there are now partially poor individuals.

Besides the linear decreasing function in $\left(v_j^{(L)}, v_j^{(H)}\right]$, there could further non-linear functions be taken. Among them are sigmoid, logistic, gaussian, exponential or irregularly shaped functions that can be used for fine-tuning. Some not only require defining $v_j^{(L)}$ and $v_j^{(H)}$ but also the flex or crossover point to be associated with a degree of membership of 0.5 (Martinetti, 2006, p. 101).

The determination of the values for $v_j^{(L)}$ and $v_j^{(H)}$ for indicator $i$ is done by the investigator. An approach given by Cerioli and Zani (1990, p. 274) is to set $v_j^{(L)}$ equal to the minimal amount of goods that are required for living, and $v_j^{(H)}$ at the observed mean number of goods. With this rule, individuals with a number of goods below the lower limit are regarded as poor, and individuals with a number of goods between the boundary values are partially poor.

Schaich and Münnich (1996, p. 449) mention that it is convenient to set $v_j^{(L)}$ to zero for economically well-developed countries because a physiologically conceived subsistence minimum is provided by public institutions, e.g. by paying income support. For developing countries, $v_j^{(L)} > 0$ should be equal to the minimum amount of goods that are required to secure the physiological subsistence minimum for the sake of not having a developed social system. I would argue that with the placement of the lower and upper limits, it is controlled if relative or absolute poverty is measured. For example, relative poverty could be measured if the lower limit is set equal to zero or the smallest observed amount of goods and the upper limit to the highest observed amount of goods or above. Absolute poverty is, in my view, measured when values in between are taken for the lower and upper limits.

For categorical indicators that are measured on an ordinal scale, a modification of the membership function from Equation 3 can be used. This process is explained by Cerioli and Zani (1990, pp. 275-276). Looking at an indicator $V_j$ with $s_j$ ordered categories then a score $\psi_j^{(r)}$ with $r = 1, \dots, s_j$ can be assigned to each category, so that

$$\psi_j^{(1)} < \ldots < \psi_j^{(r)} < \ldots < \psi_j^{(s_j)}, \tag{4}$$

applies. A straightforward assignment is

$$\psi_j^{(r)} = r, \tag{5}$$

which assumes that the categories of $V_j$ are equally spaced, but other scores assignments are also possible. Having these scores, the membership function is then defined as

$$\mu_j(\psi_j) = \begin{cases} 1 & \text{if } \psi_j \leq \psi_j^{(L)}, \\ \dfrac{\psi_j^{(H)} - \psi_j}{\psi_j^{(H)} - \psi_j^{(L)}} & \text{if } \psi_j^{(L)} < \psi_j \leq \psi_j^{(H)}, \\ 0 & \text{if } \psi_j > \psi_j^{(H)}, \end{cases} \tag{6}$$

with values $\psi_j^{(L)}, \psi_j^{(H)}, \psi_j$ as the categories $v_j^{(L)}, v_j^{(H)}, v_j$ corresponding scores.

Cerioli and Zani (1990, p. 275) explain handling of dichotomous indicators too. It is assumed that there are $k' \leq k$ dichotomous indicators, and one category in each dichotomous indicator $V_j$, $j = 1, \dots, k'$ indicates poverty. The membership degree to the set of being poor is then calculated for all dichotomous indicators similarly by using the membership function

$$\mu_{dicho}(\boldsymbol{v}) = \frac{1}{k'} \sum_{j=1}^{k'} \mu_j(v_j), \tag{7}$$

where $\mu_j(v_j) = 1$ applies if an individual is according to indicator $V_j$ poor and $\mu_j(v_j) = 0$ if an individual is not. This means that the membership function of all dichotomous indicators comprises individual membership functions. This idea is used in the following to combine the membership functions of all indicators into a single membership function.

In multidimensional poverty measurement, it is preferable to end up with a single index for each individual. This index will be referred to as the **poverty index** in this thesis. It can be calculated for each individual in a population to reflect how poor the individual is dependent on the measured indicators.

The poverty index of an individual composes of the membership degrees calculated with the membership functions. In TFA can, according to Cerioli and Zani (1990, p. 276), Formula 7 be used to calculate the poverty index $\mu(\boldsymbol{v})$ from the membership degrees $\mu_j(v_j)$. It is just necessary that the membership degrees for each indicator are calculated with its data type corresponding membership function. When the poverty index is calculated with Formula 7, equal importance is attached to each variable. For example, suppose poverty is measured with the indicators "Owning a Television" and "Income". In that case, both indicators equally impact the poverty index, although one indicator might be more important to determine if an individual is poor.

The issue of the previous example leads to the weighted version. Cerioli and Zani (1990, p. 276) define the poverty index as

$$\mu(\boldsymbol{v}) = \frac{\sum_{j=1}^{k} \mu_j(v_j) w_j}{\sum_{j=1}^{k} w_j}, \tag{8}$$

where $w_j$ are weights that determine the impact of indicators since it is reasonable to assume that some indicators are more important when determining if an individual is poor. Two mentioned ways to define the weights $w_j$ are

$$w_j = \frac{1}{f_j}, \tag{9}$$

and the modification

$$w_j = \ln\left(\frac{1}{f_j}\right), \tag{10}$$

where $f_j$ is the rate of individuals exhibiting deprivation according to indicator $V_j$ (Cerioli and Zani, 1990, p. 277). It is said that taking the logarithm of $\frac{1}{f_j}$ avoids assigning large weights to variables with a small $f_j$ that would consequently dominate the poverty index. A precise description of how to derive $f_j$ is not given, which makes it unclear how to calculate it, as it is unclear in this fuzzy poverty measurement method which individuals are exhibiting deprivation in this indicator. It could be those with a membership degree below zero or below one.

It is explained that both definitions of weight see poverty as a matter of relative deprivation (Cerioli and Zani, 1990, p. 277). If a great fraction of the population owns some goods, and just some individuals do not, it makes sense to weight the corresponding indicator higher. For example, “Having a Bathroom” is achieved by a large fraction of the population because it is essential for living and “Owning a Car” is achieved by a smaller proportion as there are alternative ways of getting around. Then it makes sense to give the indicator “Having a Bathroom” a larger weight because an individual without a bathroom is relatively more deprived than an individual without a car, and therefore more individuals put a greater effort into having a bathroom.

### 3.2.3 Totally Fuzzy and Relative (TFR)

Another fuzzy poverty measurement approach is the **TFR approach** introduced by Cheli and Lemmi. Like before, the membership function is different depending on the indicator, and according to Cheli and Lemmi (1995, p. 124), dichotomous indicators are handled the same as already described in the TFA approach. Handling continuous and ordinal indicators is different according to Filippone et al. (2001, pp. 2-3) since the membership function is now defined as

$$\mu_j(v_j) = \begin{cases} H(v_j) & \text{if the degree of poverty grows as } V_j \text{ increases,} \\ 1 - H(v_j) & \text{otherwise,} \end{cases}$$

where $H(v_j)$ is the distribution function of indicator $V_j$, and the modalities of the ordinal indicators are ordered in increasing order. In the case of ordinal indicators, where the frequency associated with a category is quite high, it is advised to adopt a normalised version given by

$$\mu_j(v_j) = \begin{cases} 0 & \text{if } v_j = v_j^{(1)}, \\ \mu_j(v_j^{(r-1)}) + \dfrac{H(v_j^{(r)}) - H(v_j^{(r-1)})}{1 - H(v_j^{(1)})} & \text{if } v_j = v_j^{(r)} \text{ and } r > 1, \end{cases} \tag{11}$$

where $v_j^{(r)}$ with $r = 1, \ldots, s_j$ are the modalities of $V_j$ sorted in increasing order. The poorest individual, according to indicator $V_j$, has in the TFR approach a membership degree of one, and the richest has a membership degree of zero. By definition, the mean of the membership degrees is always 0.5 (Betti, Cheli, Lemmi and Verma, 2006, p. 118), and if it is desired to change the mean to a specified value, it is further advised to raise the membership function to some power $\alpha \geq 1$. It is added that larger values of $\alpha$ result in more weight to the poorer end of the distribution.

The membership degrees are combined into a single poverty index as before in the TFA approach with Formula 8, but now it is clear how the weights are calculated. According to Cheli and Lemmi (1995, p. 126) the weights are defined as

$$w_j = \ln\left(\frac{1}{\overline{\mu_j}}\right), \tag{12}$$

where

$$\overline{\mu_j} = \frac{1}{n}\sum_{i=1}^{n} \mu_j(v_{ij}), \tag{13}$$

represents the fuzzy proportion of poor individuals concerning the indicator $V_j$. Cheli and Lemmi (1995, pp. 126-127) note that the weighting system is a generalisation of the TFA system, and it is furthermore stated that the whole TFR approach can be seen as a generalisation of the most widespread poverty measurement techniques.

### 3.2.4 Vero and Werquin (VW)

The last fuzzy approach mentioned is the **VW approach**. The information about it is taken from Deutsch and Silber (2006, pp. 156-157).

The difference to previously explained approaches is that the poverty indices of the individuals are not weighted averages of the membership degrees of the indicators. Instead, all indicators are included in a single membership function from the beginning. The approach is explained with the following example adopted from the literature. Given are the three variables $V_1$ the individual does not have a bathroom, $V_2$ the individual does not have a car and $V_3$ the individual does not have a phone. Table 1 shows example data for $n = 6$ individuals. The values of the variables are equal to 1 if the statement is true and 0 if the statement is false. The values $f_i$ in the last column indicate the proportion

| Individual | $V_1$ | $V_2$ | $V_3$ | $f_i$ |
|---|---|---|---|---|
| 1 | 0 | 1 | 1 | 4/6 |
| 2 | 1 | 1 | 1 | 1/6 |
| 3 | 0 | 1 | 1 | 4/6 |
| 4 | 0 | 0 | 0 | 6/6 |
| 5 | 0 | 1 | 1 | 4/6 |
| 6 | 1 | 0 | 0 | 2/6 |

Table 1: Example data that is used to explain the VW approach.

of individuals who are at least as poor as individual $i$ when all indicators are considered. The proportion $f_3$ is calculated as an example. Considering all indicators, individual two is poorer, and individuals one and five are as poor as individual three. Therefore, the third individual is together with the first and the fifth, the fourth poorest of the six individuals. Notice that individual six is poor in variable $V_1$ but rich in variables $V_2$ and $V_3$, so it can only be said that this individual is richer than individual two. It is impossible to say whether this individual is more or less poor than individuals one, three and five.

With the proportions $f_i$, the deprivation indicator

$$m(i) = \frac{\ln\left(\frac{1}{f_i}\right)}{\sum_{i=1}^{n} \ln\left(\frac{1}{f_i}\right)} \quad \text{if } 0 < f_i < 1, \tag{14}$$

is then calculated for each individual $i$. Afterwards, the membership function

$$\mu(\boldsymbol{v}_i) = \frac{m(i) - \min\limits_{1\leq i\leq n}\big(m(i)\big)}{\max\limits_{1\leq i\leq n}\big(m(i)\big) - \min\limits_{1\leq i\leq n}\big(m(i)\big)}, \tag{15}$$

is used to calculate the poverty index for each individual.

A few poverty measurement approaches have now been introduced and explained in this chapter. In the next chapter, these will be discussed and compared.

## 3.3 Discussion

The previous chapter aimed to show how the poverty measurement approaches work, but the problems were not discussed.

The first thing that distinguishes the traditional approach from the others is that only a single poverty indicator, such as income or expenditure, can be used. This is problematic because according to Miceli (2006, p. 195), it is very likely that income alone tells not very much about an individual's living conditions, and the same applies to expenditures. It goes on to say that it should not be automatically considered that individuals with lower consumption expenditures are poorer, as it is the choice of each individual whether or not to buy certain goods or services or participate in certain activities. The study by Panek (2006, p. 233) on the poverty measurement in Poland confirms this. One reason why measuring poverty with a single indicator is insufficient is that each indicator reflects only one particular aspect of poverty (Miceli, 2006, p. 195). The second point that makes the traditional approach problematic is that a poverty line $z$ is required. It was mentioned in Chapter 3.2.1 that this value is chosen in reference to the mean or the median income of the population, but according to Berenger and Celestini (2006, p. 139), arbitrariness is then inherent in the identification of poor and non-poor. That any cut-off point is somewhat arbitrary is also pointed out by Mack and Lansley (1985, p. 41) with the statement "it is likely that there is a continuum of living standards from the poor to the rich".

All the other methods mentioned in the previous chapter overcome the first issue since multiple poverty indicators can be used. Still, they are not standard algorithms that can be applied to data sets without making assumptions. Using them requires fundamental understanding at each step to establish a close connection between the contents of the theoretical concepts under examination and their representation through fuzzy set theory (Martinetti, 2006, p. 112). The TFA approach is an excellent example because several decisions relating to the membership functions $\mu_j(v_j)$ need to be made, like picking a linear or a non-linear membership function between $v_j^{(L)}$ and $v_j^{(H)}$, deciding which poverty indicators need to be included, which weights $w_j$ are used or setting $v_j^{(L)}$ to zero for economically well-developed countries.

Next, the TFA approach is discussed, which differs from the other approaches in that it requires defining $v_j^{(L)}$ and $v_j^{(H)}$. One could say that this approach's weak points could also be considered as benefits depending on the literature. Therefore it could have an advantage or disadvantage over the TFR approach. On the one hand, the choice of the two thresholds is again arbitrary, and using the linear function between the threshold values is justified only based on its simplicity, without any theoretical basis or empirical evidence to support it (Cheli and Lemmi, 1995, p. 123). On the other hand, Martinetti (2006, p. 101) sees both points as an advantage because it is possible to maintain linearity, and it makes it possible to include minimum and maximum thresholds. This allows adopting the states of an indicator to different realities or circumstances. This can be explained with an example of calorie income in a well-developed and least-developed country with fictive numbers. In a well-developed country, an individual with a calorie income below 2,000 kcal might be considered deprived, while in a least-developed country, an individual with a calorie income below 1,700 kcal. If the same lower threshold $v_j^{(L)}$ is used for analysing both countries, the result would either be that a large proportion of the population would be considered poor in the least-developed country or the opposite in the well-developed country.

Another issue to note about the TFA and TFR approaches is that an individual needs the highest membership degree for each indicator to have a poverty index equal to one (Qizilbash, 2006, p. 19). This means that for an increasing number of indicators, it is less likely to have a poverty index equal to one because it is likely that an individual has, for at least one indicator, a membership degree larger than one, regardless of how important this indicator is. One could conclude that this creates the problem of selecting the right amount of indicators. Selecting the right indicators gets even more complex since the indicators' correlations are not considered in the TFA and TFR approach (Silber and Deutsch, 2005, p. 150). This can be derived by thinking about the weighting systems from both approaches. The weights are determined based on the measured values of the respective indicator. So, if there are two or more highly correlated indicators, they are not weighted down and instead treated like independent indicators.

The VW approach from Chapter 3.2.4 solves the problem of having highly correlated attributes (Silber and Deutsch, 2005, p. 150). For this reason, it is argued by Silber and Deutsch (2005, p. 170) that this algorithm may be ultimately more reliable than other approaches that ignore the problem of correlated attributes. The authors came up with this statement because, with their data, they observed differences between fuzzy poverty measurement methods that account for correlation, like the VW approach, and those that do not, like TFA and TFR. They noticed that methods that account for correlation categorised different households as poor. Since it was not proven that the VW approach is more reliable, it is added that further empirical illustrations are required to confirm this statement.

I would further argue that a disadvantage of the relative approaches TFR and VW is that, depending on the weighting system, the poverty indices of individuals from different populations are not directly comparable. This issue can appear when studies are performed in two different countries or at different times. The issue is that the membership degrees or weights are calculated relative to the population. So if the population changes, the poverty indices change too.

To sum up the criticism of fuzzy poverty measurement, it has to be said that the user's decisions, like membership function and chosen attributes, strongly impact the result. The results are indifferent if the approach has a completely different procedure, like the VW approach. In my opinion, this raises the question of whether fuzzy poverty measurement approaches are a reliable way to calculate poverty indices for individuals, as the results of the methods vary, and ultimately it is impossible to say which method best reflects reality.

The following chapter continues with the theory about logistic regression and logistic lasso regression as preparation for the BS approach and to understand a later introduced modification of logistic regression.

# 4 Regression Analysis

One main goal of regression analysis is to analyse the influence of the independent variables $x_{i1}, \ldots, x_{ik}$ on the mean value of the dependent variables $y_i$, $i = 1, \ldots, n$. So, the conditional expected value $\mathrm{E}(y_i \mid x_{i1}, \ldots, x_{ik})$ of $y_i$ is modelled as a function of the independent variables $x_{i1}, \ldots, x_{ik}$. Note that categorical variables are assumed to be dummy coded from now on. For the class of linear regression models with a linear function $f$, it then applies

$$\mathrm{E}(y_i \mid x_{i1}, \ldots, x_{ik}) = f(x_{i1}, \ldots, x_{ik}) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}, \tag{16}$$

where $\beta_0, \ldots, \beta_k$ are the coefficients. $\beta_0$ is referred to as intercept, and $\beta_1, \ldots, \beta_k$ are the slopes. As a result, for given data $(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i)$ it then applies that

$$y_i = \mathrm{E}(y_i \mid x_{i1}, \ldots, x_{ik}) + \varepsilon_i, \tag{17}$$

whereby $\varepsilon_i$ the random errors are that the independent variables can not explain. It applies that $\varepsilon_i \sim N(0, \sigma^2)$ with $V(\varepsilon_i) = \sigma^2$. The random errors $\{\varepsilon_i \mid i = 1, \ldots, n\}$ are assumed to be independent and identically distributed. In the regression analysis, it is then the task to estimate the linear function $f$ for the given data (Fahrmeir et al., 2009, pp. 19-21).

In BS, the dependent variables $y_i$ take the values one and zero. However, the linear regression model is recommended only when the dependent variables $y_i$ are continuous and ideally approximately normally distributed (Fahrmeir et al., 2009, p. 30). For this reason, **logistic regression** is used instead. Logistic regression is explained in more detail in the following chapter, as it is an important part of the BS procedure. Logistic lasso regression will be introduced afterwards.

## 4.1 Logistic Regression

The information in this chapter is mainly taken from Fahrmeir et al. (2009, pp. 189-201) if not differently declared.

For binary dependent variables $y_i \in \{0, 1\}$, it is now the goal to model and estimate the effect of the independent variables $x_{i1}, \ldots, x_{ik}$ on the conditional probability

$$\pi_i = \mathbb{P}(y_i = 1 \mid x_{i1}, \ldots, x_{ik}) = \mathrm{E}(y_i \mid x_{i1}, \ldots, x_{ik}). \tag{18}$$

The second relation holds because $y_i$ are binary, and then the following applies

$$\begin{aligned} \mathrm{E}(y_i \mid x_{i1}, \ldots, x_{ik}) &= 1 \cdot \mathbb{P}(y_i = 1 \mid x_{i1}, \ldots, x_{ik}) + 0 \cdot \mathbb{P}(y_i = 0 \mid x_{i1}, \ldots, x_{ik}) \\ &= \mathbb{P}(y_i = 1 \mid x_{i1}, \ldots, x_{ik}). \end{aligned} \tag{19}$$

Because the conditional probabilities $\mathbb{P}(y_i = 1 \mid x_{i1}, \ldots, x_{ik})$ take a value in $[0, 1]$ and further reasons mentioned by Fahrmeir et al. (2009, pp. 30-31), there is not the exact same relation between the conditional expectation $\mathrm{E}(y_i \mid x_{i1}, \ldots, x_{ik})$ and the linear function of the independent variables $f(x_{i1}, \ldots, x_{ik})$ that could be noticed in the linear regression model. Instead, a strictly increasing function $h$, called the response function, on the range $[0, 1]$ is used to derive a function of the dependent variables $f(x_1, \ldots, x_k)$ on this range. In the case of logistic regression, this function is defined as

$$h : \mathbb{R} \longrightarrow [0, 1],\ \eta \longmapsto \frac{\exp(\eta)}{1 + \exp(\eta)}, \tag{20}$$

and known as the standard logistic function. This leads to the logistic model

$$\begin{aligned} \pi_i &= \mathbb{P}(y_i = 1 \mid x_{i1}, \ldots, x_{ik}) \\ &= \frac{\exp(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik})} \\ &= \frac{\exp(\eta_i)}{1 + \exp(\eta_i)} \\ &= h(\eta_i), \end{aligned} \tag{21}$$

whereby $\eta_i$ is called the linear predictor.

The maximum likelihood method is used to estimate the coefficients $\beta_0, \ldots, \beta_k$. For the assumption that the the binary dependent variables $y_i$, $i = 1, \ldots, n$ are Bernoulli distributed $y_i \sim B(1, \pi_i)$, the likelihood function that is used for estimation is

$$\mathcal{L}(\beta_0, \ldots, \beta_k) = \prod_{i=1}^{n} \pi_i^{y_i} (1 - \pi_i)^{1 - y_i}. \tag{22}$$

Because of Equation 21, the likelihood function is a function of the coefficients $\beta_0, \ldots, \beta_k$. Taking the logarithm of the likelihood function to receive the log-likelihood function, which simplifies maximisation, and plugging in $\pi_i$ leads to the different form

$$\begin{aligned} \ell(\beta_0, \ldots, \beta_k) &= \ln \left( \prod_{i=1}^{n} \pi_i^{y_i} (1 - \pi_i)^{1 - y_i} \right) \\ &= \sum_{i=1}^{n} y_i \ln(\pi_i) + (1 - y_i) \ln(1 - \pi_i) \\ &= \sum_{i=1}^{n} y_i \ln \left( \frac{\exp(\boldsymbol{x}_i^t \boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_i^t \boldsymbol{\beta})} \right) + (1 - y_i) \ln \left( 1 - \frac{\exp(\boldsymbol{x}_i^t \boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_i^t \boldsymbol{\beta})} \right), \end{aligned} \tag{23}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)^t$ and $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{ik})^t$. The next step is to identify the values of $\beta_0, \ldots, \beta_k$ that maximise the log-likelihood function. These values are estimations for a given data set with the individual observations

$$(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i) = (\boldsymbol{x}_i, y_i), \quad i = 1, \ldots, n. \tag{24}$$

To receive the estimations, the gradient of the log-likelihood function with respect to $\beta_0, \ldots, \beta_k$, which is the score function, is set equal to zero to get the maximum of the log-likelihood function. According to Fahrmeir et al. (2009, p. 199), the resulting ML-equation is

$$\boldsymbol{s}(\hat{\boldsymbol{\beta}}) = \frac{\partial \ell(\hat{\boldsymbol{\beta}})}{\partial \hat{\boldsymbol{\beta}}} = \sum_{i=1}^{n} \boldsymbol{x}_i \left( y_i - \frac{\exp(\boldsymbol{x}_i^t \hat{\boldsymbol{\beta}})}{1 + \exp(\boldsymbol{x}_i^t \hat{\boldsymbol{\beta}})} \right) = \mathbf{0}. \tag{25}$$

This non-linear system of equations for $\hat{\boldsymbol{\beta}}$ is usually solved iteratively by the Fisher Scoring algorithm or another iterative numeric algorithm because it is difficult to solve this equation analytically.

The received estimations $\hat{\beta}_0, \ldots, \hat{\beta}_k$ are different to interpret compared to the linear model where a linear relationship between the independent variables $x_{i1}, \ldots, x_{ik}$ and the conditional expected value $\mathrm{E}(y_i \mid x_{i1}, \ldots, x_{ik})$ is present. In the linear regression model, the interpretation is as follows. If $x_{ij},\ j \in \{1, \ldots, k\}$ is increased by one unit in the linear model, the expected value of $y_i$ increases by $\beta_j$, suppose all other variables remain constant. The non-linearity introduced by the response function in the logistic regression model makes it just possible to interpret the effects on the odds or log odds. This can be seen with the equation

$$g(h(\eta_i)) = \ln\left(\frac{\pi_i}{1 - \pi_i}\right) = \eta_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}, \tag{26}$$

where $g = h^{-1}$ and the odds being defined as

$$\frac{\mathbb{P}(y_i = 1 \mid x_{i1}, \ldots, x_{ik})}{\mathbb{P}(y_i = 0 \mid x_{i1}, \ldots, x_{ik})} = \frac{\pi_i}{1 - \pi_i}. \tag{27}$$

Further, a multiplicative interpretation of the odds

$$\frac{\pi_i}{1 - \pi_i} = \exp(\beta_0) \exp(\beta_1 x_{i1}) \cdot \ldots \cdot \exp(\beta_k x_{ik}), \tag{28}$$

is possible too. A modification of the logistic regression, the **logistic lasso regression**, will be explained in the upcoming chapter.

## 4.2 Logistic Lasso Regression

Unless otherwise stated, the information regarding lasso regression was taken from Friedman et al. (2010, pp. 1-9).

The technique least absolute shrinkage and selection operator (lasso) shrinks some coefficients and sets others to zero (Tibshirani, 1996, p. 267). According to Tibshirani (1996, p. 268), this can be useful for two reasons. Firstly, the prediction accuracy of the lasso regression model can exceed that of the ordinary regression model. The second, in the context of BS more important reason, is interpretability, which is easier if fewer variables are included in the model. Logistic lasso regression can therefore be used to select a smaller subset of variables that still exhibits the strongest effects. Said differently, logistic lasso regression will be used for variable selection. This method uses an L1-penalty to achieve a sparse solution by forcing the absolute values of the coefficients to be smaller than a specific value

$$\sum_{j=1}^{k} |\beta_j| \leq t. \tag{29}$$

The value $t \geq 0$ controls the shrinkage applied to the estimates (Tibshirani, 1996, p. 268).

This restriction is incorporated by adding a penalty term to the log-likelihood function from Equation 23, resulting in the log-likelihood function for logistic lasso regression

$$\begin{aligned} \ell(\beta_0, \dots, \beta_k) = \frac{1}{n} \sum_{i=1}^{n} & y_i \ln \left( \frac{\exp(\boldsymbol{x}_i^t \boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_i^t \boldsymbol{\beta})} \right) \\ & + (1 - y_i) \ln \left( 1 - \frac{\exp(\boldsymbol{x}_i^t \boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_i^t \boldsymbol{\beta})} \right) \\ & - \lambda \sum_{j=1}^{k} |\beta_j|. \end{aligned} \tag{30}$$

The parameter $\lambda \geq 0$ is called the **tuning parameter** and controls the amount of shrinkage. It is related to $t$, but the exact relation will not be explained. To get the estimations for the coefficients, maximum likelihood analysis is performed. It is noted here that the additional factor $\frac{1}{n}$ in front of the log-likelihood function that is not there in Equation 23 should not affect the minimisation in my view since it is a positive constant.

Understanding logistic lasso regression is not crucial to be able to use logistic lasso regression for variable selection, but the concept of adding a penalty term needs to be understood as it is later used to build a modified logistic regression model.

This chapter has now provided the basics needed to understand BS, which is introduced in the following chapter.

# 5 Boundary Shifts (BS)

In BS, different poverty lines are placed according to the variable income, and individuals with income below each poverty line are considered poor. The proposed approach has a fuzzy view on poverty measurement since not just a single fixed poverty line is used to divide the population into poor and non-poor, like in the traditional approach, but instead, there are many poverty lines at different income values. Thus there is no fixed poverty line but many instead. The consequence of the shifting poverty line is that an increasing or decreasing number of individuals are considered poor. To highlight that not a single poverty line is used and the concept is different, the poverty lines will be called **boundaries** in the BS approach.

What also differentiates BS from the traditional approach is that after dividing the data set into poor and non-poor, a binary prediction model is applied to the data set using the additional variables provided in the data set. The parameter estimations of the binary prediction model can be interpreted to understand the drivers of poverty better. For example, suppose the logistic regression model is used, and the coefficients are estimated at many boundaries. In that case, it can be interpreted how a change of one unit of some independent variable impacts the odds of being poor at each boundary. This shows how the influence of an independent variable for dividing the population into poor and non-poor changes when the boundary is shifted upwards, and more individuals are considered poor.

The data notation is now simplified and defined as follows. Each individual $i$ out of $n$ individuals possesses a vector

$$(x_1, x_2, \dots, x_k, y) = (\boldsymbol{x}, y). \tag{31}$$

After dummy coding of the categorical variables, there are $k$ poverty indicators. According to the dependent variable $y$, individuals are separated into poor and non-poor. In the following, will $y_i$ always correspond to the income measurement of individual $i$.

There are three steps in the BS procedure

- boundary placement,
- creation of boundary-dependent data sets,
- fitting binary prediction model on the dependent data sets.

The steps are are explained in detail in the following.

## 5.1 Boundary Placement

In the first step, the boundaries $z_t$, $t = 1, \dots, T$ are defined. Since the data is split according to the dependent variable $y$, it is reasonable to demand the conditions

$$0 \leq |\{z_t \mid y_{(l)} < z_t < y_{(l+1)}\}| \leq 1, \tag{32}$$

with $y_{(l)}, y_{(l+1)}$, $l \in \{1, \dots, n-1\}$ being neighbouring ordered measurements and

$$z_t = z_{t'} \iff t = t', \tag{33}$$

where $t' = 1, \dots, T$. This ensures that two boundaries do not result in the same separation of poor and non-poor since there is a maximum of one boundary $z_t$ placed between two income measurements $y_{(l)}, y_{(l+1)}$. Further, it is guaranteed that there are no boundaries below the income poorest or above the income richest individual, resulting in a separation where all individuals are assigned to the poor or the non-poor. This is necessary for fitting the logistic regression model. It follows from these conditions that there are a maximum of $n-1$ different assignments of the individuals to the set of poor as that amount of sensible boundaries exist. The last thing that follows from this condition is that no boundary $z_t$ corresponds to an observed income. If this were otherwise, it would be unclear whether an individual should be classified as poor or non-poor.

## 5.2 Boundary Dependent Data Sets

As there are now $T$ boundaries, there can be $T$ different separations into poor and non-poor, which has $T$ different data sets as a consequence. In each data set $D_t$ then is, according to its boundary $z_t$, the membership for each individual to the set of poor calculated with

$$\tilde{y}_i = \mu(y_i) = \begin{cases} 1 & \text{if } y_i < z_t, \\ 0 & \text{if } y_i \geq z_t. \end{cases} \tag{34}$$

Since the income measurements $y_i$ are no longer required after calculating the memberships, they are removed.

For the assumption that the data set is sorted according to the income and $t = 1, \dots, T$, the structure of the data is

$$D_t = (X, \tilde{Y}_t), \tag{35}$$

with the matrix

$$X = (x_{ij})_{1 \leq i \leq n,\, 1 \leq j \leq k} \in \mathbb{R}^{n \times k}, \tag{36}$$

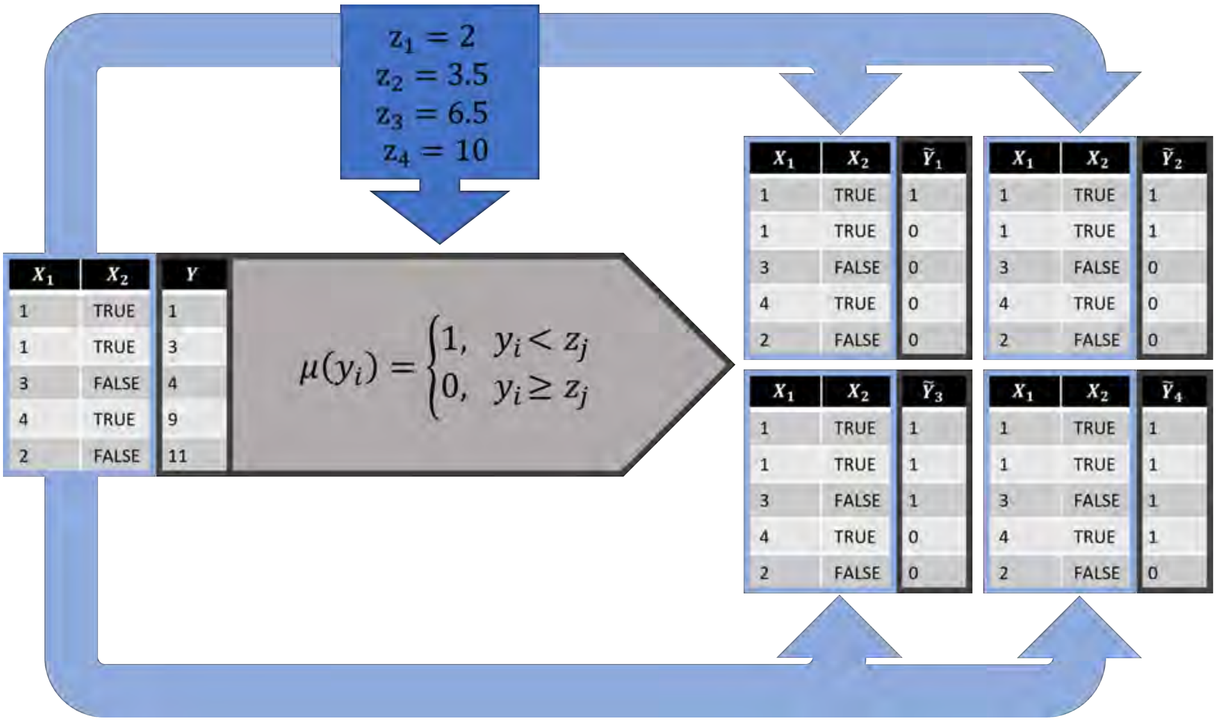
and

$$\tilde{Y}_t = \Big(1, \quad \dots, \quad \tilde{y}_{(l)}, \quad \tilde{y}_{(l+1)}, \quad \dots, \quad 0\Big)^t \in \{0,1\}^{n \times 1}, \tag{37}$$

where $l = 2, \dots, n-2$ and $\tilde{y}_{(l)} \geq \tilde{y}_{(l+1)}$. It can be seen that the income poorest individual is always assigned to the set of poor, while the opposite is true for the income richest individual because the income value $y_{(1)}$ of the income poorest individual is below every boundary $z_t$ and the income of the richest individual $y_{(n)}$ is above.

The process of getting multiple data sets is visualised in Figure 1. In this example, the data set contains five observations of three variables and the data is split according to variable $Y$ with four boundaries. After the splitting processes, there are four modified data sets. The measurements of the variables $X_1$ and $X_2$ are identical in the data sets, but the measurements of $Y$ changed according to the boundaries.

As there are now $T$ different data sets, the parameters of a binary prediction model can be estimated.


$z_1 = 2$
$z_2 = 3.5$
$z_3 = 6.5$
$z_4 = 10$

| $X_1$ | $X_2$ | $Y$ |
|---|---|---|
| 1 | TRUE | 1 |
| 1 | TRUE | 3 |
| 3 | FALSE | 4 |
| 4 | TRUE | 9 |
| 2 | FALSE | 11 |

$$\mu(y_i) = \begin{cases} 1, & y_i < z_j \\ 0, & y_i \geq z_j \end{cases}$$

| $X_1$ | $X_2$ | $\tilde{Y}_1$ |
|---|---|---|
| 1 | TRUE | 1 |
| 1 | TRUE | 0 |
| 3 | FALSE | 0 |
| 4 | TRUE | 0 |
| 2 | FALSE | 0 |

| $X_1$ | $X_2$ | $\tilde{Y}_2$ |
|---|---|---|
| 1 | TRUE | 1 |
| 1 | TRUE | 1 |
| 3 | FALSE | 0 |
| 4 | TRUE | 0 |
| 2 | FALSE | 0 |

| $X_1$ | $X_2$ | $\tilde{Y}_3$ |
|---|---|---|
| 1 | TRUE | 1 |
| 1 | TRUE | 1 |
| 3 | FALSE | 1 |
| 4 | TRUE | 0 |
| 2 | FALSE | 0 |

| $X_1$ | $X_2$ | $\tilde{Y}_4$ |
|---|---|---|
| 1 | TRUE | 1 |
| 1 | TRUE | 1 |
| 3 | FALSE | 1 |
| 4 | TRUE | 1 |
| 2 | FALSE | 0 |

Figure 1: Example data set with attributes $X_1$, $X_2$ and income $Y$ is transformed into four modified data sets with the same $X_1$, $X_2$ measurements but binary variable $\tilde{Y}_i$ with values depending on the boundaries $z_1, \dots, z_4$.

## 5.3 Fitting the Binary Prediction Model

In BS, the logistic regression model is used as the primary binary prediction model due to its good interpretability of its parameters and simplicity compared to other methods. Therefore, logistic regression is used to explain the concepts of BS. In subsequent chapters, modifications of logistic regression models are used, but the concept of BS remains the same.

There are now different data sets $D_t$ with binary dependent variables $\tilde{y}_t$. Logistic regression can be conducted on each data set $D_t$ separately to estimate the coefficients $\boldsymbol{\beta}(z_t)$. In each model $x_1, \dots, x_k$ are the independent variables.

This means $\boldsymbol{\beta}(z_t)$ is a function of $z_t$. In the following chapters, $\boldsymbol{\beta}(z_t)$ is often written as $\boldsymbol{\beta}$ for better readability. Still, it should be remembered that the coefficients depend on the boundaries in BS. After estimating the coefficients $\hat{\boldsymbol{\beta}}(z_t)$, the coefficients can be graphically displayed with the corresponding boundary $z_t$.

A concern is noted at this point. It was said in Chapter 3.3 that using income as the only poverty indicator is questionable. Nonetheless, this is done in the first step. It is assumed that this is fine as more poverty indicators are used in the third step.

The procedure is now clear, and it is continued with the introduction of the data set.

# 6 Data Introduction

## 6.1 Data Set: Filipino Family Income and Expenditure

Taken is the data set used for the sensitivity analysis from the website Kaggle. Kaggle is the most popular data science competition platform (Banachewicz et al., 2022, p. 33) and allows users, among other things, to publish data sets. The data set is called "Filipino Family Income and Expenditure" (Flores, n.d.). It is according to the information on the website based on a survey from the Philippine Statistics Authority to provide data on family income and expenditure. This survey is conducted every three years. It is assumed that this data set resulted from the survey in 2015 since it was uploaded in 2017, and it is claimed that the data is from the latest family income and expenditure survey in the Philippines.

The data set has a usability score of 7.06 out of 10, according to Kaggle, since it has not been updated and the source is not clear. Furthermore, the file and column descriptions are missing. In addition to the points criticised by Kaggle, it is unclear what survey design was chosen, and in particular, it is not known how a unit or item non-response was dealt with. It still has been taken for the following reasons

- amount of observations,
- amount of variables,
- contains continuous and categorical variables,
- clear and self-explaining variable names,
- no missing values,
- data file format,
- relationship to poverty measurement.

With measurements of 41,544 households, it can be assumed that the data set represents the Filipino population well. Further, with over 60 variables, the data contains a lot of information regarding each individual's living conditions. Having a mixture of variable types is for the sensitivity analysis of interest since they are included differently in the logistic regression model. The self-explanatory variable names are necessary to compensate for the missing description of the variables so that a subsequent interpretation is possible. Although NAs are in the data set, there is actually no missing data, which will be explained in the next chapter. This is advantageous since imputation methods that need assumptions about the missing data mechanism are not required. The data set is supplied in tabular style, as a single comma-separated values (CSV) file, so there is no need to combine any data sets which could result in missing values. The last reason for taking this data set is that this data set is survey data collected for poverty measurement, which fits the topic of the thesis. However, I would argue that BS theoretically be used on any other data set having a continuous or ordinal variable to assess any vague predicate.

Since the origin of the data is now clear, the data set will be processed in the following chapter, and exploratory data analysis will be made.

## 6.2 Data Preparation and Exploratory Data Analysis

The following tasks are executed in RStudio (Posit team, 2023), which is an integrated development environment for the programming language and environment for statistical computing, R (R Core Team, 2022).

The data preparation involves five steps. The first step is to change the data types of the variables. The categorical variables are transformed into factors because these are supplied as the data type character. The continuous variables are supplied as integers, so no adjustment is required. The second step is to abbreviate or rename some variables and values of the categorical variables. This is not required, but it makes later analyses and graphics easier to read. In the third step, the following possible error is corrected. The variable “Highest Grade” contains the two categories “Engineering and Engineering trades Programs” and “Engineering and Engineering Trades Programs” written identically except for a capital letter. The two categories seem to result from a spelling mistake and are therefore combined into one variable.

The fourth step consists of combining two variables. The first variable is “Class of Worker”, which indicates the state of working showing if someone is self-employed, owns a family-operated farm or family business, or for whom the individual works, and in the case of a family-operated farm or business, whether the individual is paid or not, resulting in seven categories. But, some households are not assigned to any of these categories resulting in NAs. The second variable, called “Job Business”, is binary and indicates whether or not the head of the household has a job or business. Now, observations with a NA value for the “Class of Worker” variable have no job or business, and vice versa. This makes the variable “Job Business” redundant. By adding the category “No Job/Business” to the variable “Class of Worker”, the variable “Job Business” can be removed. The last step is to standardise the continuous variables, but working with interpreting standardised regression coefficient is controversial, according to Bring (1994, p. 209). The following reasons why working with interpreting standardised is controversial are taken from Bring (1994). Afterwards, it is explained why the data set is still standardised.

Comparing regression coefficients with regard to the size is a typical modelling goal, but it is not easy when variables are measured in different units. Standardising seemingly overcomes this issue as the standardised variables are measured in the same units, the standard deviations. The standardised coefficients are then interpreted as the standard deviation change in the dependent variable when the independent variable is changed by one standard deviation if all other variables are held constant.

Using the standardised coefficients to assess **relative importance** is natural since the standardised coefficient is related to the variables’ contribution to the prediction of $y$. It could be concluded that the more a variable contributes to the prediction of $y$, the more important it is. However, it is said that the question of how to quantify the contribution to the prediction of $y$ is left open in this view. Therefore the interpretation

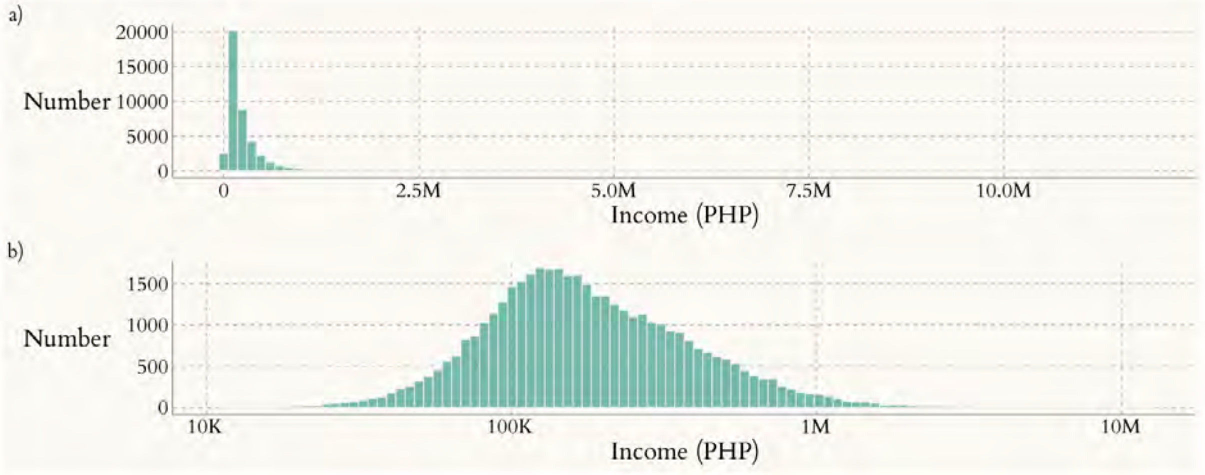


Figure 2: Logarithmically scaled plots are used in further analyses since they give a better overview.

of relative importance is unclear. More reasons against using standardised coefficients are that they are difficult to interpret and sample-specific, making them unreliable for comparing different samples.

Despite the reasons against using standardised coefficients, it is said that relative importance is a diffuse concept that can have many different meanings. Therefore a single measure of relative importance that can be used in all situations cannot be recommended.

Due to the last defusing statement and the following reason, standardised data will still be used in BS. It is not aimed at assessing the relative importance as an absolute value. Moreover, it is of interest in which direction the values of the estimated coefficients change when the boundaries are shifted, and that is more obvious if the different units are brought to the same unit.

The following exploratory data analysis provides an initial overview of the data set. In particular, the variable “Income”, measured in Philippine peso (PHP), is looked at in detail since the boundaries are placed dependent on this variable. Primarily, the correlations between the variables and the distribution of the categorical variables are analysed, and additional assumptions are made about why the correlations are high or low. In this thesis, the absolute correlation values are not given in a table but can be extracted from the electronic appendix.

The histogram in Figure 2 a) shows that the observed income of most households is just a fraction of the maximal observed income. Further, the median income of 164,080 is far below the maximum income of 11,815,988. Since measuring poverty is the main objective, giving greater attention to those with lower incomes makes sense. This is visually done by scaling the x-axis with the decadic logarithm as the lower incomes are then more spread across the x-axis, and the higher are more compressed. The histogram with the scaled x-axis is shown in Figure 2 b). Note that this is just a visual transformation. The positive skew in the histogram shows that the mean income is higher than the median

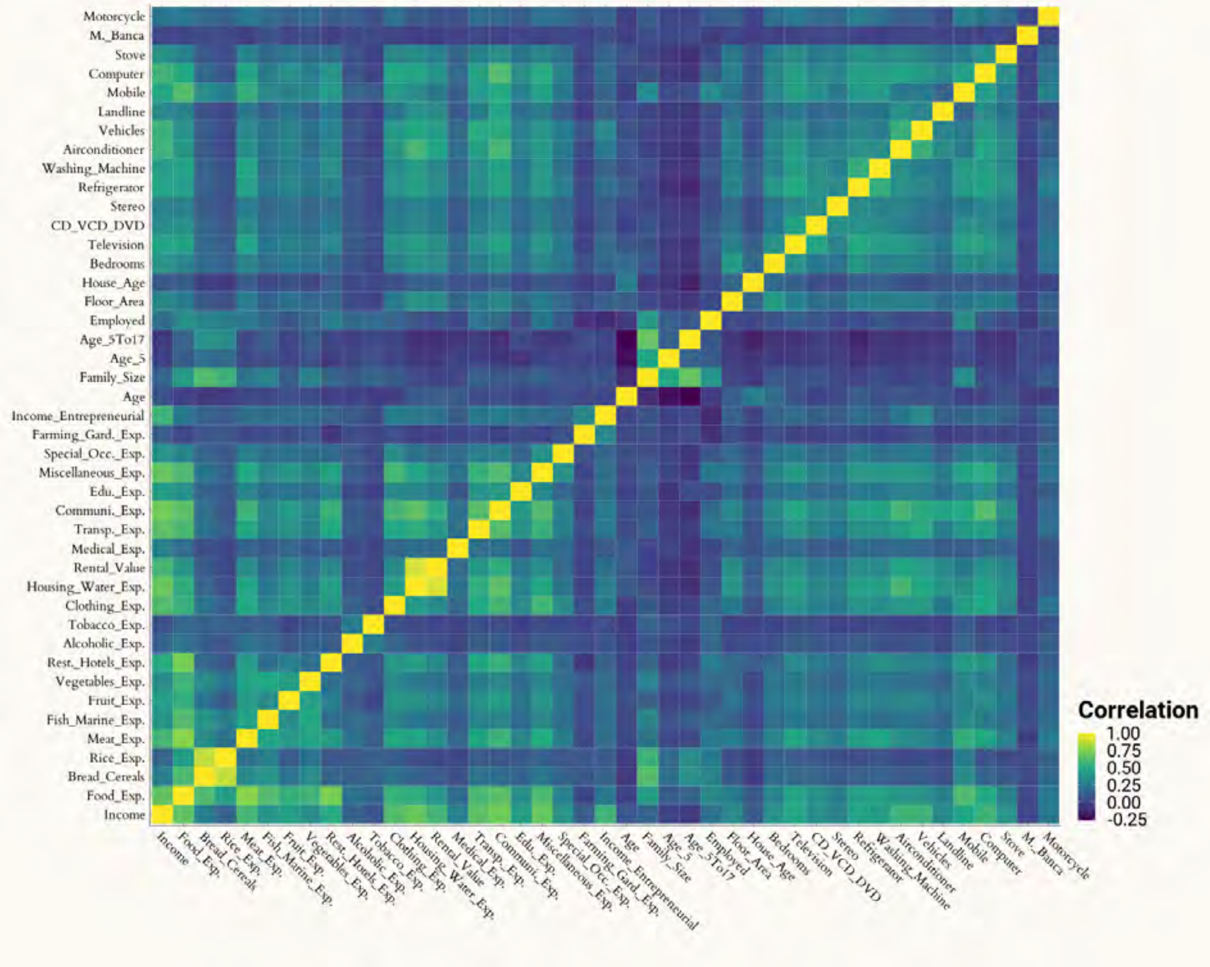


Figure 3: Heat map that visualises the correlation between the numerical variables. With a few exceptions, one can notice clusters of higher correlated variables. Expenditure variables are generally higher with each other correlated, but these variables are also relatively high correlations with variables corresponding to household machines or communication devices.

income. Furthermore, the figure shows that 90 per cent of the households have an income higher than approximately 56,000 and lower than 700,000, so the later analysis should be concentrated on this range. The usage of units is avoided from here on in the text. It is safe to conclude that income and expenditures are given in PHP.

The heat map from Figure 3 shows the correlation between the numerical variables of the data set. It can be seen that the correlations are mainly positive or slightly negative. Exceptions are the variables “Age 5to17” and “Age 5”, which indicate how many children between five and 17 years and how many children under five years live in a household, as they show the strongest negative correlation in combination with the variable “Age”, which corresponds to the age of the head of the household. One may argue that this is because as the age of the head of the household grows, so do the children. The consequence of this would be that there are fewer children in the household.

Highly correlated is “Housing Water Expenditure”, describing the water expenditures, with “Rental Value”, giving the imputed house rental value, with a correlation of 0.92, followed by the correlation of 0.88 between the variables “Rice Expenditure” and “Bread Cereals Expenditure”. The heat map shows that all expenditure variables are very correlated, excluding expenditures for legal drugs, like tobacco and alcohol, and the already mentioned for rice and bread or cereals. One could conclude that drugs are not highly correlated with other expenditures because they are consumed by the poor and the non-poor with similar frequency. The same may apply to basic foods like bread and rice. Therefore, this relation is further analysed when the income correlations are discussed. The variables that indicate the number of things a household owns, excluding “Motorised Banca”, are also stronger correlated with the expenditures. One may conclude from this small correlation of the variable “Motorised Banca” and the other expenditure variables that having one or more bancas, small boats, is independent of whether a household is poor or not. This is a reasonable conclusion since owning one or more bancas is useless for a household if it does not live near the water.

Looking at the correlations with income reveals that it correlates strongly with the expenditure variables, except for the few already mentioned. This is plausible since households with higher incomes can afford more luxury goods. Another fact is that income has a correlation of approximately zero with the variables “Age 5to17” and “Age 5”, which could mean that rich and poor have a similar amount of children. It needs further analysis to confirm this.

As previously seen with the low correlation of expenditure on rice and other expenditures, income is just slightly correlated with expenditure on rice, with a correlation of 0.16. This could mean the poor and non-poor spend similar amounts on this food. Because rice is a fairly cheap product compared to other foods, poor households can afford this food source. The higher the income, the more households can afford; therefore, the correlation might be slightly positive. For the wealthier households, the spending on rice no longer increases because these households can afford other, more expensive foods such as meat, which has a higher correlation with income. For wealthy households, spending on rice goes back down a bit, as they are likely to buy more expensive food instead of rice. The fact that food expenditure and income show a correlation of 0.66 supports this thesis, indicating that households need to compensate for their lower expenditure on rice. Figure 23 in the appendix shows the just described relationship of expenditure on rice and income. In this figure, the scaling of the two axes is logarithmic to account for very large values.

“Income Entrepeneurial”, which gives the total income of entrepreneurial activities, has a correlation of 0.56 with income but is at the same time not that correlated with the variables that are also highly correlated with income. Entrepreneurial activity is, according to the family income and expenditure survey from 2015 (Bersales, 2017, p. 81), any economic activity, business or enterprise, whether in agriculture or the non-agriculture sector, engaged in by any member of the family as an operator or as self-employed, operated by any family member as a self-employed individual or a single proprietorship, excluding formal partnerships, corporations, and registered associations. The relatively low correlation with expenses may be due to reinvesting the income in the company and not spending on excessive luxuries, but this is only speculation.

It can further be noticed that the variables, which will be derived in Chapter 7.2 and will be relevant later on, do not correspond to the ten variables with the highest correlation with income. Without pointing out which variables are involved, it is highlighted here that the variable "Food Expenditure" is the only variable in both sets. This may mean that the later used variables are good at describing all dimensions of poverty as not only variables with a strong link to income are used. But this is also pure speculation.

The last important correlation highlighted is the one between the variables "Income" and "Family Size" since the income in the data set is not individual but household related. So, with the increasing size of the family inside a household, the income should increase since it is likely that more individuals are working. Figure 24 in the appendix shows this as the family size increases approximately until the income of 110,000. Then the family does not further increase with increasing income. The correlation between income and family size is 0.37 when only households with incomes below 110,000 are considered and 0.06 when households above 110,000 are considered. This confirms that the family size increases until a certain income and afterwards remains constant.

Knowing this relation between the variables "Income" and "Family Size" allows for further speculations. One conjecture is that the variable "Mobile" might later, when BS is applied, have a high relative importance for dividing the population into poor and non-poor at low thresholds. For increasing boundaries, the relative importance could decrease but remain medium important. The reason for this speculation is that households with low incomes can not afford a mobile phone at all. If a household's income and family size increase, it can be assumed that more income must be invested in food and order essential things. Therefore, many households can still not afford one or more mobile phones.

Besides the numerical variables, the data set contains 16 categorical variables. The variables "Class of Worker", "Occupation" and "Region" are not discussed as they will not be added to the BS model. These variables contain a lot of categories which makes the plots hardly understandable. I would expect that at least occupation could be useful for poverty measurement. However, each category results in one estimated coefficient due to dummy coding in the regression model. In the case of occupation, this would result in 378 estimated coefficients, which can hardly be visualised.

The only binary variables are the gender of the head of the family (Bersales, 2017, p. 127) and Electricity, indicating that electricity is used in this household (Bersales, 2017, p. 179). For both variables, the categories are unbalanced, as almost 90 per cent of all households have electricity, and the head of household is male in about 78 per cent of households. The variable "Electricity" might be good for dividing into poor and non-poor as households with electricity have a 2.5 times higher mean income, and further, a household without it is quite disadvantaged. The same does not apply to the variable "Sex" since the mean is in both categories approximately the same.

The other variables have between three and eleven categories. More than half of the variables are dominated by one class. An example of this is the variable “Building” that indicates in which type of building or house the family resides (Bersales, 2017, p. 178). For this variable, one can see that 94 per cent of the households reside in a single house. A similar pattern applies to the variables “Martial Status”, “Roof”, “Tenure Status”, “Toilet”, “Walls” and “Water Supply”.

I came up with four reasons for small class frequencies. The first possible reason is that the category is very rare because it is very special, like the building category houseboat. It is reasonable to infer that an alternative building class within the same price segment exists, but it is not chosen for mostly money-unrelated reasons. The second reason is that something is so cheap that almost everyone but the very poorest can afford it. This probably applies to the variable toilet, where most households have some type of water-sealed toilet, and only a small number have an open or closed pit toilet or no toilet at all because these households can probably not afford to buy one. The third reason for small class frequencies is the opposite of the latter, as there are wealthy households that can afford luxury goods that are too expensive for most of the population. The last explanation for low category frequencies is that something is neither expensive nor necessary for survival, such as items held only by collectors and so infrequently owned.

The exploratory data analysis now gave a good overview of the variables included and their relationship to income to show how a boundary shift might affect the variable importance in BS.

The discussion of the first BS modelling approach occurs in the following chapter. This model contains all variables except the one already mentioned.

# 7 Sensitivity Analyses

In this chapter, mainly the fluctuation of the estimated coefficients is analysed. It is shown how different variable types impact the analysis and how different boundary placement and binary prediction models impact the BS models. A first BS model is used to explain some issues of BS and what tasks lie ahead.

From here on, for the sake of simplicity, the variables are addressed directly by their names in the formulas. In the text, the variables are not enclosed in quotes.

## 7.1 First BS Model

For the **first BS model**, the whole data set is used with all variables except class of Worker, Occupation and Region. Over the range of minimum income, $\min\limits_{1\le i\le n}(Income_i)$, and maximum income, $\max\limits_{1\le i\le n}(Income_i)$, there are $T = 998$ boundaries $z_t$, $t = 1, \dots, T$ placed with the formula

$$z_t = \exp\left(\lg\left(\min_{1\le i\le n}(Income_i)\right) + t \cdot \frac{\lg\left(\min\limits_{1\le i\le n}(Income_i)\right) - \lg\left(\max\limits_{1\le i\le n}(Income_i)\right)}{T+1}\right). \tag{38}$$

This special boundary placement accounts for the income distribution by using the decadic logarithm and scaling with the exponential function. One should note here that households with incomes equal to a boundary are assigned to the non-poor due to the R code. Since nearly all variables are included, there result 103 **poverty curves** from the estimated coefficients of the 998 boundaries. The number of poverty curves results from

- 42 continuous variables,
- 60 dummy variables generated from 13 categorical variables,
- intercept.

The term "poverty curves" is defined as follows.

**Definition 7.1** (Poverty curve)**.** Provided that the coefficients $\beta_j(z_t)$ with $j = 0, \dots, k$ are estimated for all boundaries $z_t$ with $t = 1, \dots, T$. If the estimated coefficients $\hat{\beta}_j(z_t)$ with associated $z_t$ are plotted as points in a coordinate system, the graph produced by visually connecting the points of the same coefficient is called the poverty curve of coefficient $\beta_j$.

For simplicity, each poverty curve will be directly addressed by the variable name if the poverty curve is drawn from slope estimations, i.e., the poverty curve resulting from the estimated coefficients of the independent variable mobile will be referred to as mobile poverty curve or the poverty curve of the variable mobile. The poverty curve resulting from the intercept estimations will be called the intercept poverty curve or the poverty curve of the intercept.

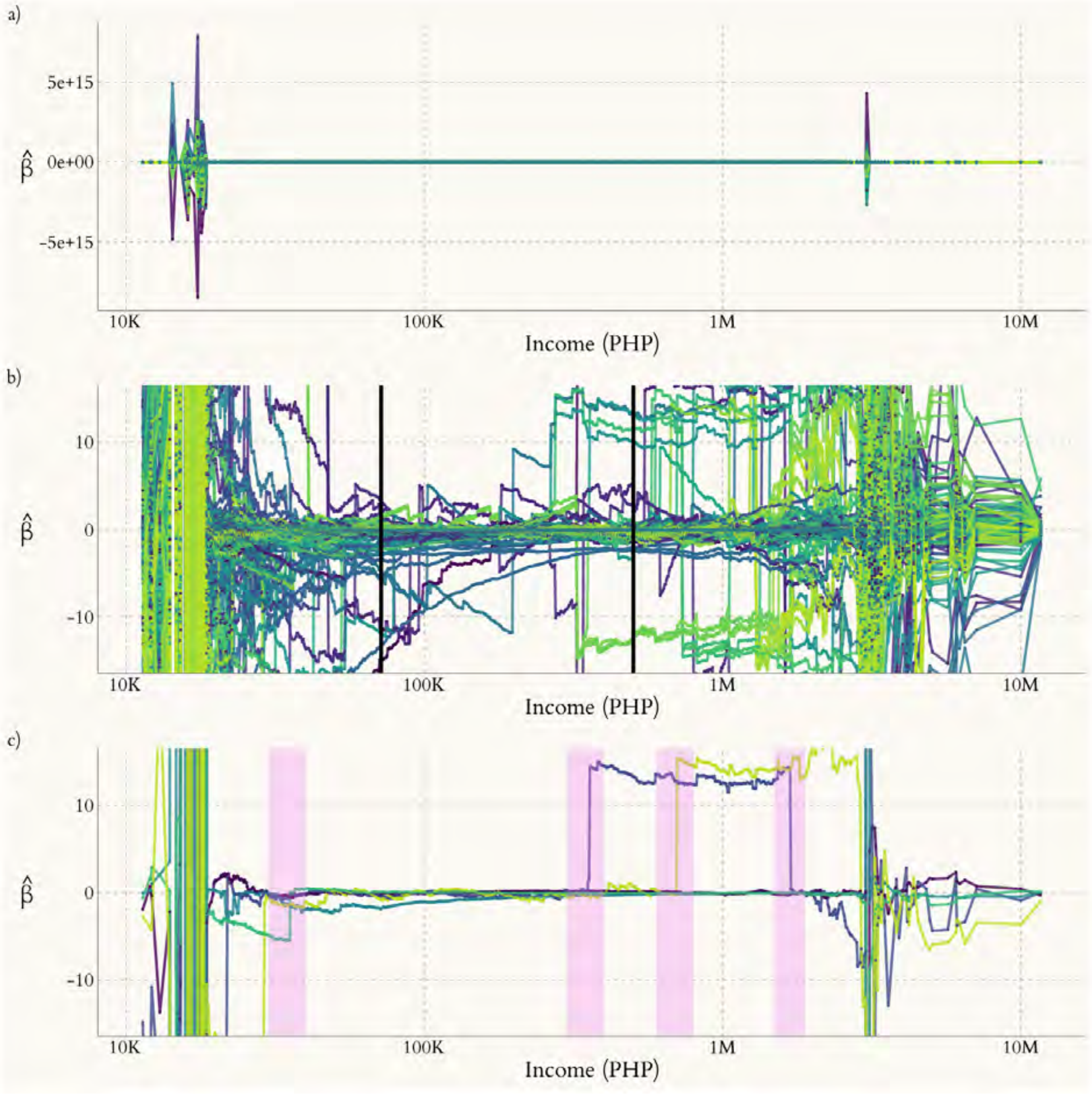


Figure 4: Poverty curves resulting from the first BS modelling approach. The plots show that for extreme boundaries, extreme coefficients are estimated. Further, one can notice extreme jumps in the poverty curves corresponding to categorical or count variables.

The poverty curves of the first model are depicted in Figure 4 in three plots in distinct ways, but always with a logarithmically scaled x-axis. It should be noted that Figure 4 is not analysed in detail, as it serves as a reference point to show the problems of the first model and the first conspicuous features of BS. Plot a) shows the complete poverty curves without restricting the y-axis to highlight the extremely small and large estimated coefficients for low and high boundaries. One can see that large estimations occur mainly for extreme boundaries. If the y-axis is restricted to the range -15 to 15, the poverty curves shown in Plot b) result. This plot is chaotic due to many poverty curves, making it hard to interpret single poverty curves, which makes it impossible to interpret the change in relative importance. It can be noted, however, that the poverty curves fluctuate strongly, with the boundaries around an income of 100,000 PHP experiencing the least fluctuation, which may be related to the fact that many households have an income around this value. This is highlighted in the figure with the black vertical lines, as 80 per cent of the households have an income in between. Plot c) depicts the same information as Plot b), but only five arbitrarily selected poverty curves are displayed. When comparing the various poverty curves, one can notice a fluctuation in all of them. Jumps are particularly high in poverty curves belonging to categorical or counting variables. The poverty curves belonging to the dummy variables roof salvaged, tenure status rent free 2, and the counting variable, stove, are examples of this observation. The jumps are highlighted in pink.

With these first insights, there will be the tasks of

1. reducing the number of poverty curves,
2. determining the source of the high jumps,
3. handling the extreme estimations of the coefficients for high and low boundaries.

When these tasks are completed, the impacts of model and data changes can be evaluated. The next chapter deals with variable selection.

## 7.2 Variable Selection with Logistic Lasso Regression

The first issue that is tackled is the reduction of poverty curves which is done by using logistic lasso regression with the purpose of variable selection. It has been explained in Chapter 4.2 that the tuning parameter $\lambda \geq 0$ controls how many coefficients are set to zero. As just a certain number of variables should remain in the model, a corresponding $\lambda$ is sought that ensures this. I think it makes sense to set all but ten coefficients to zero. One can justify this by the fact that with ten variables, which results in ten or more poverty curves depending on the number of dummy variables, there is still a certain clarity in the analysis of the poverty curves. For clarification, it can be more than ten poverty curves in the case that dummy variables are among the selected variables. That is because it has been decided that if at least one category of a categorical variable has been determined to be important by lasso logistic regression, the whole variable is added to the model having as a consequence that the for each category, a poverty curve is drawn. However, selecting ten variables is still an arbitrary decision.

Variable selection is now made in two steps. In the first step, BS is used with logistic lasso regression models instead of logistic regression models and the boundaries are placed at the mean income and the income quantiles. The tuning parameters $\lambda_{10}(z_t)$, $t = 1, \dots, 4$ depend on the boundaries and are chosen to end up with ten non-zero coefficients. The variables corresponding to the coefficients are then seen as the relevant variables.

In the second step, the intersection of the variable sets resulting from the four logistic lasso regressions is formed. One could argue that this reveals the variables that play an important role in poverty measurement on the whole income range. They did it at least for the four boundaries that cover a large income range without being too close to extremely low or high incomes. Doing BS with only four boundaries might seem arbitrary again. Still, the risk of getting a completely different set of variables increases with every additional boundary. This would result in fewer variables in the intersection of the boundary-dependent variable sets. Accordingly, I believe that employing just four boundaries, as long as they are distributed over the whole income range, is acceptable.

There are two reasons that logistic lasso regression is not primarily used in BS, but logistic regression is instead. Firstly, different coefficients are set to zero. This makes drawing the poverty curves hardly possible. The second reason is that the run time of BS increases drastically. For just four boundaries, the run took nearly six minutes. For these two reasons, using logistic lasso regression at each boundary in BS is not practicable.

The variables food expenditure, refrigerator, washing machine and mobile are in the intersection of the sets of variables. These variables are now used in the BS models for comparing the placement of boundaries and the used binary prediction models, as well as for calculating the poverty indices later on. But before, the jumps noticed in poverty curves of categorical or count variables are analysed. Since the intersection of variables does not contain any categorical variables, BS is performed with the variables from the union of the sets of variables. The analysis of the jumps in the poverty curves now follows.

## 7.3 Categorical and Count Variables

In the first BS model, there were jumps in the poverty curves of categorical and count variables observable. The term "jump" is used in the context of the poverty curves description. The term describes situations where the estimated coefficients of neighbouring boundaries are extremely different, resulting in a noticeable jump in the poverty curve. The BS union model with 15 variables is used in this chapter. Formula 38 was used again for boundary placement, so the boundaries are the same as in the first BS modelling approach. Next, the jumps in the poverty curves of the categorical variables are analysed.

### 7.3.1 Categorical Variables

Out of all the poverty and SE curves resulting from the union BS model, just those resulting from the variables electricity and walls are drawn. The **SE curves** concept is similar to the poverty curves concept. The only difference is that the estimated coefficients $\hat{\beta}_j$ related **Standard Errors (SE)** are drawn with the associated boundary $z_t$ as points in the coordinate system. To note here, the SE values are derived from the asymptotic covariance matrix $\widehat{\text{Cov}(\hat{\boldsymbol{\beta}})}$ (Fahrmeir et al., 2009, p. 134). Figure 5 shows the poverty and SE curve of the dummy variable electricity 1. The poverty and SE curves are drawn for each dummy variable resulting from the categorical variable. Therefore, there is a single poverty curve and a single SE curve for the dummy variable electricity 1. One can see three jumps in both plots of Figure 5. The major jump is between the boundaries 3,282,764 and 3,305,694. The first is at the smallest boundary, and the third is at the highest boundary. The major jump will be discussed since the others are at boundaries that lead to extremely imbalanced data sets, with just a single or a few observations being in the set of poor or non-poor. One can attribute the reason for the other jumps to the

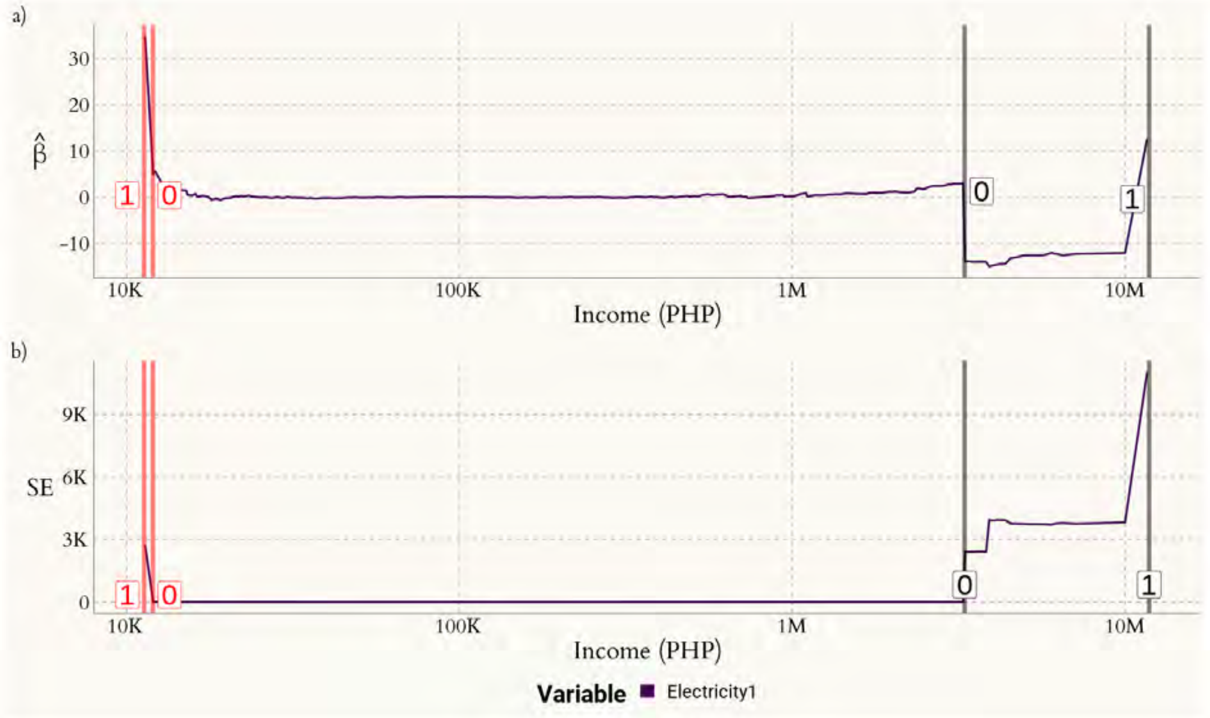


Figure 5: Poverty a) and SE b) curve corresponding to the variable that indicates if a household uses electricity and is therefore in category one. In both curves, the jumps are located at the maximum and minimum income measurement of electricity category zero households.

same problem.

Discussing the major jump, one can see that among all the households with no electricity, the household with the maximum income has an income of 3,294,322, right at the jump of the poverty curve. A vertical grey line with the label zero has been added to highlight this jump. Adding the minimum incomes of households from each category and the maximum income of households from electricity category one to the plots shows that the jumps occur at these values.

Because it is now known where the jumps occur, it can be explained what they are due to. First, however, the poverty curve is interpreted. Looking at the poverty curve, one can see that it remains at approximately zero between the maximal minimum income, 11,988, and minimal maximum income, 3,294,322, of the two existing electricity categories, zero and one. The poverty curve rises slightly when it approaches one of the two mentioned values. It is now required to explain what it means if a coefficient is close to zero. Therefore, Equation 28 is used, which shows the multiplicative impact on the odds in the logistic regressions. This equation reveals that if a coefficient $\beta_j$, $j = 0, \ldots, k$ is close to zero, then the odds of something will not change much when the corresponding value $x_{ij}$ is changed by one unit. The odds increase if $\beta_j$ is positive and $x_{ij}$ increases by one unit. Transferring this insight shows that for boundaries between 11,988 and 3,294,322, the odds of being poor change just slightly when a household belongs to either electricity category one or zero. At the few boundaries where one can observe slightly positive estimated coefficients, being in the electricity category one even has a positive multiplicative effect on the odds of being poor. This means that a household with electricity has higher odds of being poor than a household without electricity, if all other variables remain unchanged. I would argue that this makes no sense because households that are connected to the electricity grid are less disadvantaged, but no justification for this inconsistency is sought.

The estimated coefficients for boundaries above the critical value at 3,294,322 are approximately -14 and then slightly decrease to -12 until the second-to-last boundary. This means that having electricity in the house is a good indicator of not being poor since it results in a strong decrease in the odds of being poor. So, for these boundaries, being in the electricity category one is a relatively good indicator of being non-poor.

It is striking that electricity is suddenly an important indicator, although it was not important for lower boundaries. This is because the data sets $D_t$ above this boundary are **quasi-complete separable**. Quasi-complete separation is defined as follows (Lu, 2016, p. 1).

**Definition 7.2** (Quasi-Complete Separation)**.** Quasi-complete separation occurs when the dependent variable separates an independent variable or a combination of several independent variables to a certain degree.

It is added that this means that at least one category of the dependent variable has zero frequency for at least one category of an independent variable. It can occur due to continuous variables as well. This is fulfilled for boundaries above the critical value of 3,294,322 since one can assign households without electricity automatically to the poor. In the case of households that have electricity, one can not possibly say whether they are poor or not. There are methods to indicate if a data set is quasi-complete separated, and one is implemented in the R package "detectseparation" (Kosmidis et al., 2022). It is not explained here how it works, but the package description contains all the technical information. This method can check if the previous statement is true by testing the data sets resulting from the four boundaries 3,255,355, 3,294,322, 3,320,300 and 3,389,330 on quasi-complete separation. For these few examples, the test shows that quasi-complete separation occurs above the critical value because the first and second data sets are not quasi-separable, but the third and fourth are. One has to note here that the remaining variables have been removed from the data set as some lead to quasi-complete separated data sets too.

The issue of quasi-complete separation is that it can lead to some estimated regression coefficients being infinite, and adding independent variables to the data set does not remove quasi-complete separation (Mansournia et al., 2018, web appendix 2). The estimated coefficients of electricity 1 are not infinite at any boundary. Instead, it can be seen in Plot a) that they range from about -14 to 30. The estimated coefficients are probably finite due to the reason described by Mansournia et al. (2018, p. 865), that the algorithm that maximises the log-likelihood function stops when regression coefficients become numerically too large for the software to handle.

According to Mansournia et al. (2018, p. 866), another consequence of quasi-complete separation are very large SE values and, thus, the SE curve shown in Plot b) should have a jump at 3,294,322. One can indeed observe this jump, and I would therefore argue that it can be confirmed with certainty that the jumps are due to quasi-complete separation.

Some of the ways to address quasi-complete separation are to remove the variables causing the quasi-complete separation, to use the method “Exact logistic regression” as it can provide finite “median unbiased” estimates or to perform “Firth penalisation” which solves the quasi-complete separation by penalising the log-likelihood function from Equation 23 to reduce the bias of maximum likelihood estimators in generalised linear models (GLM) (Mansournia et al., 2018, p. 868). Another solution given by Allison (2008, p. 8) is to do nothing and to leave all variables in the model because the estimated coefficients, SE values and test statistics for the variables that do not cause quasi-complete separation are still valid maximum likelihood estimates. One could report the coefficient that caused the quasi-complete separation as positive or negative infinite. Furthermore, the interpretation changes if the problem variable is a dummy variable because one has to interpret the estimated coefficients for the remaining variables as the estimated coefficients of the model based on the subsample of individuals that fall into the dummy variable associated category. This means for boundaries above 3,294,322, that the estimated coefficients, which are based on the whole data set, are identical to those calculated based on the data containing just the households with electricity.

For a second example of the consequences of quasi-complete separation, the poverty curves of the variable walls are displayed in Figure 6 Plot a), and the related SE curves in Plot b). Just as there were jumps in the poverty curve of electricity 1, one can see jumps in all poverty curves corresponding to the categorical variable walls. Between 40,764 and 341,586, all poverty curves are close to zero. In Plot b), one can notice that the SE values are very large except in the mentioned interval. Again the largest value of the minimum income and the smallest value of the maximum income within the categories of the variable walls bind this interval. So the large SE values indicate quasi-complete separation.

Because of the results shown in this chapter, I am confident that the issue of quasi-complete separation is present for all categorical variables. It affects many boundaries, as the highest incomes of the households of the same category differ greatly. The same applies to the lowest incomes but not to such an extent.

One could also notice jumps in the poverty curves of the count variables. Therefore, the following chapter looks at count variables in detail.

### 7.3.2 Count Variables

Variables that have positive integer values and are used to record the quantities are called count variables in this thesis. An example of this is the variable washing machines in the data set. Although the count variables are standardised, which means that the positive integer values are transformed to continuous numbers, the variable remains discrete as there is a one-to-one connection to the set of the natural numbers possible. Therefore,

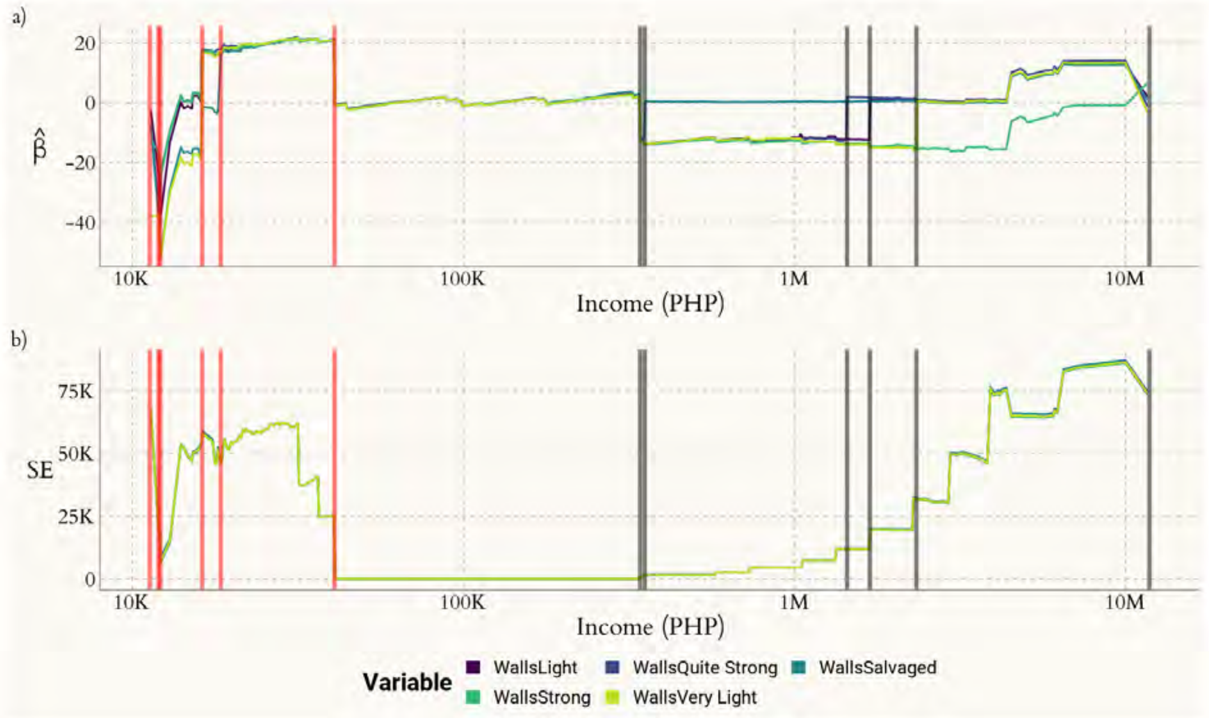


Figure 6: Poverty a) and SE b) curves corresponding to the variable that indicates which material the house walls are made of. Again, in the poverty and SE curves, one can see jumps at the categories' maximum and minimum incomes.

they will still be called count variables. One could expect that the discrete nature of the count variables is the reason for the jumps in some poverty curves of the first BS model from Chapter 7.1. The suspicion arises because there were jumps in the poverty curves of categorical variables for the same reason.

Figure 7 a) displays the poverty curve and b) the corresponding SE curve of the variable washing machine. In Plot a), one can see a few jumps in the poverty curve. One is in line with the minimum income among those who own a single washing machine at 36,569. One can also notice this jump in Plot b), as the SE of the estimated coefficients boosts from slightly above zero at the boundary 36,592.08 to approximately 300 at the boundary below. Except for the jump at the last boundary that will be ignored again, and the jumps at boundaries below 36,569, one can not notice more jumps. The other poverty curves of count variables, displayed in Figure 25 in the appendix, show the same pattern since the major jump occurs at the minimum income of the households that own the second least amount of some good.

One can explain this behaviour with quasi-complete separation again. Concentrating on the variable washing machine, for each boundary below the critical value of 36,569, one can perfectly predict if a household is non-poor. This is because if the household owns one or more washing machines, it is non-poor. If a household does not own one, it is uncertain whether it is poor or not. Above the critical value, this is not possible anymore. It is still possible to say that a household with two or more washing machines is non-poor,

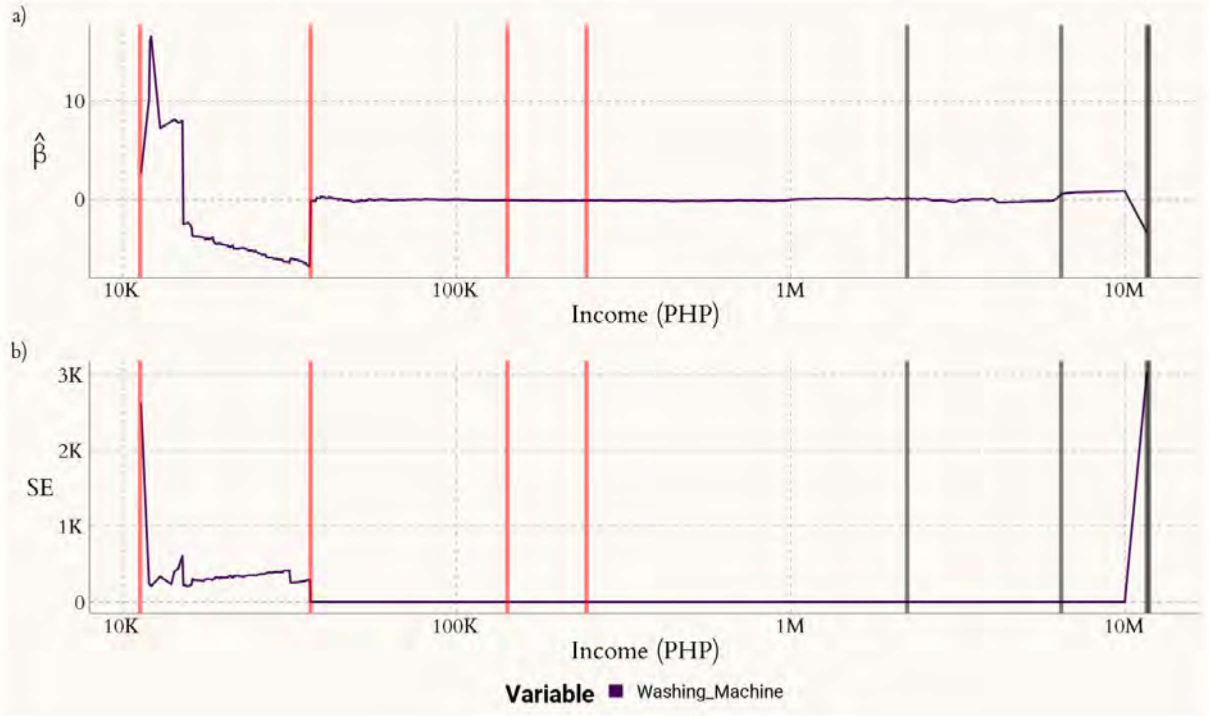


Figure 7: Poverty a) and SE b) curve corresponding to the variable washing machine. One can notice a striking jump in the poverty curve below approximately 50,000.

but for owning less than two, it is uncertain. For a quasi-complete separation, it would be necessary that not owning a washing machine leads to being poor so that there is only uncertainty if a household owns exactly one washing machine.

Quasi-complete separation can occur for each count variable individually as long the boundaries are below a variable-specific critical value. The critical values are given in Table 5 in the appendix for the variable washing machine and other count Variables. Testing these critical values with a few boundaries using the R function from package "detectseparation" (Kosmidis et al., 2022) reveals again that quasi-complete separation occurs below them. Also, the SE curve of the variable washing machine confirms that quasi-complete separation is present for this variable because one can observe large SE values for boundaries below the variable's critical value.

In Plot a), below the critical value, 36,569, one can observe further jumps in the poverty curve. My explanation for them is that they are due to the quasi-complete separation of other variables. However, since it is already clear that data sets for limits below the critical value are quasi-completely separated, no search is made for the trigger of the jumps.

As seen in Plot a), there is no jump in the poverty curve at high boundaries, excluding the jump at the last boundary. This would mean there is no quasi-complete separation for high boundaries. This may be because the quantity of some goods does not increase gradually with a household's income. To understand this speculation, what is meant by

| Number Washing Machines | Min | Max |
|---|---|---|
| 0 | 11,285 | 11,639,365 |
| 1 | 36,569 | 11,815,988 |
| 2 | 141,910 | 6,452,314 |
| 3 | 245,025 | 2,231,340 |

Table 2: Minimum and maximum income per group, defined by the number of washing machines each household owns. It can be seen that the number of washing machines increases with the minimum income in each group but the same can not be said about the maximum income.

a gradual increase must be explained.

For lower incomes, it might be the case that households can not afford the goods they need, but they can with an increasing income. However, this only applies to a certain income level at which a household is saturated with this good so that no more is bought, even though the household has enough income to buy more. The result is that grouping the households according to the number of products and examining the minimum income for each group reveals a connection between minimum income and the number of goods. The minimum incomes are ordered according to the amounts of goods, or differently said, the quantity of goods gradually increases with the income. As far as I am concerned, there is no relationship between the maximum incomes and the number of goods, and I would go even further and claim that the order is random.

At least for the variable washing machine, this can be observed. Table 2 shows the order of the minimum and maximum incomes of the groups. The minimum incomes of the groups are in the same order as the number of washing machines, and the opposite is true for the maximum incomes.

Heading back to the claim that quasi-complete separation is unlikely to occur for high boundaries, it is now clear that the number of goods does not always increase with the maximum income. As a result, the households with the highest quantity of goods just by chance have the highest income. Therefore one can not expect that quasi-complete separation occurs for high boundaries, as there is no allocation where households with the second largest quantity of goods or less are allocated to the poor, and those with the largest quantity of goods to the non-poor or poor.

Because it is now clear that quasi-complete separation causes the jumps in the poverty curves of the count variables, it is continued with the third task of handling extremely small or large estimated coefficients at low and high boundaries.

## 7.4 Boundary Limits

In upcoming analyses, extreme estimated coefficients could strongly impact calculated metrics. One way of handling them is to remove them entirely. Since the extreme estimated coefficients occur at different boundaries depending on the binary prediction model and the boundaries, one must define an **upper limit** $U$ and a **lower limit** $L$ to ensure that models are always compared on the same boundary range.

One already noticed in the previous chapter that some jumps in the poverty curves are attributed to quasi-complete separation. Based on that, quasi-complete affected boundaries can be removed, which already removes the fluctuation of the poverty curves at low boundaries. Since quasi-complete separation only affects lower boundaries in data sets that do not contain categorical variables, an additional approach is required to remove extreme estimated coefficients at high boundaries.

The other approach is to remove boundaries that lead to extremely **imbalanced data sets**. Having an imbalanced data set means that the class distributions in the dependent variable are highly imbalanced (Ling and Sheng, 2010, p. 167) or, differently said, skewed. In the following, the skew of a data set is calculated as the ratio of households in the minority class to the total number of households. Data imbalance occurs in BS by design, and the lower or higher the boundaries are placed, the more imbalanced the data sets get. In the most extreme case, one assigns just a single household to the poor or non-poor.

A consequence of data imbalance is that the SE values increase for an increasing class imbalance skew. One could conclude this from King and Zeng (2001, p. 141) as they show that observations belonging to the minority class are more statistically informative than observations from the majority class. They show it with the asymptotic covariance matrix

$$\widehat{\operatorname{Cov}(\hat{\boldsymbol{\beta}})} = \frac{1}{\sum_{i=1}^{n} \pi_i(1-\pi_i)\boldsymbol{x}_i^T\boldsymbol{x}_i}, \tag{39}$$

that is used to calculate the SE values. They further claim that most imbalanced data applications result in small estimates of $\pi_i = \mathbb{P}(y_i = 1 \mid \boldsymbol{x}_i)$ for all observations. As the factor $\pi_i(1-\pi_i)$ is in the denominator of Formula 39, it becomes clear that the denominator decreases the more $\pi_i$ deviates from 0.5, and as a result, the SE values increase.

Later, one can see that the SE curves rise with increasing class imbalance skew. To avoid large SE values, low and high boundaries must be removed. It has been decided that there will be a maximal data imbalance skew of 0.01. One has to note that this is, again, an arbitrary value.

To account for both issues, data imbalance and quasi-complete separation, which are most likely the cause of extreme estimated coefficients, some metrics are just calculated for estimated coefficients between the boundaries ranging from $L = 36,619$ to $U = 1,287,000$. The lower limit is set due to quasi-complete separation. It is calculated on the basis that only the intersection variables, derived in the variable selection with logistic lasso regression chapter, Chapter 7.2, are included in the data set. One calculates the upper limit based on the data imbalance skew of 0.01. More information on the placement of the limit is supplied in the following two chapters. However, if the derivation is not of interest to the reader, it is possible to proceed to Chapter 7.5, where the **Basic BS model** is discussed.

### 7.4.1 Quasi-Complete Separation Limits

One could observe in Chapter 7.3 that in BS, quasi-complete separation occurs for high or low boundaries that are, depending on the variable, above or below one or more critical values. As it was decided in Chapter 7.2, just the variables food expenditure, mobile, refrigerator and washing machine are used in the BS models. For this reason, it must only be dealt with quasi-complete separation resulting from count variables, which means that boundaries that are simultaneously below the critical values of the variables mobile, refrigerator and washing machine have to be removed. Table 5 from the previous chapter contains the critical values of the three mentioned count variables. Since the variable Computer, with the largest critical value, 44,313, is not included in the following BS models, the critical value, 36,569, corresponding to the variable washing machine, is the largest relevant critical value. Since the lower boundary has to be above the critical value, the lowest boundary is placed at $L = 36,619$, which corresponds to the income of the household with the next highest income.

### 7.4.2 Imbalance Limits

There are 41,544 households in the data set, and the task is to find a lower and an upper limit that the data imbalance skew is above 0.01 for each data set $D_t$, $= 1, \ldots, T$. The search is for two limits because, in one case, the poor are the minority class and, in the other, the non-poor. To get a data imbalance skew of less than 0.01, in the minority class have to be at least

$$\lceil 0.01 \cdot 41,544 \rceil = 416, \tag{40}$$

households. The 417th lowest income is 34,128; therefore, the lower limit will be placed at this value. The upper limit will be placed at 1,285,400 at the income of the 41,128th household.

The lower limit, resulting from data set imbalance, is below the lower limit resulting from quasi-complete separation. Therefore, the final, more restrictive limits are $L = 36,619$ and $U = 1,285,400$.

## 7.5 Basic BS Model

The BS model that is discussed in this chapter will be referred to as the Basic BS model. Compared to the first BS model, simply the independent variables food expenditure, refrigerator, washing machine and mobile are included. This means there is not any categorical variable used in this model. These variables are used as they have proven in the variable selection with the logistic lasso regression chapter that they are relevant for BS poverty measurement, as they were within the ten most relevant variables of each of the four used boundaries. Compared to the first BS model, one estimates the coefficients in the Basic BS model for each unique boundary that satisfies Condition 32.

Figure 8 shows the poverty curves resulting from the Basic BS model. The resulting poverty curves on the whole income range can be seen in Plot a). The purple area in the plots is in this and the following chapters, bounded by $U$ and $L$, which were derived in the

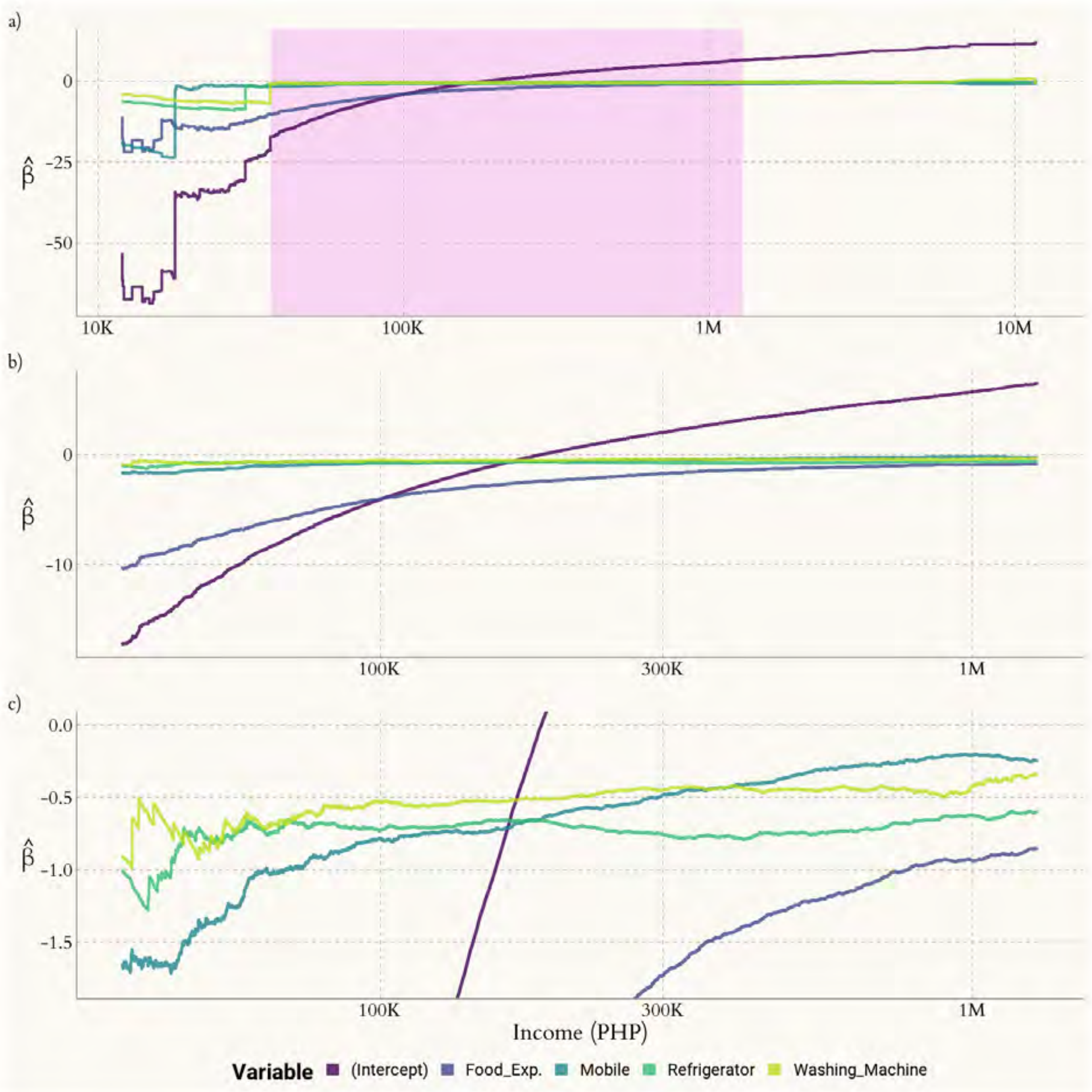


Figure 8: The poverty curves of the Basic BS model on the total income range are shown in Plot a), and between the limits $L$ and $U$ in Plot b). The poverty curves of the variables washing machine, refrigerator and mobile are shown in detail in Plot c). The poverty curves are relatively smooth between the limits, but there is still some observable fluctuation. The poverty curves of the variables shown in detail in Plot c) are in between the limits close to zero, indicating low relative importance.

previous chapter. Although the estimated coefficients and their corresponding P-values and SE values in the purple region are of main relevance when comparing models, the plots that are shown on the whole income range are used to demonstrate the form of the curves outside of the limits.

In Plot a), it can be seen that the poverty curves heavily fluctuate below the lower limit $L$, and further, in the poverty curves corresponding to the count variables, at least one jump is observable. The food expenditure poverty curve starts to fluctuate for very low incomes wildly. The trigger of the fluctuation will not be further analysed since it just occurs for extremely low boundaries. At the critical values of the count variable, jumps can also be seen in the intercept poverty curve. For low incomes, the intercept poverty curve fluctuates wildly similarly to the food expenditure poverty curve.

Generally, the intercept poverty curve is far below the poverty curves of the slopes for low incomes. Between the limits $L$ and $U$, the intercept poverty curve rises for increasing boundaries and crosses the x-axis at 186,840. The poverty curves of the slopes also rise, but not as fast, and they do not cross the x-axis. Above the upper limit $U$, the poverty curves do not start to fluctuate and continue the trend seen between the limits.

The poverty curves between $L$ and $U$ are shown in Plot b). Although the poverty curves seemed to be smooth in Figure a), it is now notable that there is some fluctuation but mostly for boundaries close to $L$. Interpreting the poverty curves, one can see that the relative importance is higher for boundaries close to $L$. One should note that relative importance has to be understood like it was explained in Chapter 6.2, and also that the estimated coefficients are less, but the absolute value is important for interpreting relative importance. For the variable food expenditure, the relative importance is very high at boundaries close to the lower limit and then decreases with increasing boundaries. Compared to the other variables, the variable food expenditure has by far the greatest effect on the odds of being poor as the estimated coefficients, for the boundaries close to $L$, are approximately -10. In contrast, the other estimated coefficients have values smaller than -2. As the relative importance of food expenditure decreases with increasing limits, this variable has near the upper limit a similar relative importance as the other variables.

I would argue that the intercept poverty curve increases with increasing boundaries due to the ratio of the poor to the non-poor. For low boundaries, for example, households are in the data set mainly assigned to the set of the non-poor, which means that if the boundary corresponding model predicts whether a household with average characteristics is poor, this household is likely non-poor. As the proportion of non-poor households decreases, households with average characteristics are more likely to be classified as poor.

The poverty curves of the variables mobile, refrigerator and washing machine are close together, making it hard to interpret them. For this reason, Plot c) shows the poverty curves in detail. One can see fluctuation at low boundaries near $L$. Although the general trend is that the relative importance decreases with increasing boundaries, the relative importance of the variable refrigerator increases before it decreases again. One can observe a strong decrease in the relative importance of the variable washing machine. For increasing boundaries then, the relative importance increases gradually and later slowly decreases again.

These changes in relative importance are probably due to differences in the quantity of a good between the poor and the non-poor. Taking the variable refrigerator as an example. Relative importance is high at low boundaries since the poor cannot afford a single refrigerator, but most non-poor households can afford at least one refrigerator.

When the boundary shifts upwards, it might be the case that the relative importance decreases because more and more households with a single refrigerator are assigned to the poor. Therefore, at these boundaries, this variable might not be that good for dividing into poor and non-poor anymore. But after further shifting the boundary upwards, there might come a point where some households can afford a second refrigerator or a freezer. Thus, one might expect that the relative importance of this variable increases again, as it is suitable for dividing the households into those that can afford one or less and those that can afford two or more.

The third notable thing is that the variable mobile, compared to the variables refrigerator and washing machine, has a larger relative importance for boundaries close to $L$ but for boundaries close to $U$ it has lesser. A reason for this could be that mobile phones are relatively cheap compared to the other two goods, and so households with a small income are able to afford at least one. Just the really income poor can not afford one, and therefore, this variable might have higher relative importance at low boundaries. However, owning one or more mobile phones, which are comparatively cheap goods, is no longer a luxury if income exceeds a certain value, as most households can afford it. From this, one could conclude that at higher boundaries, the variable mobile is no longer suitable for dividing households into poor and non-poor.

In Figure 9, the estimated coefficients associated P-values and SE values for every boundary are drawn as P-value and SE curves. Figure a) shows the P-value curves, and Figure b)-c) the SE curves on different scales.

The jumps in the P-value and SE curves are at the same boundaries as in the poverty curves. The intercept P-value curve has one outstanding peak at approximately 200,000. Heading back to Plot a) in Figure 8, one can see that the intercept poverty curve crosses the x-axis at this point, which means that the intercept estimations are close to zero. Because for calculating the P-value in a GLM, the null hypothesis $H_0 : \beta_j = 0$ is tested against the alternative hypothesis with the Wald statistic (Liu, 2016, pp. 75-76)

$$Z = \frac{\hat{\beta}_j}{\sqrt{\widehat{\text{Cov}}(\hat{\beta}_j)}}, \tag{41}$$

it is clear that this peak does not occur due to an increase in the SE values. Instead, it is attributable to the estimated coefficient being nearly zero resulting in the test statistic being close to zero. Note here that $\widehat{\text{Cov}(\hat{\beta}_j)}$ is the SE of coefficient $\beta_j$, that has been calculated with Equation 39.

This peak is not noticeable in the intercept SE curve in Figure 9 Plot b). But one can see in this plot that all SE curves are close to zero above the lower limit $L$ and seemingly also above $U$. The large SE values below the lower limit $L$ occur due to quasi-complete separation again. To see how close the SE curves are to the x-axis, one has to look at Plot c) with the scaled y-axis. One can see that the SE curves smoothly decrease and increase between the limits. While the curves are close to $L$ at approximately 0.5, they are nearly zero at $U$. The SE curves decrease and increase due to data imbalance. The

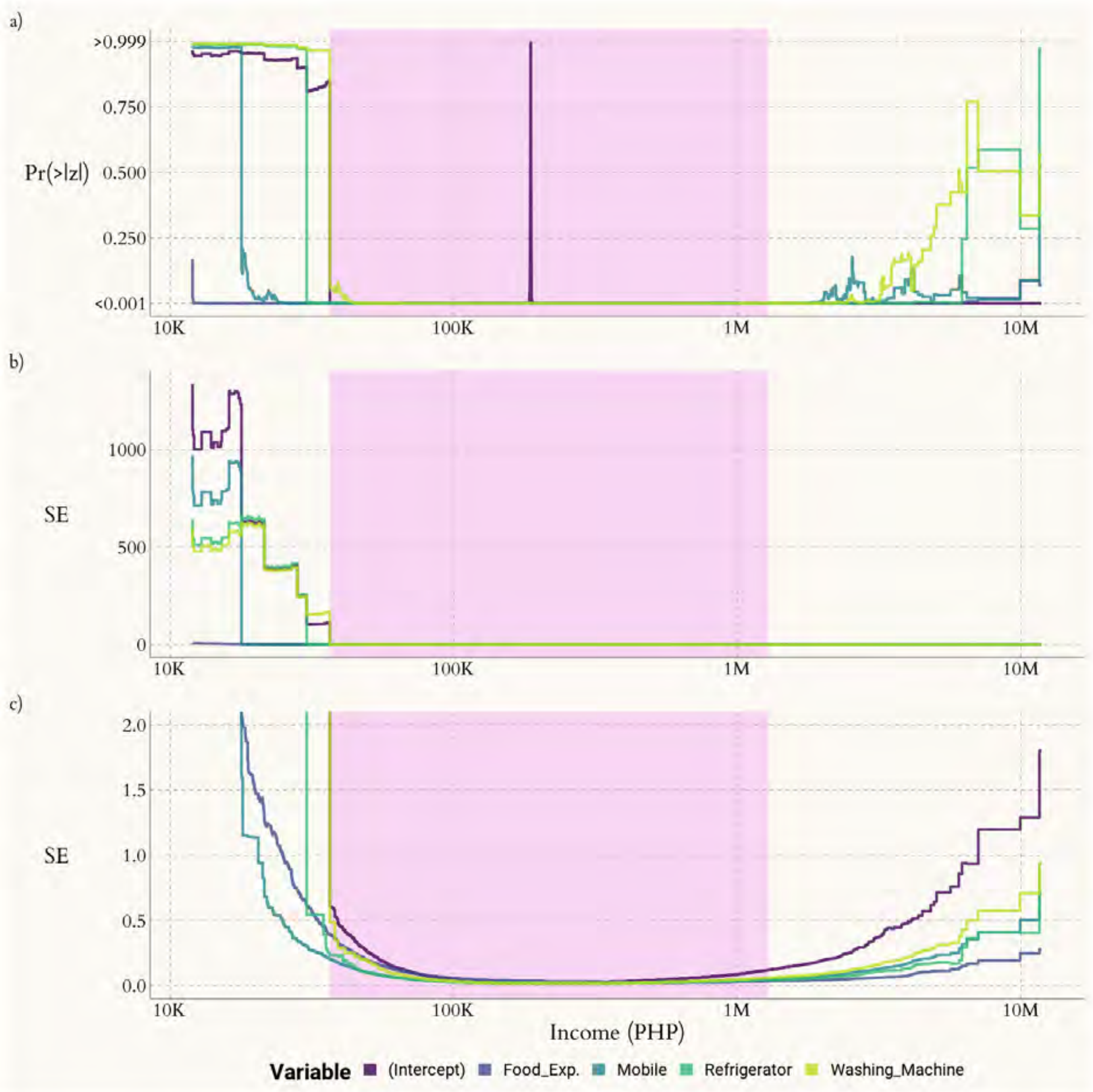


Figure 9: P-value and SE curves on the whole income range. Between the lower limit $L$ and the upper limit $U$ are all the P-value and SE curves close to zero. One can notice a peak in the intercept P-value curve due to the estimated intercept being close to zero at this boundary.

details of how the SE values are affected by data imbalance were given in Chapter 7.4 in the context of the asymptotic covariance matrix. The SE curves keep rising above the upper limit $U$, with no noticeable jumps.

For further analysis, it has been decided to analyse just the SE curves and no longer the P-value curves. I argue that it is unnecessary to look at the P-value curves as well since one derives the P-values from the SE values and the estimated coefficients, and both are in each model discussed in detail.

The fluctuation between $L$ and $U$ is now further analysed. Therefore, the estimated coefficients are not analysed, but the absolute difference of estimated coefficients of neighbouring boundaries are. The absolute difference is defined as

$$\mathbf{diff_t} = \left| \hat{\boldsymbol{\beta}}(z_t) - \hat{\boldsymbol{\beta}}(z_{t-1}) \right| , \tag{42}$$

with $t = 2, \ldots, T$. $T$ is again the total amount of boundaries and $\mathbf{diff_1} = \mathbf{0}$ applies.

To see how the fluctuation changes, the boundaries are divided into **strata** and the mean, median and standard deviation (SD) values of the absolute difference are calculated in each stratum. For readability reasons, the mean absolute difference of differences within the same stratum is referred to as the stratum mean. The same applies to the median and the SD of the absolute differences within the same stratum.

Dividing boundaries into strata means a domain is divided into sub-intervals (Saltelli et al., 2008, p. 59). Here, one uses 100 sub-intervals of different widths since the sub-intervals will be placed dependent on the boundaries. This means one places the sub-intervals so that each stratum has the same amount of boundaries. Note that if the number of boundaries is not a multiple of 100, the number of boundaries inside the strata can differ by up to one. The domain in the following analysis is the interval ranging from the lower limit $L$ to the upper limit $U$.

Before it is looked at the results, one has to mention that the mean, median and SD values of the strata are not assumed to be zero, even if there is no fluctuation in the poverty curves. This is the case since the poverty curves are increasing and decreasing. Now, if there is a steeper increase or decrease of the estimated coefficients inside a stratum, one can expect the resulting differences to be larger. This problem is partly exacerbated by the fact that one places the sub-intervals dependent on the boundaries, and the boundaries have different distances depending on the boundary placement. There will not be used a method to account for this problem.

Figure 10 displays the strata mean, median and SD values. For increasing strata, the estimated coefficient-dependent mean, median and SD values of the strata decrease to nearly zero and then increase again. The absolute differences between neighbouring estimated coefficients are, in most cases, quite small, as the highest mean value across all strata and coefficients is 0.018. One can see this value in the lowest strata for the intercept. The mean values are generally higher in the lower strata than in the middle and upper strata. One can observe the same for the SD values, which confirms the previous observation in the analysis of the poverty curves that there is more fluctuation at lower boundaries. Also, it indicates that there are probably a few larger differences. The strata median values are overall below the strata mean values. This indicates that there are a few larger differences $\mathbf{diff_t}$, which increase the mean value. One can conclude that there is indeed more fluctuation at lower boundaries, which is partly attributed to a few differences $\mathbf{diff_t}$. One could also observe that there might be some fluctuation at higher boundaries. But this is not sure because the larger differences could also be due to the increasing or decreasing trends of the poverty curves.

To clarify the two possible reasons for large strata mean values, Figure 11 shows the

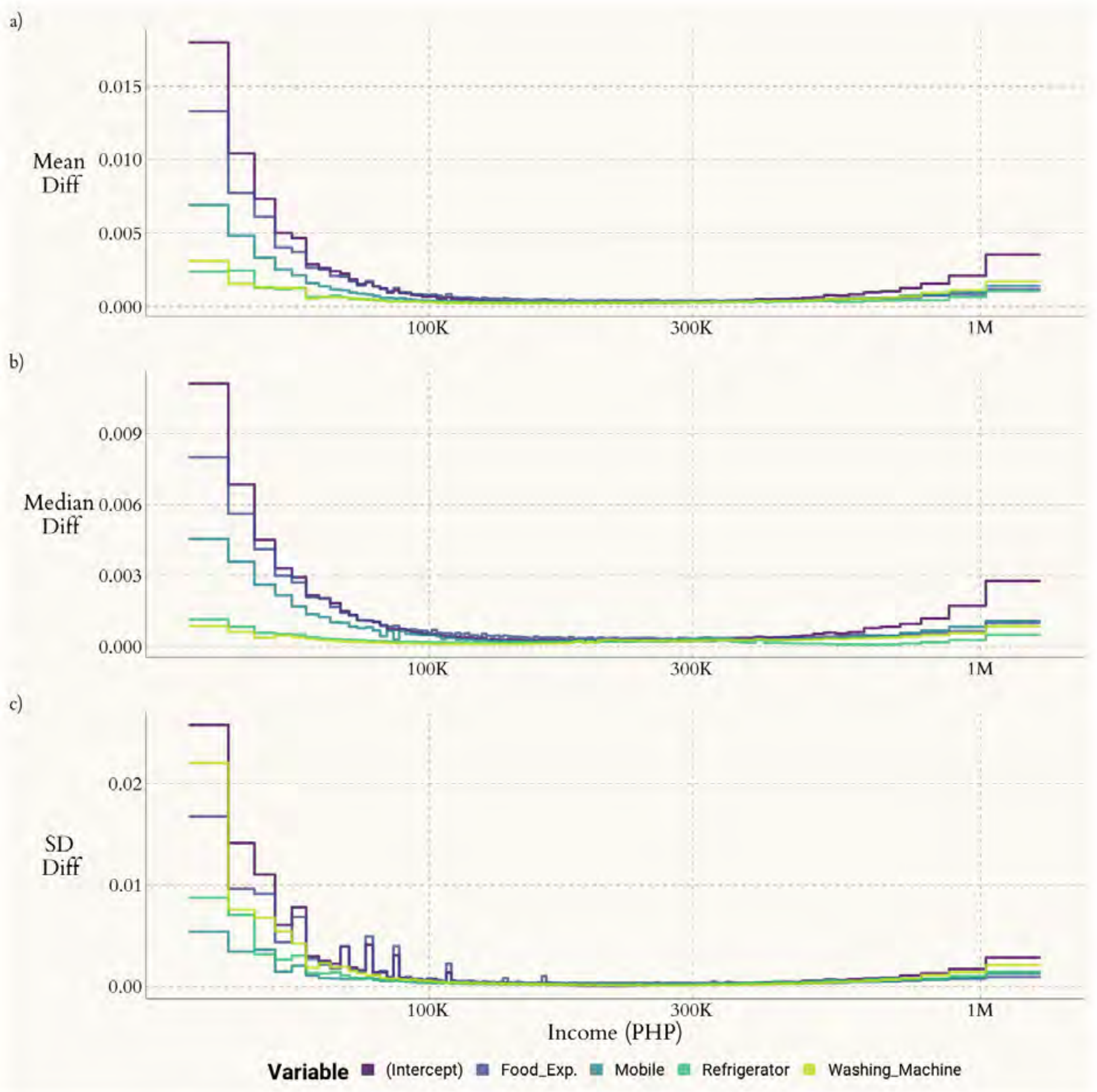


Figure 10: Figure that displays each stratum's mean, median and SD values. One can observe larger absolute differences in neighbouring estimated coefficients in the lower strata. For central strata are, the three statistics very small, which shows that there is nearly no fluctuation in the poverty curves at mediocre boundaries.

poverty curves of the intercept and washing machine for boundaries between 38,000 and 39,300. This is just a section of the first stratum, which is so small that one can still recognise individual boundaries, shown as vertical dashed lines. The first reason for large strata mean values is that the strata mean values are not supposed to be zero. One can see this in the intercept-related poverty curve in Figure 11 as the intercept estimations increase with increasing boundaries. The difference between the estimation of the first and last boundary in this section is 1.20, so if the mean value of absolute differences was calculated for this section, one could already say that it is at least $\frac{1.2}{77} = 0.015$ as the absolute

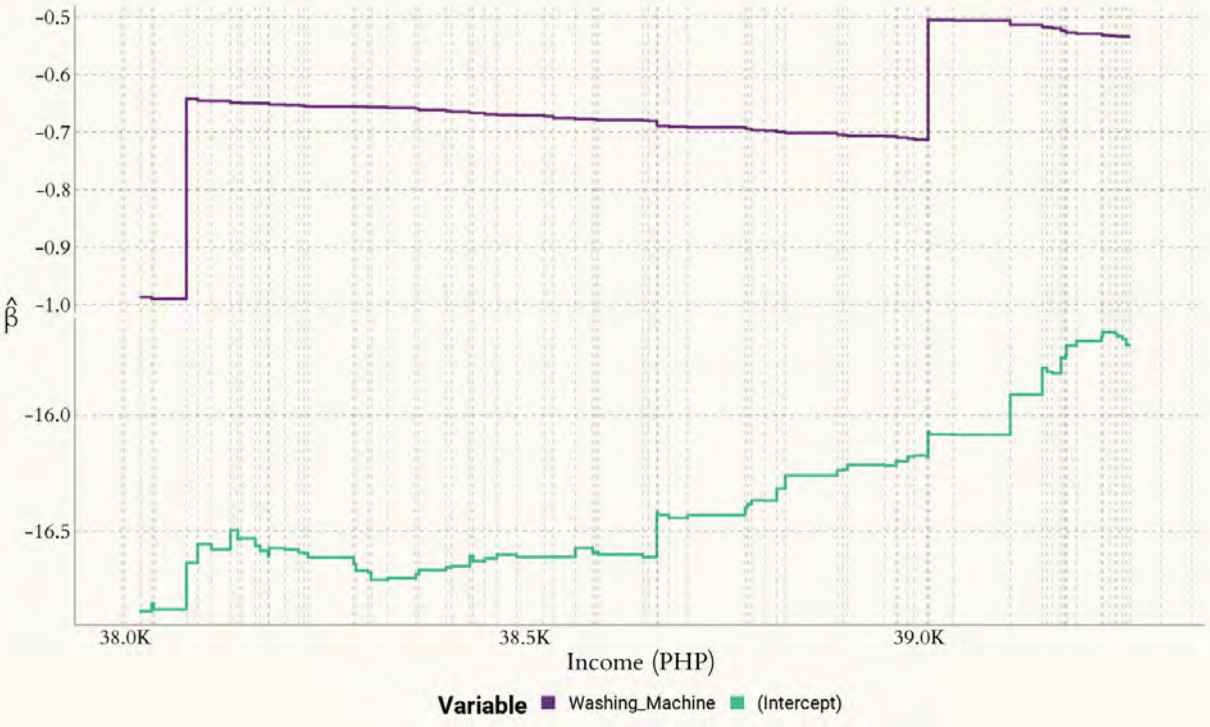


Figure 11: Plots that illustrate the two possible reasons for larger strata mean values. The first reason for the higher mean values of the strata shown in the upper poverty curve is a trend in the poverty curve. The second reason, illustrated with the lower poverty curve, is that the stratum has a few larger differences $\mathbf{diff_t}$.

differences are calculated for the intercept estimations of 78 boundaries. Comparing this value with the previously observed stratified mean of 0.018 of the intercept, it becomes clear that the large mean value is probably attributed to the trend of the intercept poverty curve.

The second reason for large absolute mean differences is actual fluctuation. One can observe this from the washing machine poverty curve as the poverty curve is relatively smooth except for two large jumps at the boundaries 38,075 and 39,011.

To conclude this chapter, one could see that below the lower limit, there were larger jumps in the poverty curves that are related to count variables. Between $L$ and $U$, there is far less notable fluctuation for all poverty curves. Still, there is some fluctuation in the poverty curves for boundaries close to $L$. But the fluctuation decreases rapidly until it drops to zero. When the absolute differences $\mathbf{diff_t}$ are used to analyse fluctuation, one must remember that larger strata mean values are not automatically attributed to fluctuation.

The next chapter continues with the first assumption change. The effects of the number and placement of boundaries on the poverty curves are investigated.

## 7.6 Boundary Placement

The Basic BS model, where one uses every possible boundary, can be time-consuming for large data sets. Reducing the number of boundaries reduces the run time on the one hand but, on the other hand, could result in a loss of information.

In the following, the poverty curves of five models, which differ in the amount and placement of the boundaries, are compared in order to analyse the effects. One should note that stair-step plots are used for the Basic BS model and line plots for the other models. That is because each data split is carried out in the Basic BS model. In the others, it is not. Using a stair-step plot highlights that due to the Condition 32, all $z_t, z_{t'} \in [y_{(l)}, y_{(l+1)}]$ with the ranked incomes $y_{(l)}, y_{(l+1)} \in y$ and $l \in 1, \ldots, n-1$ have the same data sets $D_t, D_{t'}$ and therefore the same coefficients $\boldsymbol{\beta}(z_t), \boldsymbol{\beta}(z_{t'})$ as a consequence. If not every data split is carried out, the pairs $(z_t, \hat{\beta}_j(z_t))$ are drawn as points in the coordinate system and afterwards connected with lines. The lines are drawn for clarity, and one should not use this to derive the estimated coefficients $\boldsymbol{\beta}(z)$ for a boundary that has not been used, which means $z \notin \{z_t \mid t \in 1, \ldots, T\}$. Logistic regression has been used as the binary prediction model in the compared BS models. Also, the variables food expenditure, mobile, refrigerator and washing machine are the relevant independent variables again. The compared BS models are named Basic, **Permilles**, **Sequence**, **Log998** and **Log498**. The Basic BS model is known from the previous chapter. In the model Permilles BS, one places the boundaries at the 1,000 sample quantiles, called Permilles (Walker and Lev, 1969, p. 60), of the observed incomes. $T = 4,998$ boundaries are used in the Sequence BS model and placed evenly between the minimum and maximum observed income. For the boundaries applies

$$z_t = t \cdot \frac{\max\limits_{1 \le i \le n}(Income_i) - \min\limits_{1 \le i \le n}(Income_i)}{T+1} + \min_{1 \le i \le n}(Income_i), \tag{43}$$

where $t = 1, \ldots, 4998$. The boundaries of the last BS models, Log998 and Log498, are placed with Formula 38. In the Log998 BS model is $T = 998$, and in the Log498 is $T = 498$. The boundaries of the Log998 BS model are equal to the boundaries that have been used in the first BS model.

Figure 12 displays the poverty curves on the whole income range in the first row, the poverty curves between $L$ and $U$ in the second row and the SE curves between $L$ and $U$ in the third row. The titles above the first row's Plots a)-e) identify to which model the plots in the same column belong to.

Looking at the figures in the first row, it is obvious that the parts of the poverty curves below $L$ and above $U$ differ. While the poverty curves of the BS models Log998 and Log498 look like copies of those of the Basic BS model, the poverty curves of the other two models look less detailed. This is because there are just a few boundaries above and below the limits for which the coefficients are estimated. Below $L$, one can observe the second least detailed poverty curves for the Sequence BS model, where the coefficients are estimated for ten boundaries. At the same time, the parts of its poverty curves above $U$ are very detailed because the coefficients are estimated for 298 boundaries.

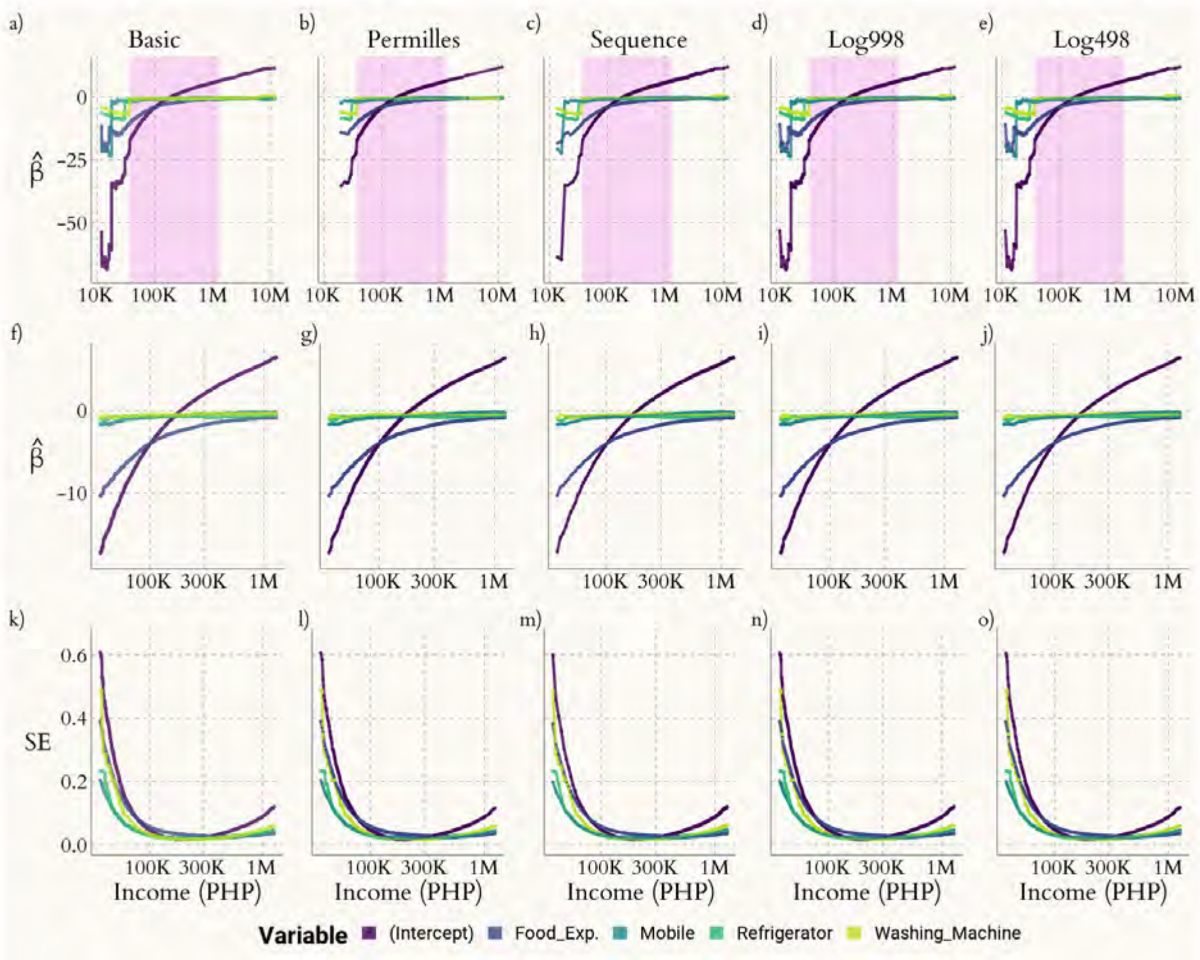


Figure 12: The poverty and SE curves are between the limits similar. Above and below the limits are the models Log998 and Log498, comparable with the Basic BS model. In the other models, Permilles and Sequence, the curves are less detailed.

This distribution of the boundaries results from the income distribution in the data. Since the boundaries in the Sequence BS model are set at even intervals, it occurs for lower boundaries that hundreds of households have an income between two boundaries. With higher boundaries, the opposite can occur, namely that multiple boundaries lie between the incomes of two households. Consequently, there are not many boundaries for lower incomes but lots for high incomes.

If multiple boundaries are between two successive income observations, they all lead to identical poor and non-poor data splits and, therefore, to identical coefficients $\boldsymbol{\beta}(z_t)$. This means that there is a reduced amount of effective splits. This is considered when the coefficients are estimated, which means that in the R code, the coefficients are only estimated for one boundary to avoid redundant calculations. This reduces the number of boundaries, and in the case of the Sequence BS model, only 833 effective boundaries are left of the originally defined 4,998 boundaries.

| | Low | Middle | High | Sum | Initial Boundaries |
|---|---|---|---|---|---|
| Basic | 525 | 37,727 | 417 | 38,669 | 38,669 |
| Permilles | 12 | 977 | 11 | 1,000 | 1,000 |
| Sequence | 10 | 525 | 298 | 833 | 4,998 |
| Log998 | 115 | 511 | 139 | 765 | 998 |
| Log498 | 66 | 255 | 87 | 408 | 498 |

Table 3: The Permilles BS model is outside the limits, not that detailed. The Sequence BS model is not detailed below $L$ but above $U$. The BS models Log498 and Log998 are satisfactorily detailed on the whole domain. Additionally, both models result in many effective boundaries, while the Sequence BS model does not.

Table 3 shows the number of the effective boundaries of the five models below $L$, above $U$ and between the limits. Further, the numbers of total effective boundaries and the numbers of initial boundaries are given. Previously mentioned numbers about the number of boundaries have been taken from this table.

Above the upper limit $U$, the Sequence BS model has, with a number of 298, the second most effective splits. However, this model has only 833 effective boundaries in total, meaning about 35 per cent of all boundaries are above the upper limit. Adding that the model originally had 4,998 boundaries, one can conclude from these numbers that the Sequence BS model might be good for analysing the change of relative importance for higher incomes but is not that detailed between the limits.

In the Permilles BS model, the number of effective boundaries equals the number of initial boundaries by design. Nearly 98 per cent of the effective boundaries are between the limits. Therefore, this model may be more suitable than the Sequence BS model for analysing the relative importance between the limits. But a disadvantage of this model is that it is not as detailed outside the limits. The BS models Log998 and Log498 seem to be somewhere between the capabilities of the other two already discussed models. Approximately 80 per cent of the initial number of boundaries result in effective boundaries, and approximately 65 per cent are between $L$ and $U$. Like Figure 12 already showed, both models can visually return the same information about relative importance on the whole income domain as the Basic BS model.

Returning to Figure 12 to analyse the poverty and SE curves between the limits, no major differences are noted by comparing Plots f)-j). It is only noticeable for the Permilles and Sequence BS model that there are not that many boundaries close to $L$. Since there are no major spikes at lower boundaries, the decreased boundary density is, in my view, acceptable. The same applies to the SE curves in Plots k)-o). It can be concluded that each model is adequately detailed between the limits since the poverty and SE curves look like copies of those of the Basic BS models.

After visually analysing and justifying the shape of the poverty curves, the fluctuation is now analysed. Again, just the fluctuation between $L$ and $U$ is looked at and the absolute differences, $\mathbf{diff}_t$, $t = 2, \ldots, T$, between the estimated coefficient of neighbouring boundaries, $\hat{\boldsymbol{\beta}}(z_{t-1})$ and $\hat{\boldsymbol{\beta}}(z_t)$, are analysed. To see how the absolute differences change over the income domain, the income domain is divided into strata, and each stratum's mean, median and SD value is calculated. One has to note here that the strata depend on effective boundaries and not the initial boundaries. If there are more boundaries per stratum, the statistics are calculated with more absolute differences. Having more absolute differences per stratum will very likely reduce the SD values. Another consequence of using effective boundaries is that the sub-intervals that define the strata are getting wider at higher boundaries where the income density is lower. This increases the mean and median values if there are trends in the poverty curves. This relationship was explained before with the help of Figure 11. This issue also arises if the boundaries are not evenly distributed due to the boundary placement of the model. This makes it hard to compare BS models with different boundaries directly, but the strata are still analysed as this analysis gives insights into the individual BS models.

Figure 13 displays the results. The rows of the figure correspond to the calculated strata statistics and the columns to the BS models. The first observation is that the strata mean, median and SD values are very different between the BS models. While the mean values in all the Basic BS models strata are very low, the maximum strata mean value of the Sequence BS model is larger than 0.75. The second observation is that the values of the strata statistics from the Log998 and Log498 BS models fluctuate and do not decrease and increase gradually, as one can notice for the other BS models. Additionally, it seems like there are more strata. This is because the exponential function was used to derive the boundaries. The x-axis in the plot is logarithmically scaled, and thus the strata seem to be evenly spread. This is not the case for the other models, and for this reason, depending on the distribution of the income and the boundaries, the strata are more or less spread.

Discussing the strata mean values, one can see that the strata mean values of the Basic and Permilles BS model are similar, as the strata mean values quickly decrease at low strata and increases at large strata. The ranking of the mean values for low strata is different since the strata mean values of the variable mobile are smaller than the strata mean values of the other variables in the Permilles BS model. The strata SD and median values ranking is also different, and the SD is generally higher since each stratum has fewer boundaries.

Discussing the other three models, one can see that the ranking of strata mean values corresponds to the ranking of the Permilles BS model. Further, the strata mean values decrease slower and do not rise at the end. That the stratum mean values do not increase indicates that at higher incomes, there are a lot of boundaries which result in data splits that differ just by a few households. The Sequence BS model is highly affected since the strata mean values vanish for increasing strata. The slow decrease of the strata mean values in the Sequence BS model is due to the same reason. This model does not attribute to the fact that income is not evenly distributed. Therefore, low boundaries result in data sets that differ by many households, which results in estimated coefficients $\boldsymbol{\beta}(z_t)$ that differ by a lot. As the boundaries increase, the data sets resulting from neighbouring boundaries become more similar, so the estimated coefficients no longer differ greatly.

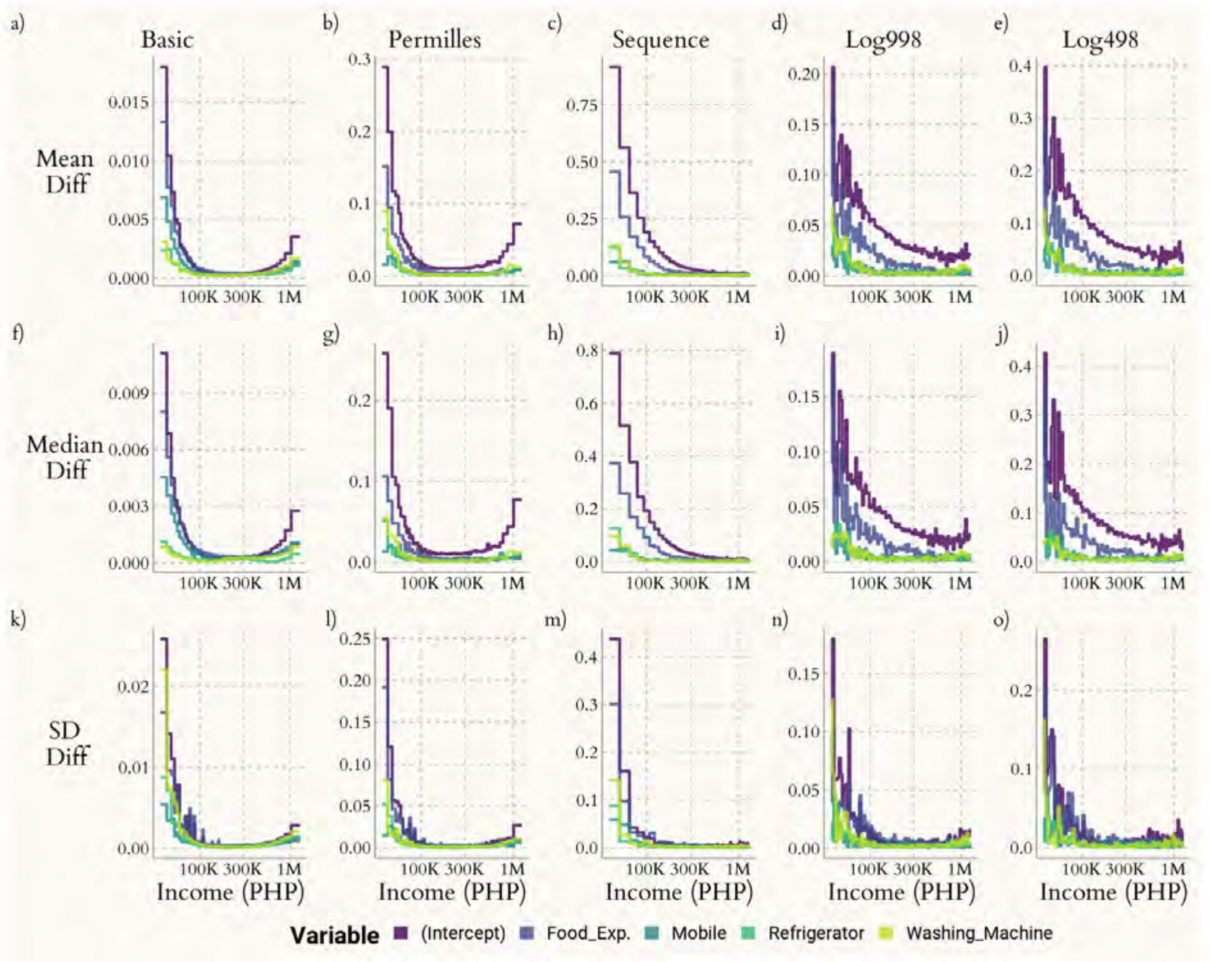


Figure 13: The strata statistics are very dependent on the boundaries, making comparing the models hard. One can still notice that the BS models Log998 and Log498 can overcome the issue of not evenly distributed boundaries. One can see this as the strata mean, median and SD values fluctuate, and additional due to the constant and not increasing values at higher strata.

This has smaller differences $\mathbf{diff_t}$ as a consequence.

The strata mean and median values in the Log998 and Log498 BS models stay equal or increase slightly for higher strata. These models seem to compensate for the unequal income distribution and find strata widths that lead to similar strata mean and median values in higher strata. This can also be noticed due to the fluctuation of the strata mean, median and SD values, which indicates the existence of a random error in every stratum, which can hardly be noticed if the strata statistics increase or decrease too fast. A difference between the Log998 and Log498 BS model is that the strata mean and median values are approximately twice as large in Log498, and the SD values are larger too, which one could explain by the reduced number of boundaries in each stratum.

As a consequence of the analysis in this chapter, the boundaries from the Log998

BS model will be used in future analyses. This can be justified by the fact that this model appears to compensate for the unequal income distribution and is more detailed than the Log498 BS model. Another reason is that the poverty curves are on the whole domain comparable with the poverty curves of the Basic BS model. Finally, one can better examine the strata statistics as the x-axis in all plots is logarithmically scaled, resulting in equal-spaced boundaries that do not lie within each other.

The following chapter presents the BS models **RobGLM** and two models resulting from Bootstrap. The goal is to find out if an alternative binary prediction model can reduce the fluctuation between the limits.

## 7.7 Comparison of Binary Prediction Models

One can add complexity to BS by changing the binary prediction model one fits on each data set $D_t$, $t = 1, \dots, T$, but this impacts the fluctuation of the poverty curves. To see if one can reduce the fluctuation of the poverty curves between $L$ and $U$ with another model, the same variables and boundaries are used in the following models so that the same conditions apply. Before the poverty curves are compared, Bootstrap and the Robust generalised linear model (GLM) are briefly introduced.

### 7.7.1 Bootstrap

Bootstrap can be used to estimate the precision of statistics by repeatedly drawing randomly with replacement from a data set (Liew, 2008, p. 2). This can be used in BS too. Randomly drawing with replacement means, in the context of BS, that from each data set $D_t$, $B \in \mathbb{N}$ Bootstrap samples $D_t^b$, where $b = 1, \dots, B$, are randomly drawn with replacement. The resulting Bootstrap samples $D_t^b$ have the same amount of observation but contain duplicate observations.

Bootstrapping in BS is used to calculate the mean estimated coefficient $\hat{\boldsymbol{\beta}}^{mean}(z_t)$ and the median estimated coefficient $\hat{\boldsymbol{\beta}}^{median}(z_t)$ from the estimated coefficients of logistic regressions that are performed on each Bootstrap sample. There are 100 Bootstrap samples $D_t^b$ used to obtain the estimated coefficients $\hat{\boldsymbol{\beta}}^{b}(z_t)$. Since $\hat{\boldsymbol{\beta}}^{mean}(z_t)$ and $\hat{\boldsymbol{\beta}}^{median}(z_t)$ is calculated, there are two different Bootstrap BS models, **MeanBoot** and **MedianBoot**.

### 7.7.2 Robust GLM

According to Ronchetti (2010), the primary goal of robust statistics is the development of procedures which are still reliable and reasonably efficient under small deviations from the model when for example, the underlying distribution lies in a neighbourhood of the assumed model. If the fluctuations in the poverty curves are due to outliers, using a robust binary prediction model could be beneficial because distributional robust and outlier resistant, although conceptually distinct, are practically synonymous terms (Huber and Ronchetti, 2009, p. 4). The binary prediction model used in the following is called the Robust GLM. In this model, robust estimations are made based on quasi-likelihood. The procedure itself will not be explained. Instead, more information about Robust GLM is

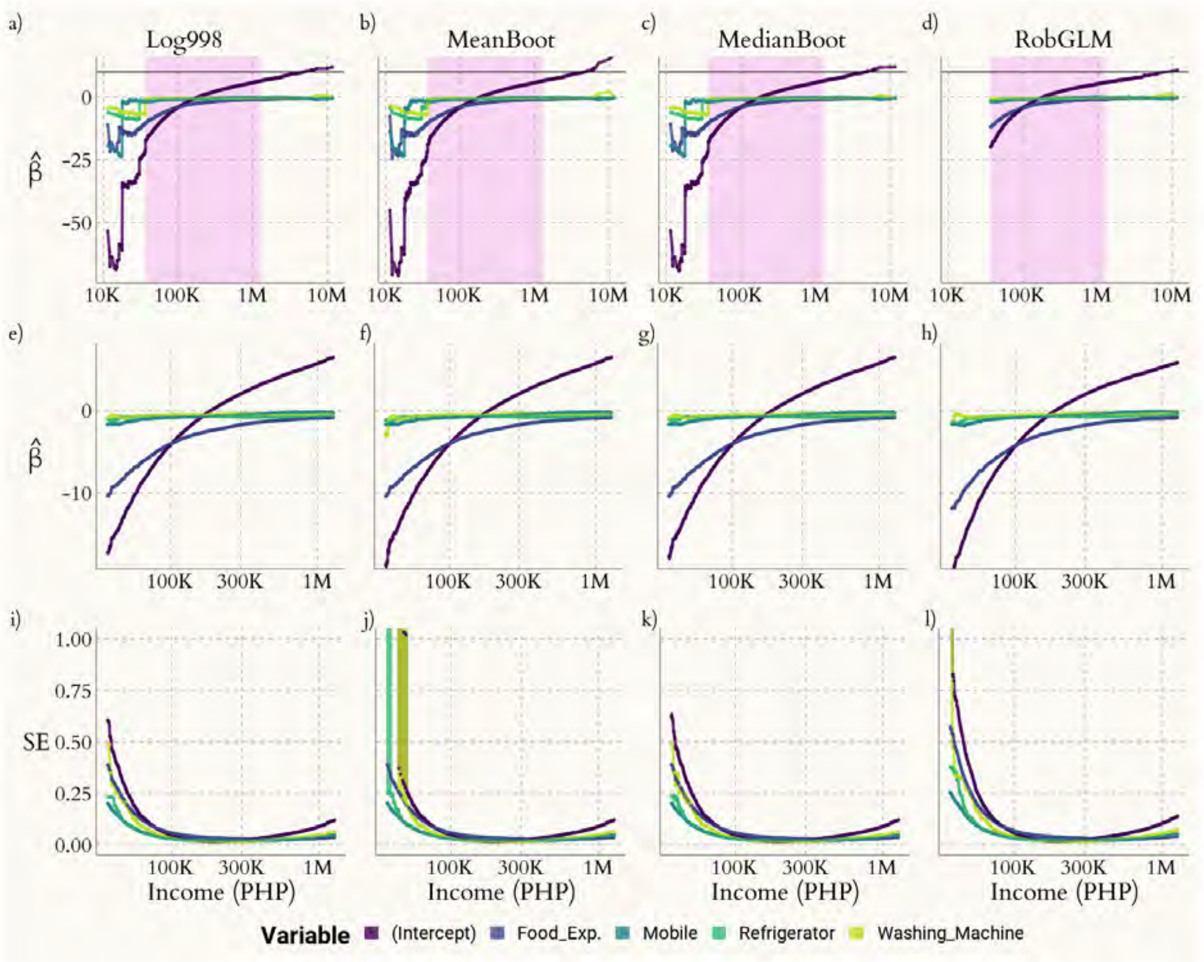


Figure 14: Below the lower limit $L$ and above the upper limit $U$, one can see some differences between the poverty curves, but the poverty curves are very similar between the limits. Close to the $L$, one can notice a heavy fluctuation in the MeanBoot BS model's SE curves due to extreme SE values in some Bootstrap samples.

given by Cantoni and Ronchetti (2001). The BS model that uses Robust GLM at each boundary will be named RobGLM.

The next chapter compares the four BS models, RobGLM, MeanBoot, MedianBoot and Log998.

### 7.7.3 Comparison

The difference to the previous chapter is that the same boundaries are used in all models. This makes the poverty curves and the strata more comparable. The first comparison of the models is a visual comparison of the poverty and SE curves. They are shown in Figure 14. Again, each column of plots corresponds to a model. Further, a horizontal line is drawn in Plots a)-d) for comparability. The poverty curves of the MedianBoot BS model are visually very similar to the ones of the Log998 BS model. The poverty curves of the MeanBoot BS model are not similar to those, as the patterns of the poverty curves change below $L$ and above $U$. For boundaries close to the maximum income, one can see that the intercept and the washing machine poverty curves are higher than the corresponding poverty curves of the Log998 BS model. For the variable washing machine, the poverty curve is even above zero, which means that owning one or more washing machines increases the odds of being poor. Below the lower limit $L$, one can observe some jumps in the poverty curves of the Log998 BS model that are now in the MeanBoot BS model a bit smoothed out.

The poverty curves of the RobGLM BS model end approximately at the lower limit since the Robust GLM only estimates the coefficients for another four boundaries below $L$ and then issues an error message. For two out of these four boundaries, the SE values, and other statistics that depend on it, are not calculated. Instead, NaN (Not a Number) is returned. In computer Science, NaNs can occur, according to Goos (1995, p. 365), from invalid operations. For the boundaries where the coefficients are not estimated at all, the Robust GLM algorithm leads to an error due to a singular matrix. A singular matrix is a square matrix that does not have a matrix inverse (Weisstein, n.d.). I would argue that the error messages and the NaNs are returned due to quasi-complete separation in the data sets. This will not be proven, but a justification for this statement is that the singularity error occurs right below $L$.

Above $L$ and close to the maximum income, one can see that the washing machine poverty curve is above zero. Also, the intercept poverty curve is for high boundaries lower than the intercept poverty curves of other models.

Plots e)-h) of Figure 14 show the poverty curves between the limits. One can only notice slight differences. For example, the MeanBoot BS model plot shows a noticeable jump in the poverty curve of the variable washing machine. Another difference is that the intercept poverty curve of the RobGLM BS model lies lower than the intercept poverty curves of the other models. However, these are just small differences.

Before it is looked at the SE curves, it has to be mentioned how these curves are derived in the Bootstrap BS models. The SE values in the MeanBoot or MedianBoot BS models for boundary $z_t$ are not calculated from the estimated coefficients $\hat{\boldsymbol{\beta}}^b(z_t)$, $b = 1, \ldots, B$ with the unbiased estimator for the SD. Instead, the mean and median of the SE values that result from estimating the coefficients of each Bootstrap sample are used. This means that the SE values of the associated coefficients of the MeanBoot BS model are calculated with

$$SE\left(\hat{\beta}_j(z_t)\right) = \frac{1}{B}\sum_{b=1}^{B} SE\left(\hat{\beta}_j^b(z_t)\right), \tag{44}$$

where $SE\left(\hat{\beta}_j^b(z_t)\right)$ the SE value of the corresponding estimated coefficient of the Bootstrap sample $D_t^b$ at boundary $z_t$ is. The SE values of the Median Boot BS model are calculated analogously with the median formula.

The SE curves are very different. Beginning with the MeanBoot BS model, the SE curves from this model look similar to those of the MedianBoot BS model and Log998 BS model. Just close to the lower limit $L$, the intercept, refrigerator and washing machine SE curves heavily fluctuate. This might be due to an unfortunate choice of Bootstrap samples having extremely large outliers as a consequence. Figure 26 in the appendix shows the maximum SE values out of the SE values derived from each Bootstrap sample. At lower boundaries, the maximum SE value is for some estimated coefficients above 200, which confirms that the large mean SE values are due to an unfortunate choice of Bootstrap samples.

The SE curves of the MedianBoot BS model in Figure 14 look like the ones of the Log998 BS model. That there are not such high SE values as seen in the MeanBoot BS model is attributed to the fact that the median is more robust to extreme values (Fahrmeir et al., 2016, p. 53). The SE curves resulting from the RobGLM model seem to be above the corresponding SE curves of other models for lower boundaries. However, the SE curves do not fluctuate as much as the SE curves of the MeanBoot BS model.

In the following, the boundaries between the limits are divided into strata again, and the strata mean, median and SD values are calculated. The results are displayed in Figure 15. The plots in the first row of the figure show the strata mean values of the different models. Looking at the strata close to the lower limit, one can see that the intercept-related strata mean values of the Log998 BS model are below those of the other models. One can notice the largest strata mean values of the intercept in the MeanBoot BS model, followed by the Median BS model. The strata mean value of the lowest strata exceeds the coordinate system since it is 0.45. The washing machine related strata mean in the same BS model is also very large, with a value of 0.50. This is conspicuous since the washing machine related strata mean values are not that large in the other BS models. Mentioning the RobGLM BS model, close to $L$, the strata mean values are also slightly larger than those of the Log998 BS model. For middle to high strata, the strata mean values are in all BS models similar, so one can not identify a model that has lower mean values.

In all BS models, the strata medians follow a similar pattern as the strata mean. In general, the strata median values are a bit below the strata mean values which again shows that there are outlier differences $\mathbf{diff_j}$. One can see something different for the strata SD values in low strata. In the lowest strata, the strata SD values of the MedianBoot and MeanBoot BS models are a bit different compared to those of the other models. Further, large strata SD values are present in the RobGLM BS model. As far as I am concerned, this contradicts the expected outcome that the outlier robust estimations of neighbouring coefficients are similar. A look at the estimated coefficients of the variable washing machine shows that the estimated coefficient jumps from -1.46 to -0.61 at one point. This is a single large absolute difference $\mathbf{diff_j}$ in this stratum which causes this large strata SD value.

But apart from the first stratum, the strata SD values in the RobGLM BS model are low compared to those of the other BS models. They are sometimes even lower than those of the Log998 BS model. The strata above the income 100,000 have comparable SD values in all models.

In summary, the least fluctuation in the poverty curves occurs in the Log998 BS

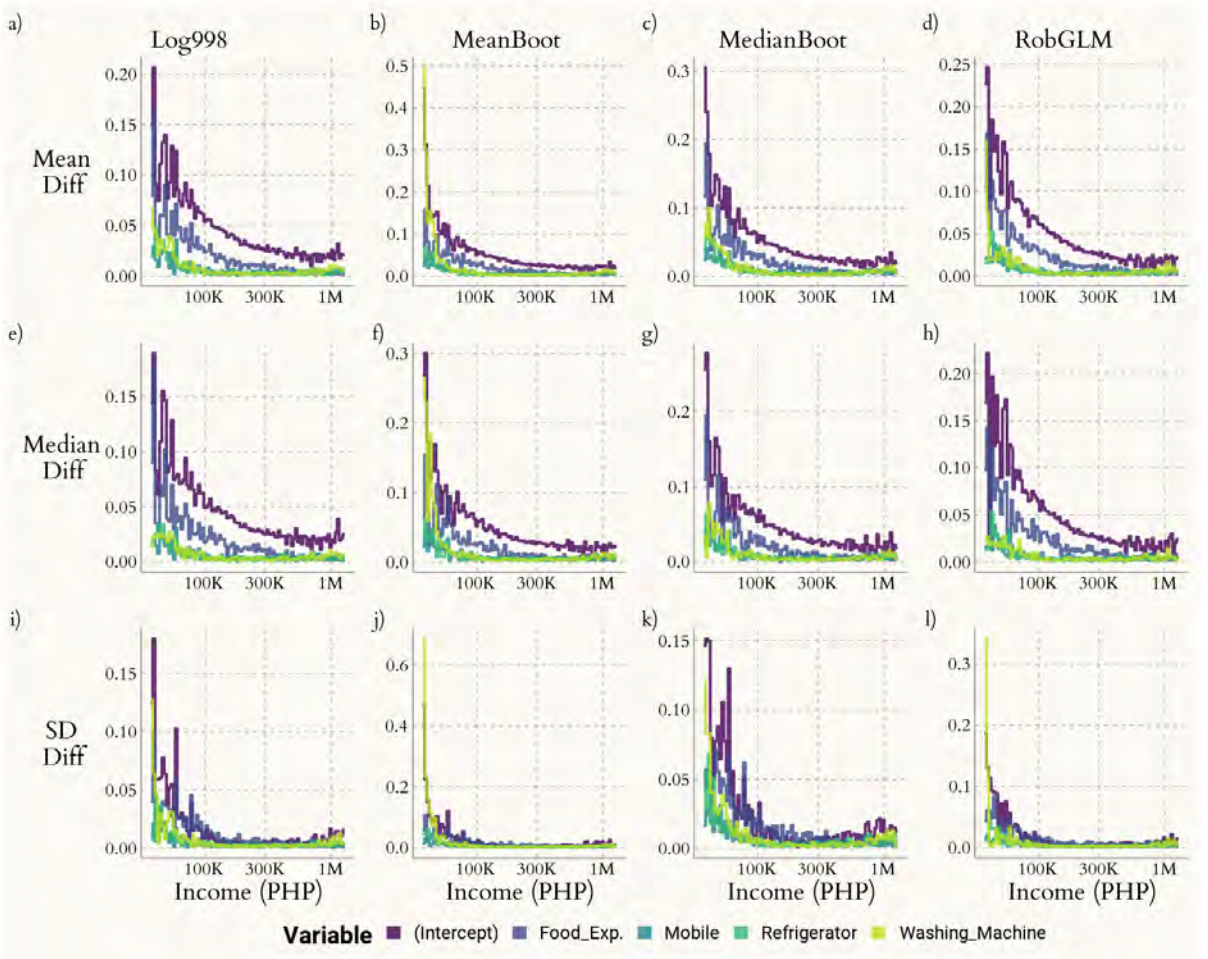


Figure 15: One can observe comparable strata mean, median and SD values. One can still notice differences for strata close to the lower limit $L$. The Log998 BS model seems to achieve the best results together with the RobGLM BS model.

model, as the strata mean values are lower or equal compared to those of the other BS models. Also, there are no larger $\mathbf{diff_j}$ differences inside the strata since the strata median values are comparable to the mean values and the strata SD values are always fairly small.

Statistics of the absolute differences $\mathbf{diff_j}$ are now calculated without stratification but still between the limits. In addition to the mean, the median and the SD value of the absolute differences, the mean and the amount of the negative differences are calculated. The results are shown in Figure 16. The bar plots in the same column correspond to the same coefficient, and bar plots in the same row correspond to the same statistic. Starting with the mean values of the absolute differences, one can see that the Log998 BS model has the smallest for every coefficient. Excluding the variable food expenditure, RobGLM has the second smallest mean values. The mean values of the two Bootstrap models are mostly slightly larger, but for the variable washing machine, the mean value of the MeanBoot BS model is almost twice as large as the second largest mean value.

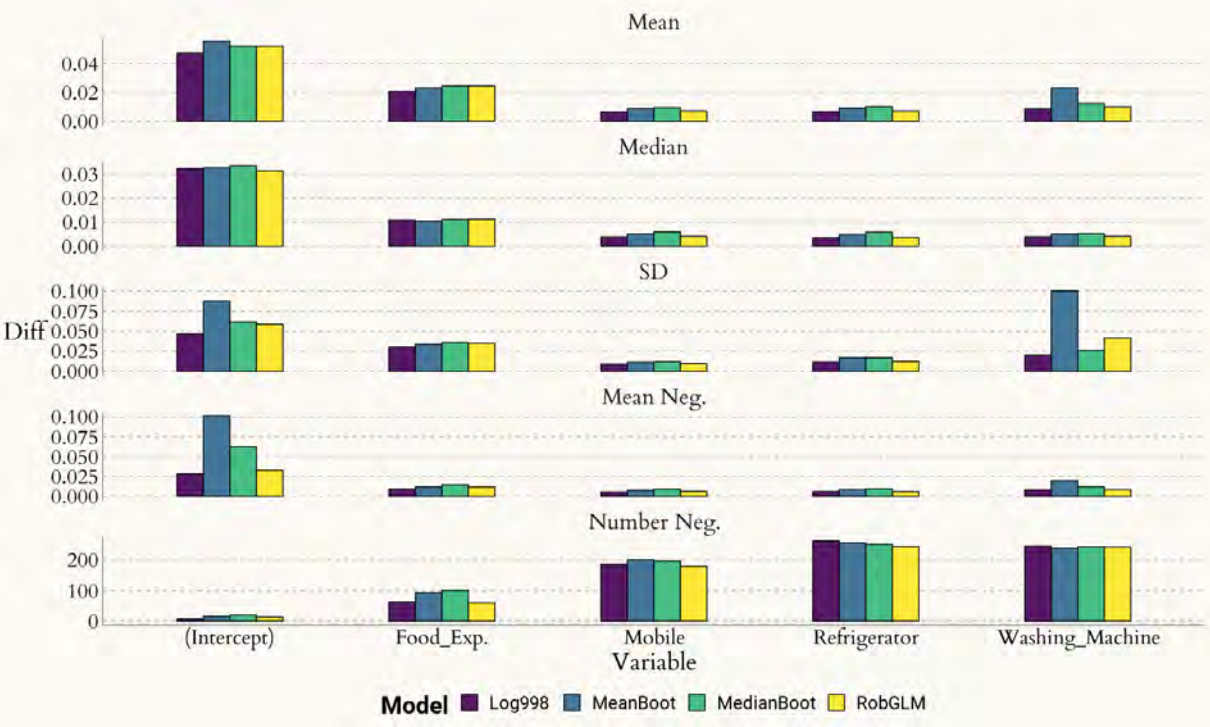


Figure 16: The Log998 BS model has the least fluctuation of the poverty curves between the limits. The small mean, median and SD values of the absolute differences indicate this. Further, it is striking that one can notice in the Log998 BS model the least negative differences of neighbouring intercept estimations. Also, the few negative differences that exist are low compared to those of other models. The statistics indicate that the RobGLM has the second lowest fluctuation in the poverty curves.

The median values of the BS models Log998 and RobGLM are overall the smallest again. This time the MedianBoot BS model has larger median values than the MeanBoot BS model. The ranking of the SD values is similar to the ranking of the absolute means, which again shows that there is less fluctuation in the poverty curves of the Log998 BS model, as there are no extreme differences $\mathbf{diff_j}$ leading to an increasing SD.

The mean value and number of negative differences are new statistics that were not used before. Note that the absolute values of the negative differences are used. Except for the intercept mean values, the mean negative differences are comparable to the mean absolute differences, so this statistic does not provide new information. But the intercept mean values of the negative differences are interesting since the intercept poverty curve is overall monotonously increasing, and therefore, negative differences should be rare and small. The Log998 BS model has the smallest mean value of the negative differences, followed by the RobGLM BS model. The mean value of the negative differences of the MedianBoot BS model is twice as large, and the value of the MeanBoot BS model is even larger. Since the number of negative differences is also the lowest in the Log998 BS model, it seems like this model is the best at modelling the intercept poverty curve in terms of fluctuation if it is assumed that the intercept poverty curve is supposed to be strictly monotonically increasing.

To conclude the results of Figure 16 and this chapter, the Log998 BS model has the least fluctuating poverty curves between the limits. This model corresponds most closely to the desire for smooth poverty curves. One could observe the second least fluctuation for the RobGLM BS model. One could notice the largest fluctuations in absolute differences in both Bootstrap BS models, whereby some large absolute differences in the MeanBoot BS model are due to outlier coefficient estimations of some Bootstrap samples. I would argue that with a larger number of Bootstrap samples $B$ one could achieve similar results as with Log998.

To reduce the fluctuation of the poverty curves to a minimum, a new binary prediction model is introduced in the following chapter.

# 8 Penalised BS

If the fluctuation of a poverty curve is measured in terms of the differences of the neighbouring estimated coefficients, one has to reduce the differences to reduce the fluctuation. The **Penalised BS model** reduces the differences by shrinking the coefficients towards the estimated coefficients of the previous boundaries. The Penalised BS model uses the self-developed binary prediction model **logistic neighbour penalisation** inspired by the lasso regression.

Instead of forcing the absolute sum of the coefficients to be less than a specific value as in Equation 29, the squared L2-norm of the difference between the $(k+1)$-dimensional parameter vector $\boldsymbol{\beta}$ and another vector $\boldsymbol{\theta}$ of the same dimension, must be less than a certain value. This is achieved again by adding a penalty term to the log-likelihood function. Analogous to the log-likelihood function of the logistic lasso regression from the Formula 30, the resulting log-likelihood function of the logistic neighbour penalisation is defined as

$$\begin{aligned}\ell(\beta_0, \dots, \beta_k) = \frac{1}{n} \sum_{i=1}^{n} & y_i \ln\left(\frac{\exp(\boldsymbol{x}_i^t \boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_i^t \boldsymbol{\beta})}\right) \\ & + (1 - y_i) \ln\left(1 - \frac{\exp(\boldsymbol{x}_i^t \boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_i^t \boldsymbol{\beta})}\right) \\ & - \lambda \|\boldsymbol{\beta} - \boldsymbol{\theta}\|_2^2,\end{aligned} \tag{45}$$

where $\lambda \geq 0$ is still the tuning parameter that controls the amount of shrinkage. Note here that the intercept $\beta_0$ is penalised too. The coefficients are then estimated as it has already been explained in Chapter 4.1 just with a modified log-likelihood function.

In the Penalised BS model, the vector $\boldsymbol{\theta}$ contains the estimated coefficients of the neighbouring boundary above, $\hat{\boldsymbol{\beta}}(z_{t+1})$, or below $\hat{\boldsymbol{\beta}}(z_{t-1})$ to get $\hat{\boldsymbol{\beta}}(z_t)$. Exactly which depends on whether the boundary is shifted up or down.

Because of this special relation, the estimated coefficients depend on all previous estimations. The initial coefficients for $\boldsymbol{\theta}$ are the estimated coefficients resulting from the boundary placed at the median income. This boundary results in a balanced poor, non-poor data split. The SE values of the estimated coefficients are therefore low. Starting from the median boundary, the boundary is shifted upwards to the upper limit $U$, downwards to the lower limit $L$, and then upwards again to the upper limit. This is the minimum amount of boundary shifts in this model. One downward and upward shift of the boundary is called a cycle. Any number of cycles can be added to the minimum number of boundary shifts, and five cycles are added in the following analysis. The coefficients are not estimated for boundaries above $U$ and below $L$ due to the already observed large jumps and the inertia of the Penalised BS model. The consequences of adding the boundaries outside the limits would be, that depending on the downward or upward shift of the boundary the estimated coefficients at the same boundary are very different.

The impact of the tuning parameter is now analysed. The three Penalised BS models with the tuning parameters $\lambda_1 = 0.1, \lambda_2 = 0.001$ and $\lambda_3 = 0.0001$ are compared. The

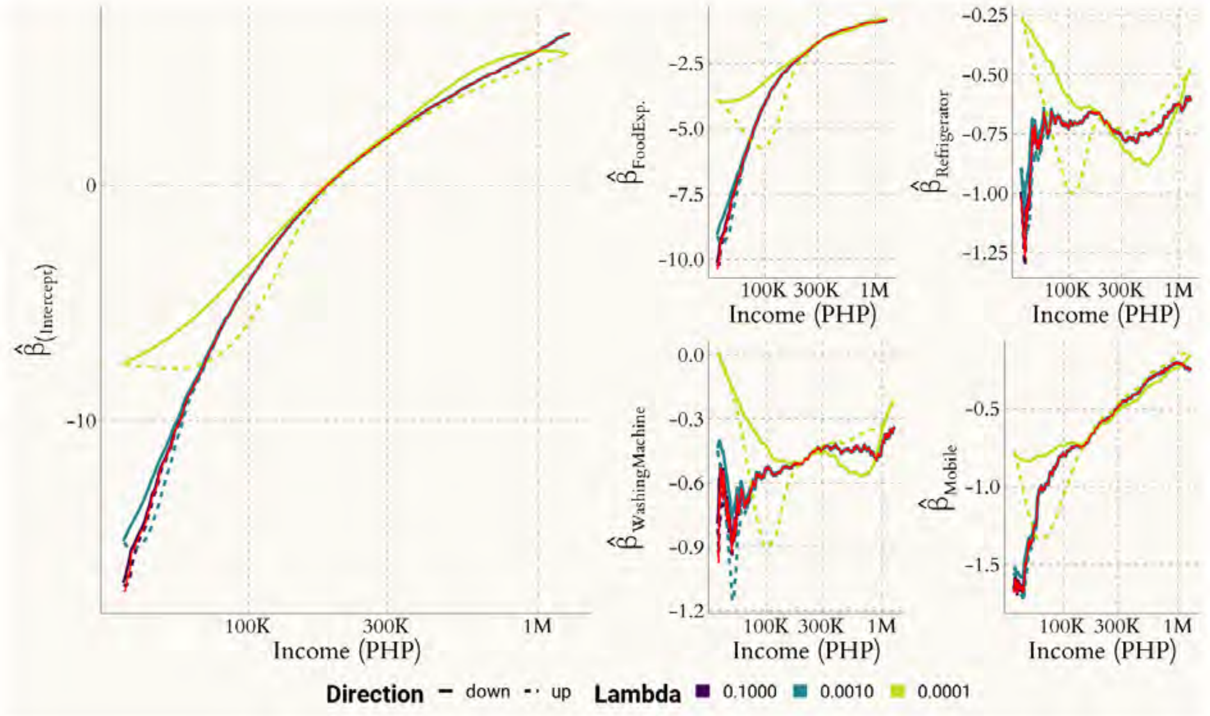


Figure 17: The poverty curves of the $\lambda_1$ BS model deviate strongly from the poverty curves of the Log998 BS model due to strong penalisation. Due to weak penalisation, the poverty curves of the $\lambda_3$ BS model and the Log998 BS model look nearly identical. A suitable degree of penalisation seems to be present in the $\lambda_2$ BS model.

boundaries that are used are those of the Log998 BS model between $L$ and $U$.

Figure 17 displays poverty curves. Each plot in the figure corresponds to a coefficient. The poverty curves of the three models are drawn in the same plot together with the poverty curve from the Log998 BS model in red for comparison. The plots contain just the poverty curves resulting from the last cycle for clearness, whereby the doted poverty curves result from shifting the boundary upwards and the solid line from shifting downwards. Since the poverty curves of the same coefficients are in the same plot, each plot used a different scale. Therefore, seemingly large differences in the refrigerator, washing machine and mobile plots should not be over-interpreted.

It can be seen in the plots that all poverty curves, resulting from shifting upwards and shifting downwards, are connected at the start and end points. This shows that an equilibrium point has been found after five cycles, and further cycles are not needed. Two other striking features are that, first, the poverty curves resulting from the downward and upward shifts diverge more sharply as the lambdas decrease. And secondly, the poverty curves at the lower and upper limits deviate more from the poverty curves of the Log998 BS model.

My theory is that both findings are related to each other. The increasing deviation from the trend of the unpenalised poverty curve of the BS model is due to the penalisation term that forces neighbouring estimated coefficients to be close together. When there is now a strong penalisation, it is more important that the estimated coefficients are close to the previously estimated coefficients. This pulls the poverty curves away from the trend of the Log998 poverty curve when the penalisation is too strong. Examples of that are the intercept and food expenditure poverty curves of the $\lambda_1$ BS model with the strongest penalisation. When the boundary is shifted downwards, the poverty curves deviate further from the trend. The result is that when the lowest boundary at $L$ is reached, and the boundaries are shifted upwards again, the poverty curves still have a downward trend until they approximately cross the poverty curves of the Log998 BS model. This applies in the same way to the upwards shifts of the intercept and the downward shifts of the mobile coefficients.

In the plots where the estimated coefficients do not differ too much, i.e. washing machine, refrigerator and mobile, another behaviour is remarkable for the poverty curves of the $\lambda_1$ BS model with the strongest penalisation. For downward shifts, the poverty curves rise while the trends of the Log998 BS model decrease; for upward shifts, the opposite is true. A hypothesis for this behaviour, which will not be proven, is that this is due to the L2-norm and the relation to the remaining part of the log-likelihood function. The L2-norm has the property that larger vector elements have a greater influence on the resulting norm due to the squaring of the elements. It is thus important to reduce the larger differences to reduce the norm. One could observe this for the intercept poverty curve where the $\lambda_1$ BS model forced the intercept estimations that would be further apart from each other, to be closer. Suppose the log-likelihood function is now disassembled into two parts. In this case, for the penalty term and the remainder, an estimate for $\boldsymbol{\beta}$ is optimal that both maximises the log-likelihood function without the penalty term and minimises the penalty term. Since, as far as I am concerned, this is only possible if $\boldsymbol{\theta}$ is equal to $\boldsymbol{\beta}$, a compromise has to be made. A possible compromise is to minimise larger coefficient differences, like the intercept and food expenditure while increasing smaller coefficient differences of the remaining variables. This could lead to a relatively small penalty term and a relatively large log-likelihood function without a penalty term to get the largest log-likelihood function overall. But this is just speculation and requires further analysis. Whether or not this speculation is correct, one could conclude that the divergence from the trend of the poverty curves of the Log998 BS model is a sign of too much penalisation, as too much attention is placed on minimising the penalty term.

The following continues with the $\lambda_2$ Penalised BS model. For lower boundaries, the intercept and food expenditure poverty curves resulting from upward and downward shifts diverge much less than the corresponding poverty curves in the $\lambda_1$ Penalised BS model. For the poverty curves of the variables washing machine, refrigerator and mobile, the non-intuitive deviations from the poverty curves of the Log998 BS model are still noticeable but now at a lower level. Therefore, for the $\lambda_2$ Penalised BS model, one can conclude that it is an improvement compared to the $\lambda_1$ model in terms of the course of the poverty curves.

The $\lambda_3$ Penalised BS model has nearly identical upward and downward shift poverty curves that nearly perfectly correspond to the poverty curves of the Log998 BS model.

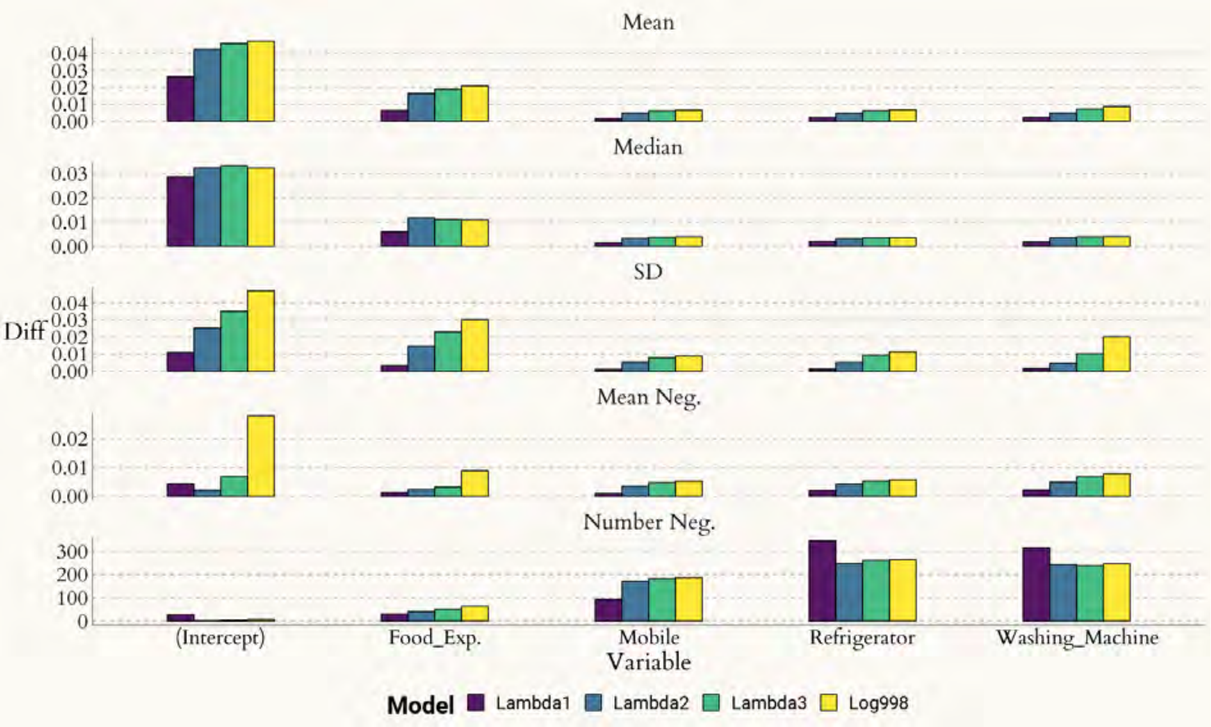


Figure 18: Looking at the statistics confirms that the fluctuation decreases with increasing penalisation. However, the effect of too much penalisation is that the poverty curves deviate from those of the Log998 BS model and have a different trend. One can notice this by looking at the number of negative differences in the intercept estimations, which should be small but is not because the poverty curve of the intercept is not monotonically increasing on the whole range.

This means that the fluctuation from the Log998 BS model, which should be reduced, is now present again to some degree in the $\lambda_3$ Penalised BS model.

The differences of neighbouring estimated coefficients are compared to measure if the fluctuation in the poverty curves of both models is comparable. Also, the differences from the $\lambda_1$ and $\lambda_2$ Penalised BS models are added to this comparison. From the Penalised BS models, the estimated coefficients resulting from downward shifts will be used to calculate the differences since Figure 17 has shown that the poverty curves from downward shifts look more like the ones from the Log998 BS model, which makes the results more comparable.

Figure 18 shows bar plots corresponding to the statistics calculated on the absolute differences $\mathbf{diff}_j$ of the estimated coefficients between the limits. The mean, the median and the SD of the absolute differences are calculated, as well as the mean and the number of negative differences of the different estimated coefficients. Looking at the statistics calculated on the positive and negative absolute differences, one can see that the fluctuation increases with decreasing penalisation. The ranking of the largest mean values of the absolute negative differences only differs for the intercept since the mean absolute negative differences of the $\lambda_2$ BS model are below those of the $\lambda_1$ BS model. One can explain this with the earlier observation in Figure 17. In the intercept plot, one could observe that the poverty curve resulting from shifting the boundaries upwards increasingly deviates from the poverty curve of the Log998 BS model. The consequence of this deviation was that the poverty curve initially continued to rise for decreasing boundaries until the poverty curve approximately crossed the poverty curve of the Log998 BS model. Since the differences are calculated by subtracting the lower boundary's estimated coefficients from the upper boundary's estimated coefficients, negative differences result for the initial boundaries.

Returning to Figure 18 to analyse the number of negative differences. The rankings in the intercept, refrigerator and washing machine plots are different too. The $\lambda_1$ BS model has the largest number of negative differences out of the BS models. In the case of intercept, this is again due to the initial rise in the poverty curve as the boundaries are shifted downwards. The number of negative differences is large for the other two coefficients due to the non-intuitive increase explained by the L2-norm and the relation to the remaining part of the log-likelihood function. Still, although there are larger amounts of negative differences, the mean values of the absolute negative differences are small for all coefficients.

One can conclude that Penalised BS is a good way to reduce fluctuation and get smoother poverty curves. As expected, the larger the value of the tuning parameter $\lambda$ is, the smaller the fluctuation gets. However, with increasing penalisation, the poverty curves resulting from the upward and downward shifts increasingly diverge from the poverty curves of the non-penalised model. Therefore, choosing too large a value for lambda is not recommended.

This was the last chapter that dealt with the fluctuations of the poverty curves. In the following chapter, poverty indices for individuals are calculated in different ways using the Log988 BS mode. Subsequently, the poverty indices are compared with those of other selected fuzzy poverty measurement methods from Chapter 3.2.

# 9 Poverty Prediction

This final chapter establishes a link to the fuzzy poverty measurement methods from Chapter 3.2. The fuzzy poverty measurement methods had in common that they were used to calculate a poverty index for each household. From BS models, poverty indices can be derived in various ways, too, and the procedure of deriving the poverty indices is explained in the following.

First, it is shown how poverty is predicted for households at each boundary. Afterwards, the predictions are combined into a single poverty index. Lastly, the poverty indices are compared to the variable income and the poverty indices of other fuzzy poverty measurement methods.

## 9.1 Prediction Curves

Logistic regression has been used in the previous chapters to analyse the influence of the independent variables $x_1, \dots, x_k$ on the binary dependent variable $y$ and therefore, the coefficients $\boldsymbol{\beta}$ were estimated. One can use the estimated coefficients to predict the probability of an object belonging to a class because

$$\widehat{f(x_1, \dots, x_k)} = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_k x_k)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_k x_k)}, \tag{46}$$

applies and one can see the function $\widehat{f(x_1, \dots, x_k)}$ as an estimator for $\mathrm{E}(\widehat{y \mid x_1, \dots, x_k})$ (Fahrmeir et al., 2009, p. 22). For given values $\boldsymbol{x}$ the prediction of $y$ will be referred to as $\hat{y}$.

In BS, at each boundary $z_t$, one can predict poverty for each household with the attributes $\boldsymbol{x}$ to get the predictions $\hat{y}(z_t, \boldsymbol{x})$. Similar to the poverty curves, if the predictions with the corresponding boundaries are drawn as points in a coordinate system, nearby predictions can be connected. The resulting curve is called the **prediction curve** of a particular household.

Figure 19 shows the prediction curves for some households of the Filipino household data set with incomes that correspond to the income percentiles. The boundaries and estimated coefficients of the Log998 BS model have been used to get the prediction curves. The colours of the prediction curves indicate the income rank, which means that the household with the lowest income has a very dark magenta-coloured prediction curve and the one with the highest income a yellow-coloured one. All the prediction curves are s-shaped, and due to the logarithmically scaled x-axis, they seem to be parallel. Although the prediction curves are seemingly monotonically increasing, there are counterexamples where the probability of being poor for a household is slightly lower at a higher boundary. A connection with the variable income can be assumed since the prediction curves become more yellow with increasing household incomes. At this position, one must note that the prediction curves are predicted with the values of the variables food expenditure, refrigerator, washing machine and mobile. The variable income itself is not a predictor variable; therefore, household income is not used when predictions are made.

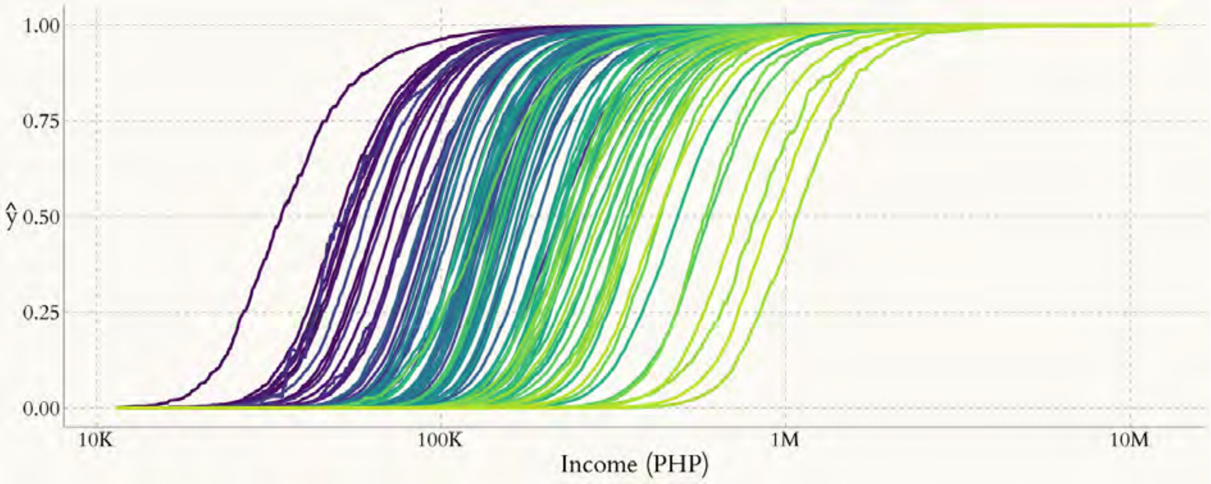


Figure 19: The plot shows the prediction curves corresponding to 100 households. The prediction curves are s-shaped and, as the colour of the curves shows, approximately ordered by income. The prediction curves appear to be completely parallel due to the logarithmic scaling of the x-axis.

One interprets the prediction curves as follows: For an increasing boundary, it is more likely that a household is poor, i.e. if the boundary is placed at 100,000, the household corresponding to the very dark magenta-coloured prediction curve has an expected probability of nearly one to be poor. But the household corresponding to the yellow prediction curve has an expected probability of approximately zero to be poor at this boundary.

Since the prediction curves are impracticable for fuzzy poverty measurement, as they can not be compared to the poverty indices of other fuzzy poverty measurement approaches, the information of each prediction curve must be summarised in a single value. One could see this single value as the poverty index. Therefore the whole process of BS, in addition to deriving the prediction curve and summarising it into a single poverty index, one could see as the membership function $\mu$ that is known from the fuzzy poverty measurement approaches. This means that BS is a fuzzy poverty measurement approach that assigns each household a poverty index.

The following chapter presents different approaches to combining the prediction curves into a single value.

## 9.2 BS Poverty Index

There are different approaches for getting poverty indices from the prediction curves. One could come up with the four procedures

- **Mean**,
- **Median**,
- **Above50**,

- **Maxslope**.

In the Median and Mean approaches, the statistics of the same name are calculated from the estimated probabilities of all boundaries for each household to get $\mu_i^{Median}$ and $\mu_i^{Mean}$. As a result, the Median poverty index is equal to the predicted probability at the median boundary. Therefore, having many boundaries would not be required since only the predictions at a single boundary are used to get the poverty indices. The Above50 method returns the boundary as the poverty index $\mu_i^{Above50}$, where the predictions are greater than or equal to 0.5 for the first time. In the fourth approach, Maxslope, the slopes $m_t$, $t = 1, \dots, T-1$ of the lines, between the two points $\big(z_t, \hat{y}(z_t, \boldsymbol{x}_i)\big)$ and $\big(z_{t+1}, \hat{y}(z_{t+1}, \boldsymbol{x}_i)\big)$ are calculated. The maximum slope $\max\limits_{1 \leq t < T} (m_t)$ is then used as the poverty index $\mu_i^{max}$.

The poverty indices from the Median and Mean approaches are subject to the condition $\mu_i^{Median}, \mu_i^{Mean} \in (0,1)$ since the poverty indices are calculated from expected probabilities. This differs from the poverty indices resulting from the Above50 and Maxslope methods as for these methods $\mu_i^{Above50} \in \{z_t \mid i = 1, \dots, t\}$ and $\mu_i^{max} \in \{x \mid x > 0\}$ applies.

The poverty index $\mu_i^{Above50}$ will be brought on the 0 to 1 scale with the formula

$$\mu_i^* = \frac{(1 - \lg(\mu_i)) - \min\limits_{1 \leq l \leq n} (1 - \lg(\mu_l))}{\max\limits_{1 \leq l \leq n} (1 - \lg(\mu_l)) - \min\limits_{1 \leq l \leq n} (1 - \lg(\mu_l))}. \tag{47}$$

This modified normalisation makes the poverty indices of the different fuzzy poverty measurement approaches comparable. Using the logarithm of the poverty index results in a preferable relation to the logarithmic income, as seen in the following chapter. Since wealthy households have a high poverty index before normalisation, the poverty index must also be subtracted from one during normalisation.

For the poverty indices resulting from Maxslope, it is not required to subtract them from one since households with lower incomes have steeper poverty curves than the ones with higher incomes. But there is still a favourable relationship between the logarithmic Maxslope poverty indices and logarithmic income, and therefore, the formula

$$\mu_i^* = \frac{\lg(\mu_i) - \lg\left(\min\limits_{1 \leq l \leq n} (\mu_l)\right)}{\lg\left(\max\limits_{1 \leq l \leq n} (\mu_l)\right) - \lg\left(\min\limits_{1 \leq l \leq n} (\mu_l)\right)}, \tag{48}$$

is used for scaling the Maxslope poverty indices.

One could expect that all approaches depend on the underlying boundaries. I would argue that the approach Mean is especially dependent on the placement of the boundaries. Because the Log998 BS model will be used for the predictions, the boundaries are not distributed according to income. Instead, the neighbouring boundaries for lower income values are closer than those for higher income values. One could speculate that this leads to poverty indices that are biased towards zero if one assumes that the poverty indices in the Basic BS model are unbiased. To note here, one could further speculate that the poverty indices of the Permilles BS model would be unbiased, but this will not be proven.

The poverty indices of the Above50 model correspond before normalisation to the boundary values, and therefore, the poverty indices are also dependent on the boundary placement. The consequence is that even after normalisation, only $T$ unique poverty indices are achievable for the 41,544 households in the case of the Filipino data set. Since $T \leq n - 1$ always applies, there are at least two households with the same poverty index which means that one can never strictly order the households according to the poverty index. Equal poverty indices are unlikely to occur in the Mean and Median approaches. The only two scenarios I can think of are either that two households have identical attributes or the unlikely case that the mean or median values calculated from the predictions $\hat{y}(z_t, \boldsymbol{x}_i)$ are indeed the same.

One could expect that the Maxslope poverty indices depend on the boundaries because the prediction curves are approximately s-shaped and not linear increasing. Consequently, the slope $m_t$ for $t = 1, \ldots, T - 1$ of the line between two points changes if $z_t$ or $z_{t+1}$ is different.

Next, the different poverty indices derived from BS poverty measurement are compared. Also, they are compared with the variable income and poverty indices of non-BS poverty measurement methods.

## 9.3 Comparison

This chapter compares the poverty measurement methods Mean, Median, Above50, Maxslope, TFR and VW. The BS poverty indices will be calculated for each method in two ways. One is to calculate the poverty indices from predictions of boundaries between the limits. The other is to calculate them from the predictions of all boundaries, resulting in eight BS poverty measurement methods. The TFR and VW poverty indices will also be derived in two ways. In the first, the variable income is used in addition to the variables food expenditure, refrigerator, washing machine and mobile. In the second, the variable income is not added to the models to see if the results are similar. Both approaches are implemented as described in Chapter 3.2. One should note that the methods are directly addressed by their names, meaning that the BS methods with poverty indices calculated from boundaries between the limits are named limited BS methods. The same applies to the non-BS methods, which are addressed as income non-BS methods and non-BS methods.

### 9.3.1 Comparison to Income

First, the poverty indices are compared with income. Figure 20 shows the income and poverty index pairs drawn as points in a coordinate system. The x-axis of the plots is again logarithmically scaled. The strong relationship between the various poverty indices and income stands out. A decreasing trend is visible in all plots, even in the VW and TFR plots where the variable income has not been used. One can notice a linear downward trend in the limited and unlimited Mean and Above50 plots. The trend is s-shaped in the

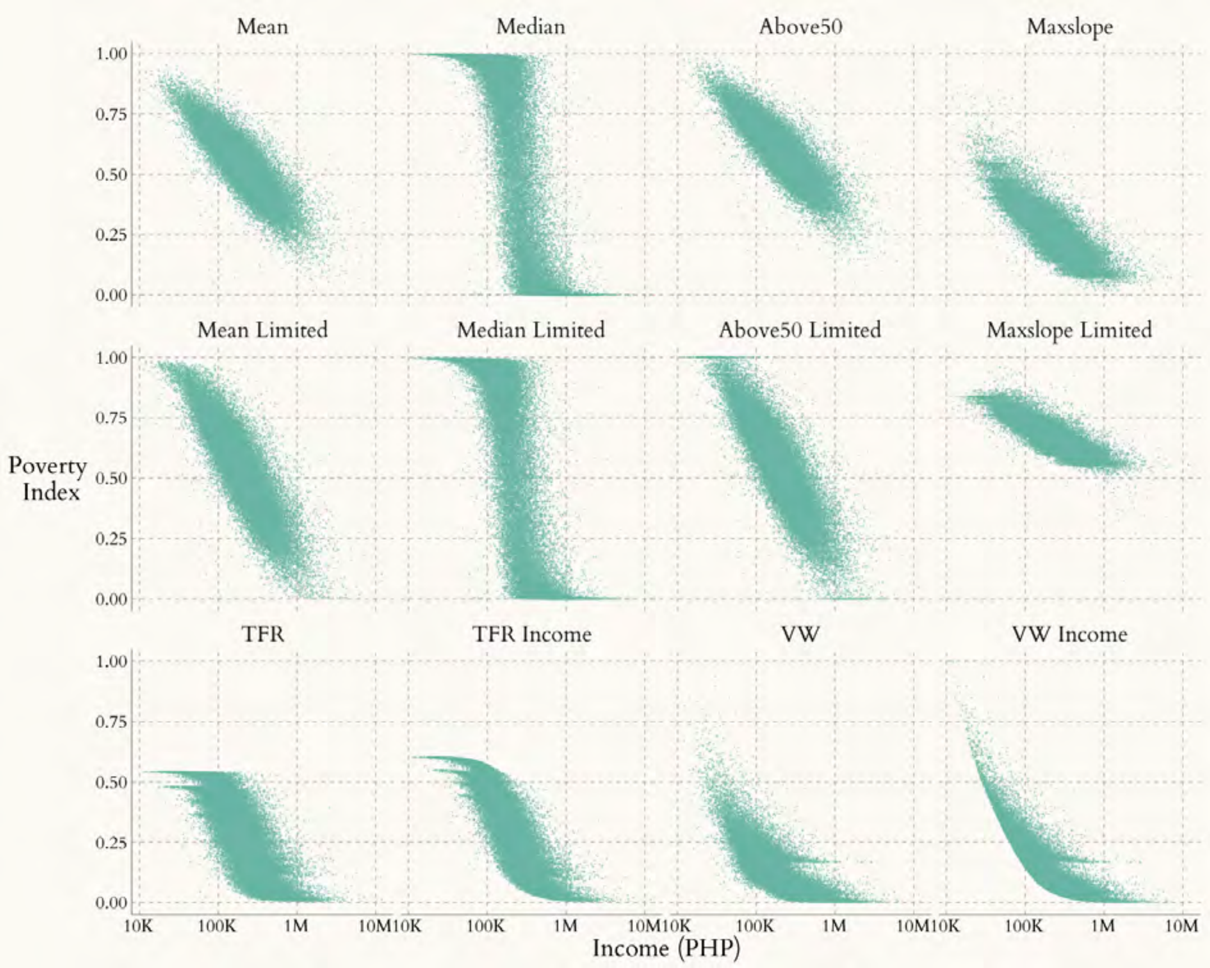


Figure 20: Poverty index and income of households drawn as points in a coordinate system. One can see a linear decreasing trend for the methods Mean and Above50. The poverty indices calculated with the limited Maxslope method are mostly larger than 0.5. The poverty indices of the non-BS methods decrease with increasing income too. Even when income has not been used in the calculation.

Median plots, and one can describe the trend in the unlimited Maxslope plot with the function $f(x) = \frac{1}{x}$ on the domain $\{x \in \mathbb{R} \mid x > 0\}$. An explanation for this special trend can not be given. The trend in the limited Maxslope plot is decreasing but not as steep, resulting in most poverty indices being in the range of 0.5 to 1. Just 46 households have a poverty index below 0.5. Moreover, households with very low incomes have poverty indices below one, while some households with slightly higher incomes have poverty indices close to one.

It follows a possible explanation for the different poverty indices of the limited and unlimited Maxslope method. One can assume that the prediction curves that were defined in the previous chapter are s-shaped and therefore have commonalities with the Sigmoid function

$$sig(x) = \frac{1}{1 + e^{-x}} \tag{49}$$

where $x \in \mathbb{R}$. The Sigmoid function is a strictly monotonic increasing function with its steepest slope at the inflexion point. Assuming that each prediction curve is a function with its inflexion point, then the slope at the inflexion point is the resulting poverty index from the Maxslope method. Now, if the x-coordinate of a prediction curve's inflexion point is below the lower limit or above the upper limit, the slope at this point can no longer be the poverty index. Instead, the steepest slope of a point on the prediction curve with an x-coordinate between the limits has to be used. Because the slope of the prediction curve increases until the inflexion point and decreases afterwards again, the x-coordinate of the point with the steepest slope corresponds to the limit that is closer to the inflexion point's x-coordinate. This means for the limited Maxslope's poverty indices that, if the x-coordinate of the steepest slope is below the lower limit or above the upper limit, the steepest slope of the prediction curve is not taken. Instead, the slope of the prediction curve at the upper or lower limit is returned as the poverty index.

The consequence is that for extremely poor households, which have most likely very low incomes, the normalised limited Maxslope poverty index is below one, as households with more income have a steeper slope. For non-poor households, where the steepest slopes are not so steep, the steepest slope reduces even further.

One can hardly see any difference in the Median plots because the trend is the same. A property of the median is that if the $i$ with $i < \frac{n}{2}$ largest and, simultaneously, the $i$ smallest values of a sample with $n$ observations are removed, the median of the reduced sample stays equal. Since there are approximately the same number of boundaries above the limit as below the limit, the median boundary changes slightly when the boundaries outside the limits are removed. As the Median poverty indices correspond to the poverty predictions at the median boundary, the consequence of limiting is that the predicted probabilities at a boundary close to the initial boundary are taken. These predictions at this slightly different boundary should not differ that much.

Heading to the methods Mean and Above50. Most of the poverty indices from the unlimited models are larger than zero and smaller than one, but poverty indices close to zero or one are rare. The poverty indices calculated with the methods limited Mean and limited Above50 are more evenly distributed between zero and one. One could explain the more even distribution in the limited Mean plot with the fact that for boundaries below the lower limit and above the upper limit for many households, respectively, the expected probability of being poor is approximately zero and one. But for example, some predicted probabilities might already be larger than zero for very poor households at boundaries right below the lower limit. Suppose now the expected probabilities for boundaries outside the limits are not used. In that case, the poverty index is calculated in both methods with mostly probabilities of approximately one and just a few below one resulting in a poverty index close to one. One can explain this in the same way for richer households.

The seemingly better spread of poverty indices in the limited Above50 plot can not be explained, but that there are many households with a poverty index of exactly zero or one can be explained. The situation is that there are poorer households, where the estimated probability of being poor is above 0.5 at a boundary below the lower limit, or, in the opposite case, wealthier households, where the estimated probability of being poor is the first time above 0.5 at a boundary above the upper limit. To be able to assign a value to these households, it has been decided to assign the poorer households the lower limit as the poverty index and the wealthier households the upper limit.

Figure 27 in the appendix shows the prediction curves that lead to a poverty index of zero or one when the method limited Above50 is used. One can see that some prediction curves cross the horizontal line at 0.5 below the lower limit, and some cross the horizontal line above the upper limit. The upper and lower limit are visualised as vertical lines. The colour of the prediction curves indicates if a household has an income above $500,000$. This shows that the prediction curve is very high for a few households with an income below $500,000$. However, the exact reason for this is not being researched.

Heading to the plots resulting from non-BS methods, one can notice a s-shaped trend in the TFR plots. A trend that corresponds to the function $f(x) = \frac{1}{x}$ on the domain $\{x \in \mathbb{R} \mid x > 0\}$, one can see in the VW plots. An explanation for the trends can not be given. Both TFR plots show that most poverty indices are below approximately 0.6. Since the points in the plot that displays the result of the TFR approach without income are a bit more spread than in the plot of the TFR approach with income, one can see that there is, as expected, a weaker relation to income. One can conclude the same for the VW plots. It is conspicuous in the VW plot where income is used that a curve seemingly limits the points in the plot, meaning that there are no points below a certain curve. To explain this, one should remember that in the VW method, the information on all variables is used from the beginning to determine the proportion of households that are at least as poor as the household itself. Since the VW income poverty indices are drawn with the corresponding income value in the coordinate system, one can say at each observed income that the poverty index of a household with less income has to be larger since the proportion of households that are at least as poor as the individual itself is smaller. Due to calculating the deprivation indicator and normalisation as explained in Chapter 3.2.4, the downward limitation with the falling trend occurs. Finally, mentioning the poverty index range of the VW approaches, one can see that the poverty indices range from zero to one, but most poverty indices are below approximately 0.8.

Table 4 shows the correlation between the poverty indices resulting from the different methods and income. The correlation between the poverty indices and the logarithmic income is also given.

The poverty indices of all methods are extremely negatively correlated with the logarithmic income. With income itself, the poverty indices are less negatively correlated. One can observe the least negative correlation with income, of approximately -0.46, for the VW method without income and the second least for the VW method with income. The highest negative correlations with income of approximately -0.65 are observable for the Above50 and Mean methods. The remaining correlation values are not conspicuous. The correlations with the logarithmic income range from -0.76 for the VW without income to -0.87 for the methods TFR with income, Mean and Above50. This is a surprise as the method income TFR had the fifth highest correlation with income. This is a sign that

| Method | Income Correlation | Log Income Correlation |
|---|---|---|
| Mean | -0.65 | -0.87 |
| Median | -0.61 | -0.80 |
| Above50 | -0.65 | -0.87 |
| Maxslope | -0.59 | -0.84 |
| Mean Limited | -0.64 | -0.87 |
| Median Limited | -0.60 | -0.81 |
| Above50 Limited | -0.64 | -0.87 |
| Maxslope Limited | -0.62 | -0.84 |
| TFR | -0.58 | -0.81 |
| TFR Income | -0.62 | -0.87 |
| VW | -0.46 | -0.76 |
| VW Income | -0.49 | -0.83 |

Table 4: Correlation between the different poverty indices and income or logarithmic income. The correlation with logarithmic income is higher for all methods. Furthermore, all methods have a more similar correlation with logarithmic income. The non-BS methods, which do not include income, show the lowest correlation with logarithmic income.

income is highly weighted in the income TFR approach.

One can summarise from Figure 20 and Table 4 that there is a strong relationship between the different poverty indices and income. For the non-BS methods, the relationship is stronger if the variable income is used to calculate the poverty indices. One could observe this in the figure by the greater dispersion of the points and in the table by the stronger correlation. For the non-BS methods, where income has not been used in the poverty index calculation, it could further be seen that the TFR approach has a stronger relation to the logarithmic income than the VW approach. The Mean and Above50 poverty indices have a very similar and strong relationship with logarithmic income, although the methods differ.

Suppose the aim is to identify the extremely income poor or income rich. In that case, the Median method is probably not good since only households with medium to high or medium to low incomes have poverty indices ranging from zero to one. With this method, most households will have a poverty index of zero or one. The Above50 approach has the favourable property that the poverty indices linearly decrease with logarithmic income and that the extremely income poor or rich have one or zero poverty indices. Although the poverty indices of the limited Maxslope approaches are strongly negatively correlated with income, it seems like this approach is unpractical since the majority of poverty indices are between approximately 0.5 and 0.9, which is due to extreme values that occur because of the limitation.

This has now shown that the BS methods have a generally high relation to income and can therefore be good at identifying income poor. Regarding correlation, the BS and non-BS methods are similarly correlated with the logarithmic income. Whether a strong relation to income is good in fuzzy poverty measurement is questionable. It was said at the beginning that fuzzy poverty measurement should overcome the issue of seeing poverty as a one-dimensional problem. If the logarithmic income now corresponds to the poverty indices, one could come up with two conclusions in my view. Firstly, contrary to the opinion of many experts, using just income as an indicator of poverty is sufficient to determine if a household is poor. Secondly, the fuzzy poverty measurement method can not capture all dimensions of poverty. Since the correlations of poverty indices of the BS methods with the logarithmic income is not that much higher than the correlation of the non-BS methods, where the variable income has not been used, with income, I would argue that BS can capture more dimensions than just one.

The following chapter continues by comparing the poverty indices of the methods.

### 9.3.2 Comparison Between Methods

The correlations of the poverty indices resulting from the different methods are analysed to see if there are extreme differences in the ordering of the households according to the poverty indices between the methods.

The correlations are displayed in the heat map in Figure 21. As expected, the methods limited Mean, limited Above50, unlimited Mean and unlimited Above50 are perfectly pairwise correlated. The other two BS methods, Maxslope and Median, are just with their limited counterparts nearly perfectly correlated. The same counts for the non-BS methods as they are nearly perfectly correlated with their counterparts where the variable income has not been used. One can see low correlations between the VW methods and the remaining methods. The lowest correlations of 0.71 to 0.74 are apparent between the VW and Median methods. The pairwise correlations between both VW methods and the other eight methods range from 0.80 to 0.93. The correlation between the Maxslope methods and the Median methods is, compared to the correlations with the other methods, not too high as the correlations range from 0.89 to 0.91. On average, the Maxslope methods, with correlations ranging from 0.97 to 0.98, are highly pairwise correlated with the Mean and Above50 methods. The remaining pairwise correlations that have not been mentioned range from 0.92 to 0.95.

In terms of correlation, the heat map has shown that using the limited BS methods or including income in the non-BS method has just a small impact. Further, the ordering of the households according to the poverty indices of the methods Mean and Above50 is nearly the same. The correlation between other methods is a bit lower. As the correlation is not a good indicator for similar ranking of the households according to the poverty index, if there is no linear relation between the poverty indices, the poverty indices of the methods are pairwise plotted.

The results are shown in Figure 22. The limited BS and non-BS methods with income are excluded from this comparison due to the similar poverty indices of the methods' counterparts. One can perceive a gradual arrangement of points in the plots where non-BS methods are compared to other methods. This is probably due to the integer values of the variables mobile, refrigerator and washing machine, which cause a step-wise increase in the poverty index. It will not be confirmed if this is the actual reason. As already

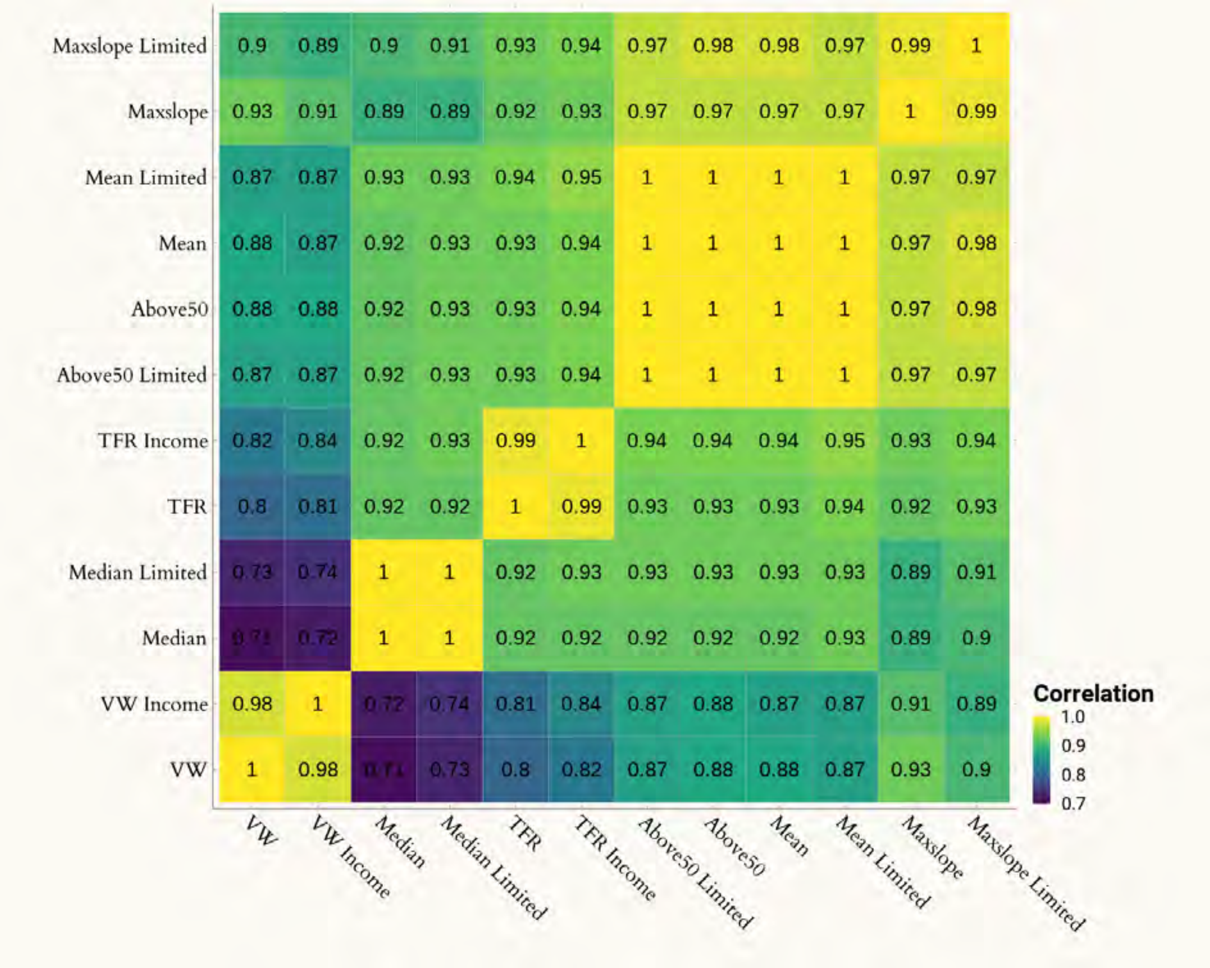


Figure 21: The poverty indices calculated with the same method but slightly different assumptions regarding using income or excluding boundaries outside the limits are nearly perfectly correlated.

concluded from the heat map, there is a nearly perfect linear relationship between the poverty indices of the Mean and Above50 methods. Although the pairwise correlation between the Median methods and the Mean methods and also Above50 were not too high, one can still see that the rankings of the households according to the poverty indices are more similar than expected. Just the scaling is very different, resulting in an s-shaped trend. The points in the plot that shows the Median versus Maxslope poverty indices are s-shaped arranged too, but not as clearly. That indicates a stronger relationship between the two approaches than one might have thought of after looking at the correlation.

The high correlation between the poverty indices of the methods Above50 and Maxslope, which one could notice in the heat map, can also be seen here. One can see something similar in the Mean versus Maxslope and the Above50 versus Maxslope plots, but not as clear. In these two plots, the points are gradually arranged, resulting in seemingly shifted lines of points.

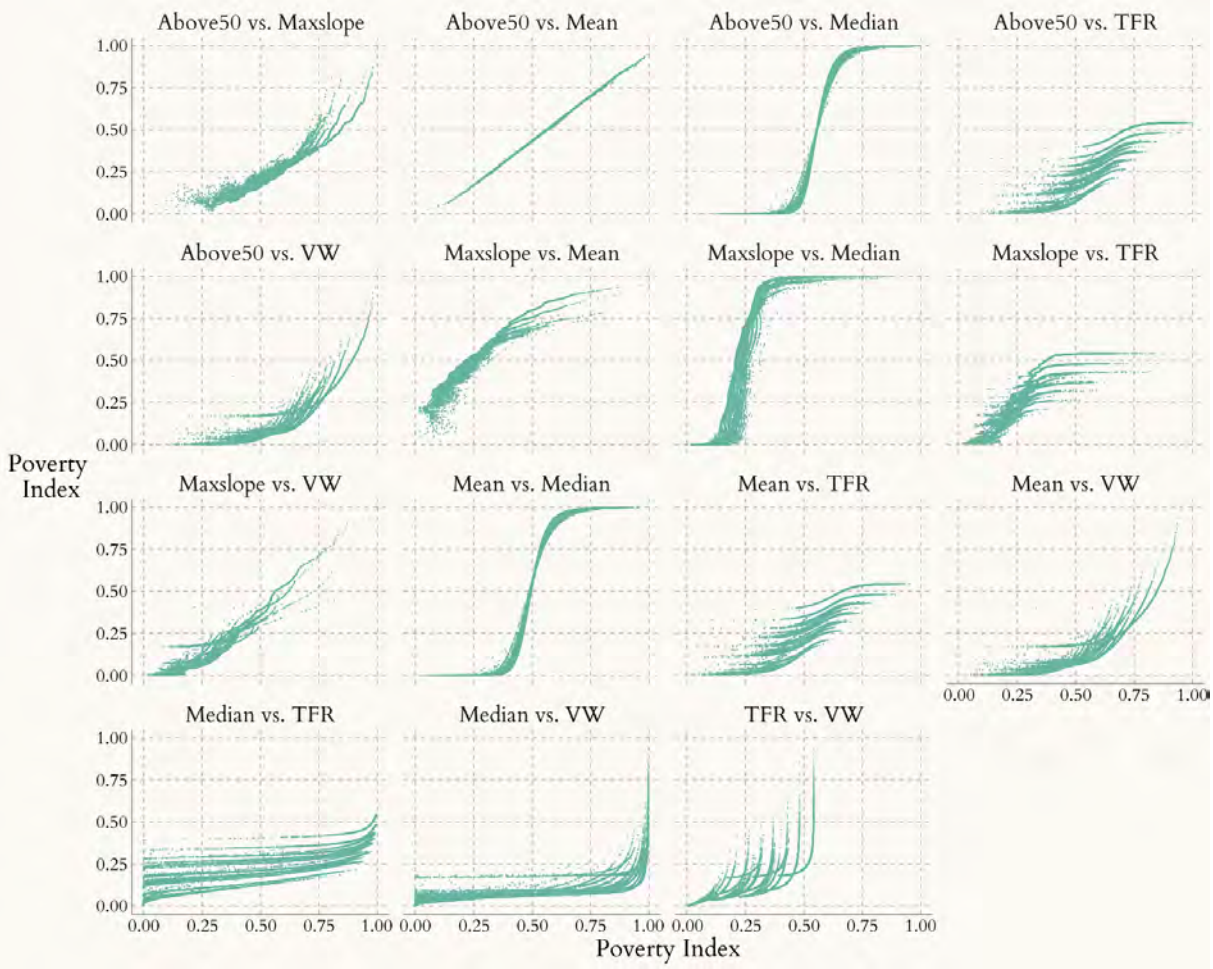


Figure 22: The poverty indices of the methods Mean and Above50 are nearly identical. Between the poverty indices of Mean or Above50 and other BS methods is still a non-linear relationship. In the plots where BS and non-BS methods are compared, the points are not properly spread around a line, showing that the different modelling approaches tend to return different poverty indices.

In the remaining plots, the poverty indices of at least one non-BS method are drawn against the poverty indices of another method, and as previously mentioned, one can perceive gradual arrangements of points. Suppose one sees these plots as a total. In that case, it is obvious that the correlations with non-BS methods poverty indices were smaller compared to other correlations since the points are just approximately spread around a non-horizontal straight line. This suggests that the same households' poverty indices differ for BS and non-BS methods. But this is not a big surprise since it was already mentioned in the fuzzy poverty measurement discussion in Chapter 3.3 that research has shown that the approaches VW and TFR categorised different households as poor. Therefore, it was certain that BS could also classify different households as poor. One can still see that the poverty indices of the VW and BS approaches are more similar than those of the TFR approach. Examples are the Mean versus VW and the Maxslope versus VW plot, as one can see an increasing trend. One can observe the increasing trend in the Mean versus

TFR plot too, but not as clearly.

For interpretation, there are now the following consequences. Firstly, suppose there is a linear relationship between the poverty indices of two different methods, which is present when the pairs of poverty index values lie on a straight line. In that case, poverty index values from one method can easily be transformed into poverty index values of other methods, and relative poverty index differences can be interpreted similarly. Secondly, suppose the pairs of poverty indices lie on a curved line. In that case, relative poverty index differences can not be directly compared between the methods but the rank of households according to the poverty index is still comparable. Therefore, one could use any method to determine which of the two randomly chosen households is poorer, as the methods produce similar rankings.

It was shown in this chapter that although the poverty indices are differently derived, some are still highly pairwise correlated and very correlated with income. For the BS methods, it could be established that the order of the households according to the poverty indices of the BS methods is comparable. The poverty indices of the methods Mean and Above50 are even nearly identical. Comparing non-BS methods poverty indices with other BS or non-BS methods poverty indices has shown that the derived poverty index differs between models. Nevertheless, the VW methods tend to produce more similar poverty indices to the BS methods than the TFR methods.

# 10 Conclusion

Various fuzzy poverty measurement-related topics were covered. Firstly, poverty itself was analysed to understand that fuzzy poverty measurement methods can be an alternative to conventional poverty measurement approaches as they can deal with the vagueness and multidimensional nature of poverty. Afterwards, fuzzy sets were introduced to understand the fuzzy poverty measurement methods TFA, TFR and VW, and their issues. The main issues with fuzzy poverty measurement were that many decisions must be made when these methods are applied. In some methods, decisions are arbitrary. The consequence is that the poverty indices of the models can be very different, meaning that different individuals are considered poor depending on the method.

Subsequently, logistic regression analysis was briefly introduced to refresh knowledge on estimating the coefficients. In addition, logistic lasso regression was introduced to perform variable selection and better understand a later chapter. In the Boundary Shifts chapter, the procedure of defining boundaries, splitting the data set according to the boundaries and performing logistic regression to get the boundary-dependent coefficients was described.

Afterwards, the data set was introduced, standardised and cleaned. In the subsequent exploratory data analysis, it was found that expenditure-related variables, variables describing the number of goods and income, are pairwise moderately correlated. Furthermore that the categories in categorical variables are very unbalanced.

The first BS model revealed that in the case of data sets with many variables, a variable selection is required to get clear results. Therefore, the data set was reduced using logistic lasso regression to end up with the data set containing the variables food expenditure, mobile, refrigerator, and washing machine in addition to income.

In the first BS model, it was also observed that there are jumps in the poverty curves of categorical and count variables. It was found that these jumps occur due to quasi-complete separation, which has been handled by not using categorical variables and removing the boundaries that cause it. As BS by design is affected by data imbalance, causing larger Standard Errors, even more boundaries have been removed.

The Basic BS model, where boundaries are placed at all meaningful incomes, showed that the variables' poverty curves fluctuate and that the relative importance of the variables decreases the higher the boundaries are set. Because placing boundaries at each meaningful income value can be time-consuming, subsequent sensitivity analyses were performed to determine how boundary placement affects the fluctuation of the poverty curves. It was observed that a logarithmic placement is preferable as it can account for the distribution of income, but other placements can also have their advantages. Using other binary prediction models in BS than the logistic regression model was also investigated. However, other binary prediction models could not reduce the fluctuation in the poverty curves.

To still reduce fluctuation, Penalised BS was introduced. In this binary prediction model, a penalty term is added to the log-likelihood function to shrink the coefficients towards the estimated coefficients of the previous boundary. This BS model was indeed able to reduce the fluctuation of the poverty curves. Still, strong penalisation caused a strong deviation from the trend of the poverty curves of the unpenalised BS model.

Subsequent pairwise comparisons of the BS and non-BS fuzzy poverty measurement methods showed that the poverty indices determined from the methods correlate strongly with the variable income. A strong relationship between the BS poverty indices was also found. A strong relationship with the non-BS methods was not observable. This showed that the poverty indices of the BS methods deviate from those of other methods.

It was seen that BS could be an alternative to other methods, which offers the added value of being able to see which variables play a major role in the separation of poor and non-poor at different boundaries. Firstly, this could enable the government or other organisations to determine what distinguishes those with a low income from those with a higher income in order to provide specific goods to compensate for the income deficit. And secondly, the possibility of identifying the poorest individuals based on the poverty index and then compensating them.

BS is also practical because no variable selection is required to get a poverty index. It is only advised for the interpretation of the relative importance. This makes BS less arbitrary, and each individual's situation can be recorded as accurately as possible. Nevertheless, arbitrary decisions still have to be made regarding the number and placement of the boundaries, the binary prediction method and the method of deriving a single poverty index from the poverty predictions.

One issue that has not been addressed is whether the poverty indices of the households change if the boundaries are placed according to an expenditure variable and not income. Further research could examine how the number and location of boundaries affect the poverty index or whether using another binary prediction model in BS adds value. Further work could also investigate whether the use of a norm other than the L2-norm or the standardisation of the coefficients could improve the Penalised BS model.

To finish this thesis, a final comment and appeal. Poverty is a disease. You do not choose to be affected by it, but if you are, you have to suffer. Since it is difficult to escape poverty, it is our responsibility to assist those in need rather than to cause more individuals to fall into it via selfishness.

# A Appendix

This figure corresponds to the exploratory data analysis showing the relationship between income and rice expenditure.

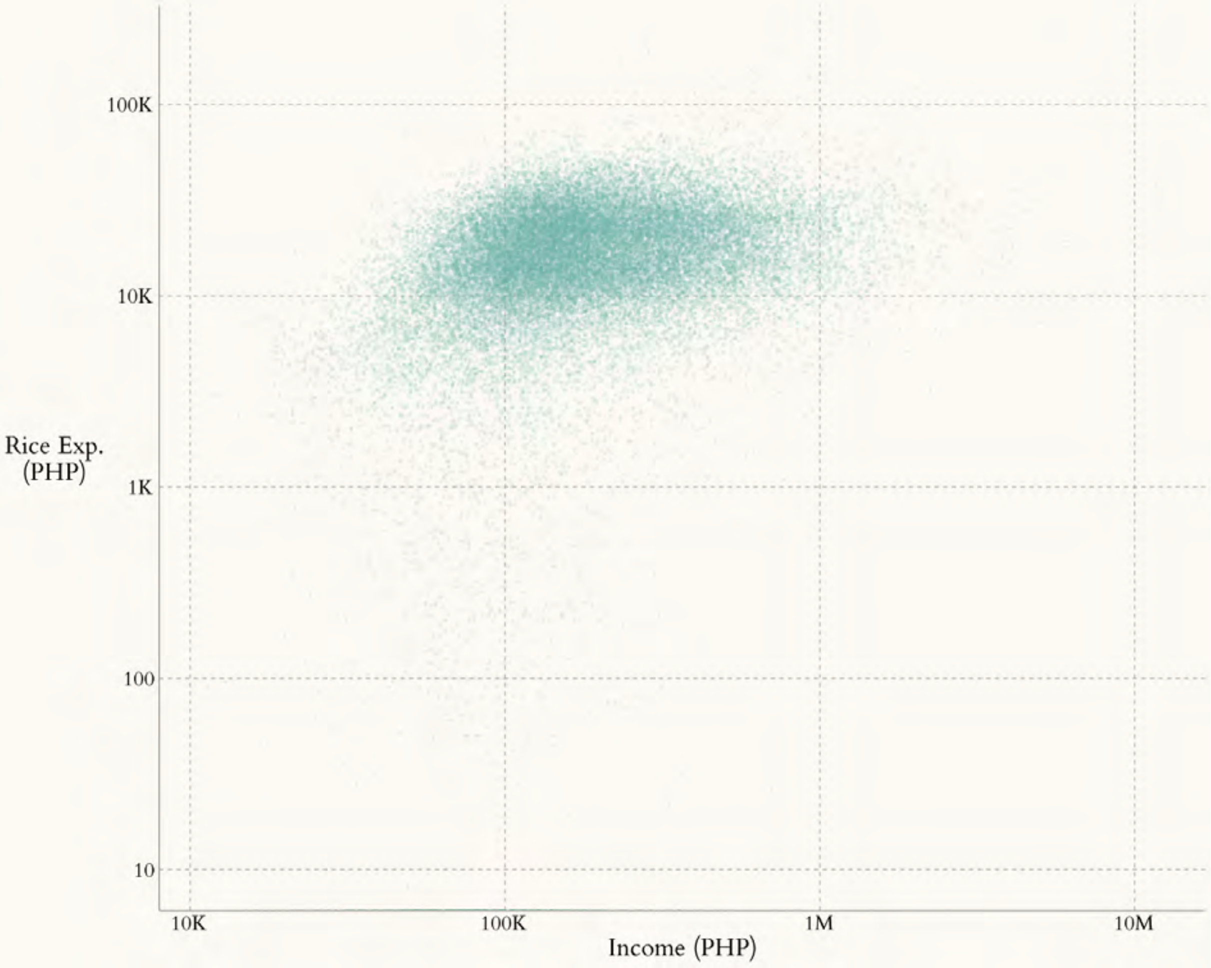


Figure 23: Each point represents a household's income and rice expenditure. With increasing income, households can afford more rice, but at a certain income, households no longer buy more rice, and the rice expenditure thus remains constant. At this certain point, one can expect that households will buy more expensive food in addition to rice. For income-rich households, it might even be the case that they buy more expensive food instead of rice, and therefore, the rice expenditure slightly decreases at some point again.

The following figure corresponds to the exploratory data analysis. It shows that the family size increases as income increases, which could be due to the fact that the more individuals live in a household, the more can work.

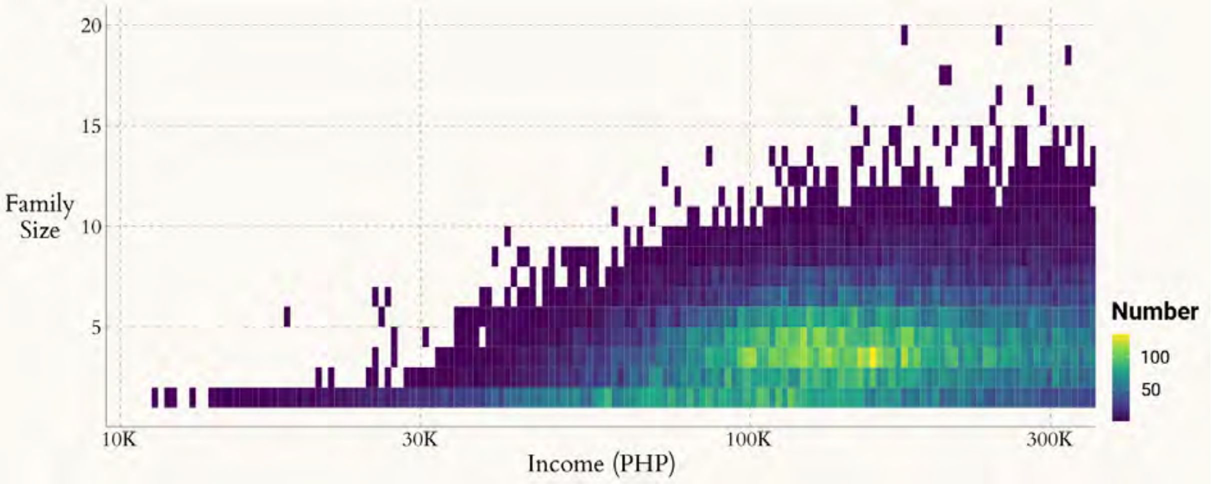


Figure 24: This figure must be interpreted as a two-dimensional histogram where the colour of the tiles indicates how many households are in a certain income range and how large the family size is. One can see that with increasing income, the family size increases to an income of approximately 110,000.

The figure shows the poverty curves of some count variables. This figure is used in Chapter 7.3.2 to show commonalities in the variables.

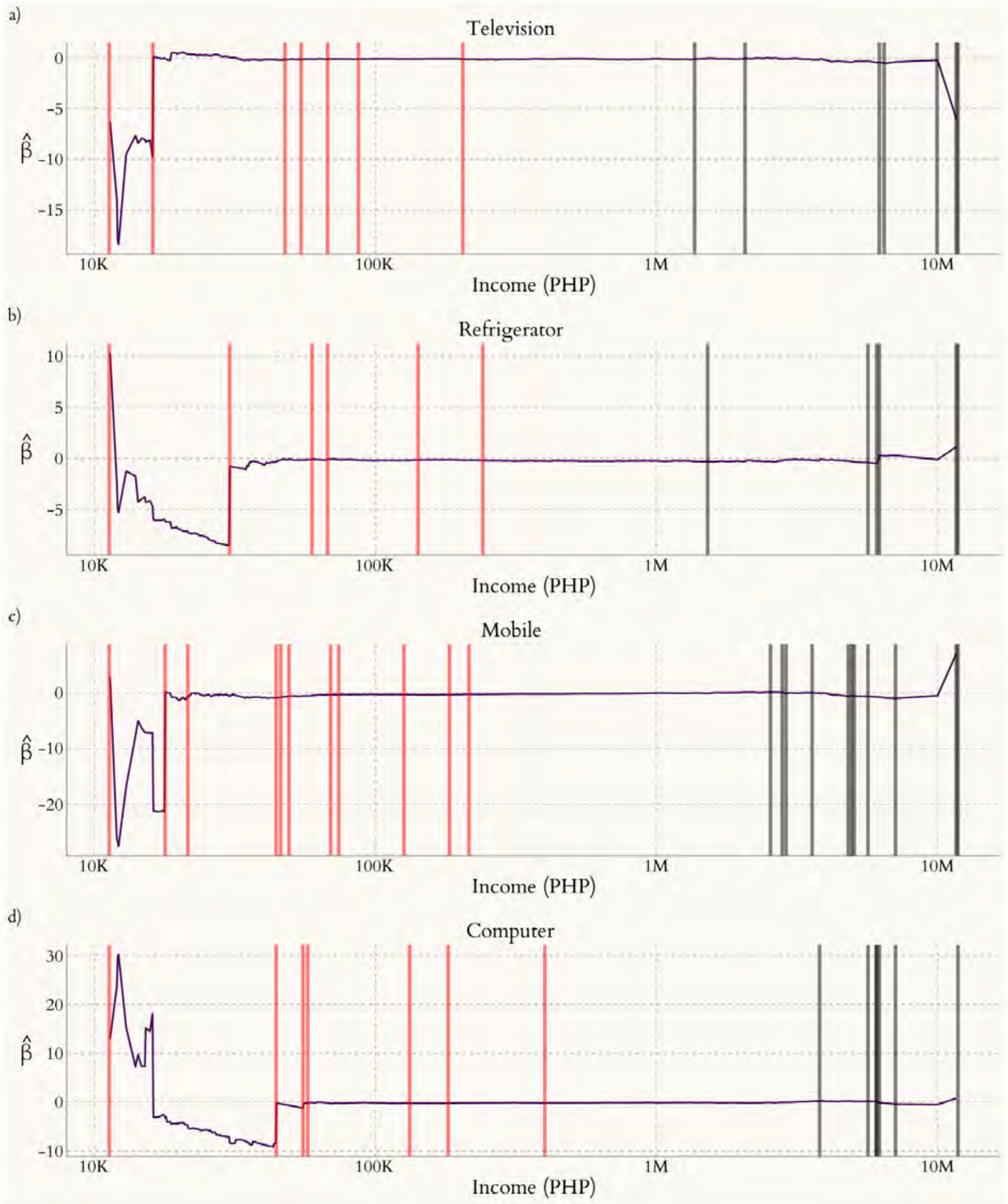


Figure 25: Poverty curves of a selection of count variables. At the second red vertical line is always a jump in the poverty curve observable. In any case, this vertical line is at the minimum income of households that own the second least amount of goods.

A table that contains the critical values of a few selected count variables used in Chapter 7.3.2.

| Variable | Min |
|---|---|
| Television | 16,137 |
| Refrigerator | 30,294 |
| Washing Machine | 36,569 |
| Mobile | 17,840 |
| Computer | 44,313 |

Table 5: Critical values at which jumps in the poverty curves of count variables occur. Note here that the count variable computer has the highest critical value.

This figure shows the SE curves of the maximum SE values from the MeanBoot BS model.

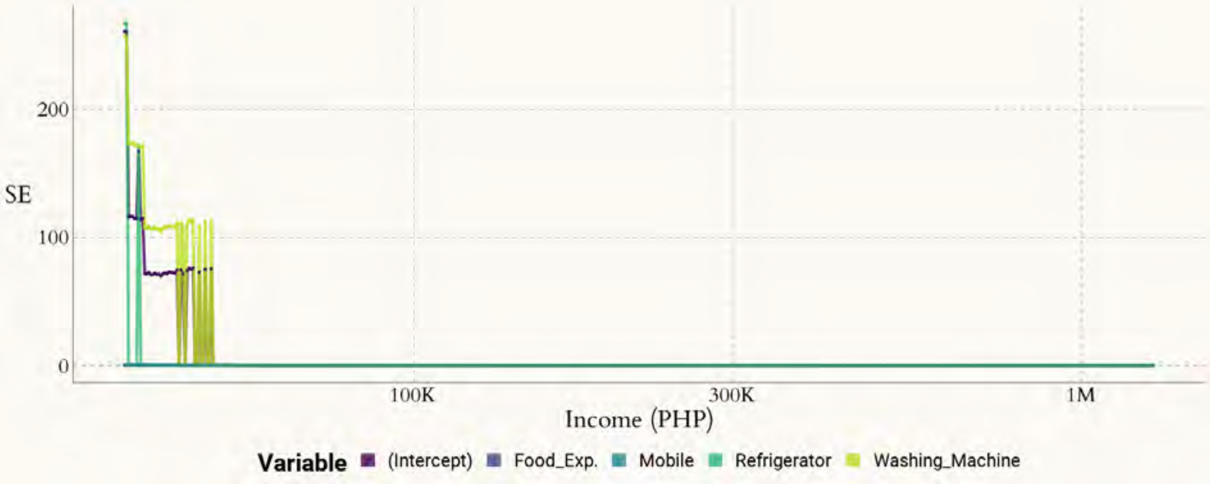


Figure 26: For boundaries close to the lower limit, the SE values of some estimated coefficients resulting from some bootstrap samples are very large. This results in large mean SE values in the MeanBoot BS model.

Prediction curves that cross the horizontal line at 0.5 below the lower or upper limit. Consequently, the poverty indices in the limited Above50 are set to the value $L$ or $U$.

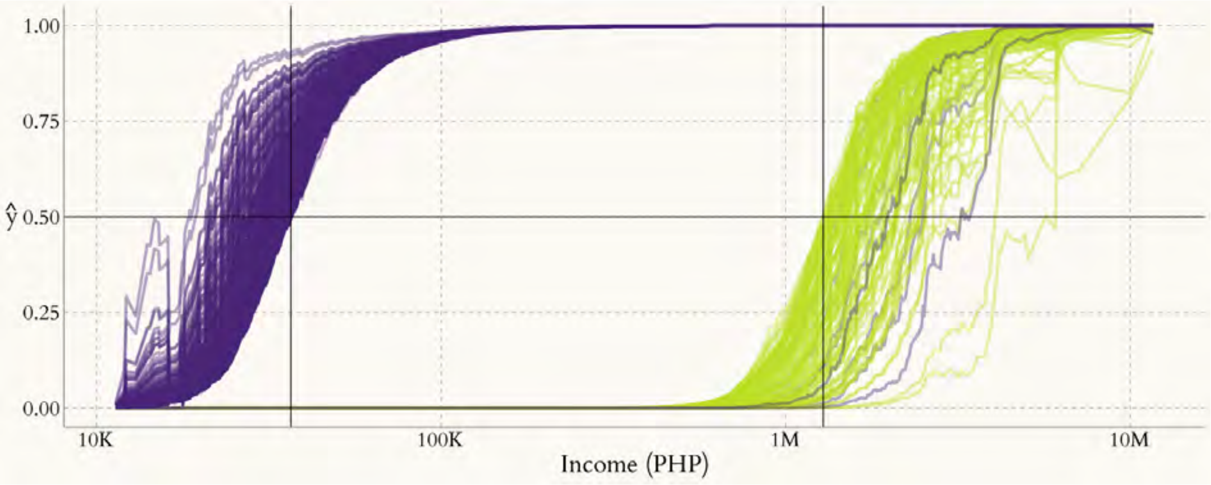


Figure 27: Horizontal black line is placed at 0.5 and the vertical black lines at $L$ and $H$. As the households' prediction curves cross the horizontal line above or below the limits, their limited Above50 poverty index is set to the lower or upper limit.

# B Electronic Appendix

The data, code and figures are in electronic form in the folder “BS_programme” provided and can be extracted from the storage medium on the last page of the thesis. This folder contains the sub-folders “data”, “figures”, “results” and “tables”. Further the files “constants”, “exploratory_data_analysis”, “main”, “main_evaluations”, “main_functions”, “penalised_BS” and “read_data”. The additional R Project file, “BS_code”, should be used to start the program. It is only briefly explained what the files contain, as comments are added to the code. Technical details about the code are not given, only the purpose of the code.

The Filipino family income and expenditure data, the figures and tables generated by the code, are the contents of the folders “data”, “figures” and “tables”. The folder “results” contains the estimated coefficients, poverty scores and related estimation of the different BS and non-BS models. These results are generated in the file “main”. For the calculation of the results in “main”, the further files “main_functions”, “constants”, “penalised_BS” and “read_data” are required. The R-file “main_functions” contains every function that has been defined in this programme. There are larger functions, i.e., those used to fit the BS models, but also smaller auxiliary functions. Functions to create the figures and calculate the statistics are defined in this file as well and, in most cases, very general. This means that BS could be applied to other data sets in the same manner if data is prepared. The only function that is not defined in the file “main_functions” is the one used to get the estimated coefficients of the Penalised BS model. Penalised BS model-related functions are defined in “penalised_BS”.

The file “constants” is an auxiliary file used to prepare the results of the logistic lasso regression to get the variables used in the BS models. The file “read_data” reads in the data set from the folder “data”. Further, all required packages are loaded, variables and their categories are renamed, and the data set is manipulated with this file.

In the file “exploratory_data_analysis”, the exploratory data analysis is performed, and the acquired figures and tables are saved in the “figures” and “tables” folder. The same applies to the file “main_evaluations” as it is used to create figures and tables from the results saved in the folder “results”. This means in the file “main_evaluations” occurs the actual data analysis. If values mentioned in the thesis are not provided in the tables or figures, this file or the file “exploratory_data_analysis” must be executed to get them. This does not take long since the models have already been calculated and saved in the “results”. Because the analysis in this script contains short descriptions and is performed in the same order as in the thesis, the desired values should be found quickly.

# References

ACC (1998). Statement of commitment for action to eradicate poverty adopted by administrative committee on coordination.

Allison, P. (2008). Convergence failures in logistic regression, *Proceedings of the SAS® Global Forum 2008 Conference*, Vol. 360, SAS Institute Inc.

Banachewicz, K., Massaron, L. and Goldbloom, A. (2022). *The Kaggle Book*, Packt Publishing Ltd.

Berenger, V. and Celestini, F. (2006). French poverty measures using fuzzy set approaches, *in* A. Lemmi and G. Betti (eds), *Fuzzy Set Approach to Multidimensional Poverty Measurement*, Springer New York, NY, pp. 139–154.

Bersales, L. (2017). 2015 family income and expenditure survey.

Betti, G., Cheli, B., Lemmi, A. and Verma, V. (2006). Multidimensional and longitudinal poverty: an integrated fuzzy approach, *in* A. Lemmi and G. Betti (eds), *Fuzzy Set Approach to Multidimensional Poverty Measurement*, Springer New York, NY, pp. 115–137.

Betti, G., D'Agostino, A. and Neri, L. (2006). Modelling fuzzy and multidimensional poverty measures in the united kingdom with variance components panel regression, *in* A. Lemmi and G. Betti (eds), *Fuzzy Set Approach to Multidimensional Poverty Measurement*, Springer New York, NY, pp. 257–275.

Bring, J. (1994). How to standardize regression coefficients, *The American Statistician* **48**(3): 209–213.

Cantoni, E. and Ronchetti, E. (2001). Robust inference for generalized linear models, *Journal of the American Statistical Association* **96**(455): 1022–1030.

Cerioli, A. and Zani, S. (1990). A fuzzy approach to the measurement of poverty, *in* C. Dagum and M. Zenga (eds), *Income and Wealth Distribution, Inequality and Poverty*, Springer Berlin Heidelberg, pp. 272–284.

Cheli, B. and Lemmi, A. (1995). A "totally" fuzzy and relative approach to the multidimensional analysis of poverty, *Economic Notes* **24**(1): 115–134.

Deutsch, J. and Silber, J. (2006). The "fuzzy set" approach to multidimensional poverty analysis: Using the shapley decomposition to analyze the determinants of poverty in israel, *in* A. Lemmi and G. Betti (eds), *Fuzzy Set Approach to Multidimensional Poverty Measurement*, Springer New York, NY, pp. 155–174.

Eskelinen, T. (2011). Relative poverty, *in* D. Chatterjee (ed.), *Encyclopedia of Global Justice*, Springer Netherlands, pp. 942–943.

European Commission (n.d.). At-risk-of poverty rate (AROP). Accessed: 2023-06-28.
**URL:** *https://ec.europa.eu/social/main.jsp?catId=818&langId=en&id=8*

Fahrmeir, L., Heumann, C., Künstler, R., Pigeot, I. and Tutz, G. (2016). *Statistik*, 8 edn, Springer Spektrum Berlin, Heidelberg.

Fahrmeir, L., Kneib, T. and Lang, S. (2009). *Regression*, 2 edn, Springer Berlin, Heidelberg.

Filippone, A., Cheli, B. and D'Agostino, A. (2001). Addressing the interpretation and the aggregation problems in totally fuzzy and relative poverty measures, *ISER Working Paper Series 2001-22*, Institute for Social and Economic Research.

Flores, F. (n.d.). Accessed: 2023-04-20.
**URL:** *https://www.kaggle.com/datasets/grosvenpaul/family-income-and-expenditure?select=Family+Income+and+Expenditure.csv/version/1*

Friedman, J., Hastie, T. and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent, *Journal of Statistical Software* **33**(1): 1—-22.

Fustier, B. (2006). The mathematical framework of fuzzy logic, *in* A. Lemmi and G. Betti (eds), *Fuzzy Set Approach to Multidimensional Poverty Measurement*, Springer New York, NY, pp. 29–47.

Goos, G. (1995). *Vorlesungen über Informatik*, Springer Berlin, Heidelberg.

Hornby, A., Deuter, M., Bradbery, J. and Turnbull, J. (2015). *Oxford Advanced Learner's Dictionary*, 9 edn, Oxford University Press, Oxford.

Huber, P. and Ronchetti, E. (2009). *Robust Statistics*, 2 edn, John Wiley & Sons, Inc., Hoboken, New Jersey.

Kim, J., Sosa, E. and Rosenkrantz, G. (2009). *A Companion to Metaphysics*, 2 edn, Blackwell Publishing Ltd.

King, G. and Zeng, L. (2001). Logistic regression in rare events data, *Political Analysis* **9**(2): 137–163.

Kosmidis, I., Schumacher, D. and Schwendinger, F. (2022). *detectseparation: detect and check for separation and infinite maximum likelihood estimates.* R package version 0.3.
**URL:** *https://CRAN.R-project.org/package=detectseparation*

Liew, V. (2008). An overview on various ways of bootstrap methods, *MPRA Paper* .

Ling, C. and Sheng, V. (2010). Class imbalance problem, *in* C. Sammut and G. Webb (eds), *Encyclopedia of Machine Learning*, Springer US, pp. 171–171.

Liu, X. (2016). *Methods and Applications of Longitudinal Data Analysis*, Elsevier Inc., chapter 3.

Lu, X. (2016). Correcting the quasi-complete separation issue in logistic regression models, *Proceedings of the SAS® Global Forum 2016 Conference*, SAS Institute Inc.

Mack, J. and Lansley, S. (1985). *Poor Britain*, George Allen & Unwin.

Mansournia, M., Geroldinger, A., Greenland, S. and Heinze, G. (2018). Separation in logistic regression: Causes, consequences, and control, *American Journal of Epidemiology* **187**(4): 864–870.

Martinetti, E. (2006). Capability approach and fuzzy set theory: Description, aggregation and inference issues, *in* A. Lemmi and G. Betti (eds), *Fuzzy Set Approach to Multidimensional Poverty Measurement*, Springer New York, NY, pp. 93–113.

Miceli, D. (2006). Multidimensional and fuzzy poverty in switzerland, *in* A. Lemmi and G. Betti (eds), *Fuzzy Set Approach to Multidimensional Poverty Measurement*, Springer New York, NY, pp. 195–209.

Panek, T. (2006). Multidimensional fuzzy relative poverty dynamic measures in poland, *in* A. Lemmi and G. Betti (eds), *Fuzzy Set Approach to Multidimensional Poverty Measurement*, Springer New York, NY, pp. 233–255.

Posit team (2023). *RStudio: integrated development environment for R*, Posit Software, PBC, Boston, MA.
**URL:** *http://www.posit.co/*

Qizilbash, M. (2006). Philosophical accounts of vagueness, fuzzy poverty measures and multidimensionality, *in* A. Lemmi and G. Betti (eds), *Fuzzy Set Approach to Multidimensional Poverty Measurement*, Springer New York, NY, pp. 9–28.

R Core Team (2022). *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria.
**URL:** *https://www.R-project.org/*

Ronchetti, E. (2010). Robust inference. Accessed: 2023-05-25.
**URL:** *http://encyclopediaofmath.org/index.php?title=Robust_inference&oldid=50961*

Sachs, J. (2005). *The End of Poverty*, Penguin Group Inc.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M. and Tarantola, S. (2008). *Global Sensitivity Analysis. The Primer*, John Wiley & Sons, Ltd.

Schaich, E. and Münnich, R. (1996). Der Fuzzy-Set-Ansatz in der Armutsmessung, *Jahrbücher für Nationalökonomie und Statistik* **215**(4): 444–469.

Silber, J. and Deutsch, J. (2005). Measuring multidimensional poverty: An empirical comparison of various approaches, *Review of Income and Wealth* **51**(1): 145–174.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society. Series B (Methodological)* **58**(1): 267–288.

UNDP (2022). Human development report 2021-22, *Technical report*, United Nations Development Programme.

UNDP (n.d.). Human development index (HDI). Accessed: 2023-06-28.
**URL:** *https://hdr.undp.org/data-center/human-development-index#/indicies/HDI*

UNDP and OPHI (2022). 2022 global multidimensional poverty index (mpi): Unpacking deprivation bundles to reduce, *UNDP (United Nations Development Programme)* .

Vero, J. (2006). A comparison of poverty according to primary goods, capabilities and outcomes. evidence from frencli school leavers' surveys, *in* A. Lemmi and G. Betti (eds), *Fuzzy Set Approach to Multidimensional Poverty Measurement*, Springer New York, NY, pp. 211–231.

Walker, H. and Lev, J. (1969). *Elementary Statistical Methods*, 3 edn, New York, Holt, Rinehart and Winston.

Weisstein, E. (n.d.). Singular matrix. Accessed: 2023-05-26.
**URL:** *https://mathworld.wolfram.com/SingularMatrix.html*

World Bank (2022a). Measuring poverty. Accessed: 2023-06-28.
**URL:** *https://www.worldbank.org/en/topic/measuringpoverty#1*

World Bank (2022b). *Poverty and Shared Prosperity 2022*, World Bank.

Zimmermann, H.-J. (2001). *Fuzzy Set Theory-and Its Applications*, 4 edn, Springer Dordrecht.

# Declaration of authorship

I hereby declare that the report submitted is my own unaided work. All direct or indirect sources used are acknowledged as references. I am aware that the Thesis in digital form can be examined for the use of unauthorized aid and in order to determine whether the report as a whole or parts incorporated in it may be deemed as plagiarism. For the comparison of my work with existing sources I agree that it shall be entered in a database where it shall also remain after examination, to enable comparison with future Theses submitted. Further rights of reproduction and usage, however, are not granted here. This paper was not previously presented to another examination board and has not been published.

Location, date

Name