Thesis submitted in fulfilment of the requirements for the degree

Dr. rer. pol.

on the topic

A Selection of Statistical Methods for Interval-Censored Data with Applications to the German Microcensus

to the

Chair of Applied Statistics
School of Business and Economics
Freie Universität Berlin

submitted by

Paul Walter

Berlin, 2018

Paul Walter, *A Selection of Statistical Methods for Interval-Censored Data with Applications to the German Microcensus,*
October 2018

Supervisors:
Prof. Dr. Timo Schmid (Freie Universität Berlin)
Prof. Nikos Tzavidis, Ph.D. (University of Southampton)

Location:
Berlin

Date of defense:
December 19, 2018

Publication List

The publications listed below are the result of the research carried out in this thesis titled, "A Selection of Statistical Methods for Interval-Censored Data with Applications to the German Microcensus."

1. Walter, P., Groß, M., Schmid, T., and Tzavidis, N., (2018). **Estimating Linear Mixed Regression Models with an Interval-Censored Dependent Variable using a Stochastic Expectation-Maximization Algorithm applied to German Microcensus Data.** *Working paper*, to be submitted.

2. Walter, P. and Weimer, K., (2018). **Estimating Poverty and Inequality Indicators using Interval-Censored Income Data from the German Microcensus.** *Working paper*, to be submitted.

3. Walter, P., Groß, M., Schmid, T., and Tzavidis, N., (2018). **Small Area Estimation with Interval-Censored Income Data.** *Working paper*, to be submitted.

4. Walter, P. (2018). **The R Package smicd: Statistical Methods for Interval-Censored Data.** *R package vignette*, to be submitted.

Contents

Introduction

In its Global Risks Report 2017, the World Economic Forum identifies rising income and wealth disparity as one of the top five global development trends that potentially causes unemployment, underemployment, and profound social instability (World Economic Forum, 2017). To counteract this development, it is necessary to quantify inequality and to analyze the distribution of income and wealth. In order to measure inequality and identify factors that significantly impact income or wealth, governments and statistical offices collect data by conducting surveys and censuses. However, collecting data on rather private topics such as income can lead to high item non-response rates. Therefore, it is tempting for survey designers to collect information on income using income bands as opposed to detailed income (Micklewright and Schnepf, 2010). This kind of data is commonly known as interval-censored data, grouped data or banded data. It is defined as observing only the lower and upper bound of an income variable with its exact value remaining unknown. Collecting only the interval information instead of continuous data offers a higher degree of data privacy protection to survey respondents, which lowers response burdens and thus leads to lower item non-response rates and higher data quality. This kind of data is already being collected by a number of surveys and censuses. Among them is the biggest annually survey in Europe, the German Microcensus (Statistisches Bundesamt, 2018a), and the censuses of Australia (Australian Bureau of Statistics, 2011), Colombia (Departamento Administrativo Nacional De Estadística, 2005), and New Zealand (Statistics New Zealand, 2013).

While data quality is increased, analyzing interval-censored data requires more advanced statistical methods. This is due to the fact that only the interval information is observed and the underlying data distribution within each interval remains unobserved. For instance, well-established and widely used statistical methods, such as linear and linear mixed regression, require a continuous response variable. Furthermore, formulas to estimate statistical indicators, such as the mean, rely on metric data. While regression models are commonly applied to analyze income and wage, the estimation of statistical indicators from interval-censored data is of particular interest for the Federal Statistical Office and the Statistical Offices of the German States in order to measure and monitor the regional distribution of poverty and inequality (Stauder and Hüning, 2004). This work therefore proposes new statistical methodology for the estimation of linear and linear mixed regression models with an interval-censored response variable and for the estimation of statistical indicators from interval-censored data, e.g., German Microcensus data.

In Part I of the thesis, theory is developed to infer the properties of a population with linear and linear mixed models using sample data. In particular, in Chapter 1, theory is proposed to estimate the regression parameter and its standard errors of linear and linear mixed models with an interval-censored response variable. For the estimation of the parameters, a novel stochastic expectation-maximization (SEM) algorithm is proposed. In order to estimate the standard errors of the regression parameters, two different bootstraps are introduced. A non-parametric bootstrap for the linear regression model and a parametric bootstrap for the linear mixed regression model. Both the introduced bootstraps account for the additional uncertainty that is caused by the interval censoring of the dependent variable. The theory is applied to analyze interval-censored personal income data collected by the German Microcensus with a linear mixed regression model. By applying the newly proposed methodology, different components that significantly affect income are discovered.

In Part II, new methodology is proposed for the direct estimation (without covariates) and the prediction of statistical indicators, for instance, poverty and inequality indicators. For the direct estimation of statistical indicators, an iterative kernel density algorithm is proposed in Chapter 2. The proposed algorithm generates metric pseudo samples from the interval-censored target variable. From these pseudo samples, any statistical indicator of interest can be estimated. The estimation of the standard errors is facilitated by a non-parametric bootstrap that accounts for the additional uncertainty coming from the interval censoring. The method is applied to estimate poverty and inequality indicators at the federal state level from interval-censored household income data collected by the German Microcensus. For valid indicator estimates, survey and household equivalence scale weights are incorporated into the algorithm and used in the analysis.

When samples sizes are small, e.g., in small geographic areas, direct estimators of statistical indicators might be unreliable. Furthermore, some areas of interest might not even be sampled. In these situations, small area estimation (SAE) methods can provide reliable estimates for the desired indicators (Rao and Molina, 2015). One particular SAE method that has been used in this context is the empirical best predictor (EBP) method (Molina and Rao, 2010). This method is based on the use of a linear mixed regression model estimated with income as a response variable that is measured on a continuous scale. To enable the use of the EBP method with an interval-censored response variable, the SEM algorithm proposed in Chapter 1 is applied to estimate the model parameters in Chapter 3. The EBP method crucially depends on the normality assumption of the residuals. Therefore, the SEM algorithm is further developed to facilitate the use of the data-driven Box-Cox transformation (Box and Cox, 1964). The estimation of the mean squared error of the EBPs is facilitated by a parametric bootstrap that accounts for the additional variability coming from the uncertainty from estimating the transformation parameter of the Box-Cox transformation and the uncertainty resulting from working with limited information due to interval censoring. The newly introduced SEM algorithm in conjunction with transformations and the modified EBP approach is then used to estimate disaggregated poverty and inequality indicators from interval-censored income data in Chiapas, one of the poorest states in Mexico.

In Part III, the implementation of the proposed theory in the programming language R is presented (R Core Team, 2018). Implementing new methodology is valuable in order to enable other researchers, data analysts, and practitioners to easily use the newly introduced statistical theory. Therefore, the theory is implemented in the R package `smicd` available on the Comprehensive R Archive Network. In Chapter 4, the package, its functionality, and its usage is presented in detail.

Part I

Linear and Linear Mixed Regression Models with an Interval-Censored Dependent Variable

Chapter 1

Estimating Linear Mixed Regression Models with an Interval-Censored Dependent Variable using a Stochastic Expectation-Maximization Algorithm applied to German Microcensus Data

1.1 Introduction

In statistics, linear and linear mixed regression analysis are approaches for modeling the linear relationship between a dependent variable $\mathbf{y}$ and explanatory variables (or independent variables) $\mathbf{X}$. While linear regression models only contain fixed effects, linear mixed models (also called hierarchical linear models or multilevel models) extend linear regression theory by containing fixed effects and random effects, see for example Goldstein (2003) or Snijders and Bosker (2011). These kinds of models are applied when the data is not independent – which is a crucial assumption of linear regression models – but clustered. Linear mixed models allow for all kinds of clustered data that occurs when measurements are made of related statistical units (e.g., students within schools or people of different nationalities) or when repeated measurements are made of the same statistical unit (longitudinal data). Since clustered data is common in many disciplines, linear mixed models are widespread not only in the field of econometrics and the social science, but also in physics, biology or medicine. Parameter estimates of these models are commonly obtained by maximum likelihood (ML) or residual (restricted) maximum likelihood theory (REML) (Lindstrom and Bates, 1990). However, when the dependent variable is not measured on a continuous scale, but rather censored to specific intervals (also known as grouped or banded data), standard ML or REML theory cannot be applied without adjusting for the unobserved data distribution within each interval.

In the field of econometrics and in the social sciences, data is frequently collected as

interval-censored data because of confidentiality constraints or to avoid item non-response and thus increase data quality (Micklewright and Schnepf, 2010). Item non-response can be avoided because interval-censored data offers a higher level of data privacy protection that is of particular concern if sensitive questions (e.g., about income) are asked in a survey or census (Moore and Welniak, 2000; Hagenaars and Vos, 1988). Therefore, many surveys and censuses ask for interval-censored data, for example, the German Microcensus (Statistisches Bundesamt, 2017), the Australian census (Australian Bureau of Statistics, 2011), the Colombian census (Departamento Administrativo Nacional De Estadística, 2005) and the census from New Zealand (Statistics New Zealand, 2013). Asking for interval-censored data lowers the amount of item non-response, but leads to less informative data because the distribution of the data within each interval remains unobserved.

For dealing with interval-censored dependent variables in the linear regression context, various approaches and statistical methods are described in the literature. A naive approach is ordinary least squares regression (OLS) on the midpoints of the intervals (Thompson and Nelson, 2003). However, this approach leads to biased parameter estimates because the unobserved distribution of the sample data within each interval is neglected in the estimation (Cameron, 1987). The performance of this approach depends heavily on the number of intervals. As the number of intervals increases to infinity the bias vanishes (Fryer and Pethybridge, 1972). Another approach is to conceptualize the model as an ordered logit- or probit regression (McCullagh, 1980). However, this means switching to models with different link functions, which alters the interpretation of the coefficients. In order to stick to the linear modeling framework and to overcome the drawbacks of OLS regression on the midpoints, there exists methodology for left-censored (Tobin, 1958), right-censored (or both) (Rosett and Nelson, 1975) and grouped (or interval-censored) (Stewart, 1983) dependent variables. Stewart (1983) describes an algorithm for attaining the maximum likelihood solution when the dependent variable of a linear model is interval censored. The algorithm can be seen as a special case of the expectation-maximization (EM) algorithm introduced by Dempster et al. (1977). Therein, monotonic convergence is guaranteed (Burridge, 1981).

For linear mixed models with an interval-censored dependent variable, OLS regression on the midpoints and conceptualizing the model as an ordered logit- or probit regression is feasible as well. However, to avoid biased estimation results (OLS regression on the midpoints) or switching to models with different link functions (logit- or probit regression), we propose a stochastic expectation-maximization (SEM) algorithm for the parameter estimation in linear mixed models based on Celeux and Dieboldt (1985) and Celeux et al. (1996). In the SEM algorithm, the analytical expectation step from the EM algorithm is replaced by the drawing of pseudo samples. The algorithm can also be applied for parameter estimations in linear models without random parameters.

The linear (mixed) regression model relies on certain model assumptions, e.g., normality of the residuals. These assumptions are also valid when the model is estimated with an interval-censored instead of a continuous dependent variable. When dealing with departures from the model assumptions, the proposed algorithm allows for the use of transformations on

the dependent variable.

For the estimation of the standard errors of the fixed effects, a parametric bootstrap is proposed that accounts for the additional uncertainty that comes from the interval-censored dependent variable. The validity of the proposed methodology is demonstrated via several model-based simulations.

To show the strength and flexibility of the SEM algorithm, it is applied to model income in Germany using Microcensus data from 2012. The German Microcensus is a representative sample of the German population with a sample size of about 380,000 households and 820,000 household members that is carried out annually as a replacement for a full census. Since income, the dependent variable in the working model, is interval censored the SEM algorithm is applied to estimate the parameters of the linear mixed model. With this example, we substantiate that asking for or providing only interval-censored data does not impair valid inference when working with linear mixed models with an interval-censored response variable.

The paper is structured as follows. In Section 1.2, the German Microcensus data set that is used in the application is presented. In Section 1.3, the SEM algorithm and the parametric bootstrap is introduced. In Section 1.4, the method is empirically evaluated by several model-based simulation studies. In Section 1.5, interval-censored income data is modeled with linear mixed models based on data from the German Microcensus. And finally, in Section 1.6, the main results are summarized and further research directions are presented.

1.2 The German Microcensus

The derivation of new statistical methodology for linear mixed regression models is motivated by the German Microcensus data set that contains interval-censored income data. The German Microcensus is a survey that is conducted annually by the German Federal Statistical Office (Statistisches Bundesamt, 2018a). The survey has a long history and was first carried out in 1957 (Statistical Offices of the Federation and the Federal States, 2016). The total sample size is equal to 1% of the German population. This amounts to about 380,000 households and 820,000 household members. It is the largest annually conducted household survey in Europe (Statistisches Bundesamt, 2018b). The large sample size is required to estimate statistical measures with high accuracy for small subdomains. Hence, for the analysis of small subdomains, the Microcensus is superior to other, smaller, surveys (Boehle, 2015). The aim of the Microcensus is to provide data on a regular short-term basis. Topics covered by the survey are: demographic background, migration, employment, education, and vocational training (Schwarz, 2001). For most questions, answering is compulsory by law, however, there are also questions that are answered on a voluntary bases, such as questions about health status, health insurance, housing situation, and retirement programs (Statistical Offices of the Federation and the Federal States, 2016). The results are published in several governmental reports, like the annual report of the German Council of Economic Experts, the employment report, and the Federal Governments Annual Pension and Insurance report. Furthermore, the data is used to estimate EU indicators on employment policy (Statistisches Bundesamt, 2018b).

However, the data is highly valuable not only for governmental institutions but also for researchers from various fields, e.g., econometrics or the social sciences. Researchers appreciate the Microcensus data set for very low non-response rates and high data quality (Schwarz, 2001). While low non-response rates are guaranteed by mandatory responses for most questions, high data quality is achieved with face-to-face interviews. Although the Microcensus is valued by many researchers, analyzing the data properly is problematic when the research focuses on income. This is due to the fact that both personal and household income are only observed as an interval-censored variable. Furthermore, the censoring scheme and the number of intervals has changed over time, which makes the longitudinal analyses even more complicated (Boehle, 2015). Some researchers even say that because of the interval censoring of the income variable, the Microcensus is unsuitable for valid research on the topic of income (Stauder and Hüning, 2004).

To overcome these problems, we propose an SEM algorithm that enables the parameter estimation in linear and linear mixed regression models with an interval-censored response variable, independently of the censoring scheme. We demonstrate the applicability of the SEM algorithm in Section 1.5. As a demonstration data set the scientific use file (SUF) of the German Microcensus from 2012 is used (Statistisches Bundesamt, 2017). The SUF is a 70% sample of the German Microcensus. In contrast to the original Microcensus data set, some variables are aggregated to assure anonymity for small subgroups. For example, regional information is only available on a higher geographical level (federal state level). Also nationalities with less than 50,000 residents in Germany are aggregated to groups of nationalities (e.g., Belgium and Luxembourg form a group).

In the application, interval-censored personal income is modeled using a linear mixed regression model. The distribution of the interval-censored income variable is given in Figure 1.1 and in Table 1.5 in Appendix 1.7. It can be seen that the interval widths differ. The lower intervals are very narrow, while the interval width increases with higher income. For instance, the first interval is $(1, 150]$ and the penultimate is $(10000, 180000]$. The last interval is omitted in the plot because its upper bound is $+\infty$.

In the analysis, we do not aim for a perfect income equation from an economical standpoint. The focus of this paper is rather on the introduction of new statistical theory and the application serves as a motivation for its development. Nevertheless, the selection of explanatory variables is taken seriously and is based on relevant literature on the subject of modeling income and wage (as a component of income). The Mincer equation is the classical wage equation in the field of economics. In the Mincer equation, wage is a function of education and experience (Mincer, 1958, 1974). In its standard version, log wage is modeled and experience is included as a quadratic term in the equation in order to control for its decreasing marginal effect (Heckman et al., 2003; Lemieux, 2006). Many authors have extended the classical Mincer equation through variables such as region, sex, job, and age (Vijverberg, 1986; Charlotte and Steiner, 1999; Bell et al., 2002; Corrado, 2007). There are also studies on the immigration wage gap in Germany that point out the need to include identifiers for nationalities (Aldashev et al., 2008). Based on these studies, the variables education, job, sex, age, region, and nationality described

in Table 1.1 are included in the working model. The variable experience from the classical Mincer equation is not included in the working model since it is not collected by the German Microcensus.

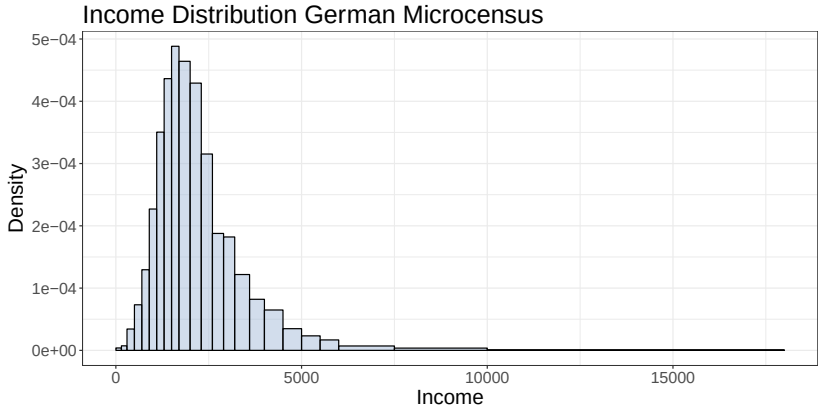


Figure 1.1: Interval-censored personal net income distribution of the German Microcensus. Since the upper bound of the upper interval is $+\infty$ it is omitted for visualization purposes.

Table 1.1: Variables from the SUF of the German Microcensus used in the application.

Variable name	Description
Dependent variable:	
Income	Interval-censored personal net income from last month
Independent variables:	
Education	Level of education measured on the International Standard Classification of Education (ISCED) scale from 1997
Job	Job status measured on the international Standard Classification of Occupations (ISCO-08) scale
Sex	Sex (male or female)
Age	Age
Region	East Germany, West Germany, Berlin
Random intercept:	
Nationality	First foreign nationality

Nationality is included as random intercept v_j to control for the within-cluster correlation in the data. We have decided against including it as a fixed parameter because we are not interested in interpreting its effect. Furthermore, age is included as a squared term to control for its decreasing marginal effect. The linear mixed model is given by

$$\begin{aligned} log(\text{Income}_{ij}) = \beta_0 + \beta_1 \text{Education}_{ij} + \beta_2 \text{Job}_{ij} + \beta_3 \text{Sex}_{ij} \\ + \beta_4 \text{Age}_{ij} + \beta_5 \text{Age}_{ij}^2 + \beta_6 \text{Region}_{ij} + v_j + e_{ij}, \end{aligned}$$

for $i = 1, \ldots, n_j$ and $j = 1, \ldots, D$, where n_j is the number of people with nationality j and D is equal to the number of different nationalities in the data set. After the exclusion of unemployed people (no income from work), observations with missing values, and nationalities

which cannot be uniquely identified, the number of observations equals $n_{total} = \sum_{j=1}^{29} n_j = 311659$ from 29 different identifiable nationalities in the sample. Some nationalities are not uniquely identifiable due to the mentioned anonymity constraints of the SUF.

Detailed descriptive statistics for the explanatory variables are given in Table 1.11, 1.12, 1.13, 1.14 and Table 1.15 in Appendix 1.7. The educational level ISCED 3b (apprenticeship or vocational qualifying degree at a full-time vocational school, annual school of health care) is attained by the largest amount of people in the sample (45.4%), while the highest level of education ISCED 6 (doctorate) is only attained by 2.2% of the people in the sample. The variable job is measured on the ISCO-08 scale. The modus of the variable is Technicians and Associate Professionals with 25.8% of the respondents working in that field. In the analysis, it is expected that a higher level of education has a positive effect on income. Furthermore, people that work in a job with a higher status, e.g., managers, are expected to earn more. It is notable that 74.8% of the sample are male, while only 25.2% are female. This is due to the fact that personal income is only measured for the head of the household, who is mostly male. We expect higher income for males than for females (gender pay gap). The minimum age in the data set is 16, the maximum is 93, and the median 45. In the sample data, 80.8% live in the West, 15.4% in the East, and 3.8% in Berlin. The East is defined as the federal states of the former German Democratic Republic. All other states belong to the category West, except the federal state Berlin. Since Berlin is the only federal state that was divided between East and West, an extra category is defined for it. Since the East includes the states from the former German Democratic Republic, incomes are expected to be lower than in the West. Also, descriptive statistics for nationality, the variable specified as random intercept, are given in Table 1.16 in Appendix 1.7. Not surprisingly, the nationality of most people is German (91.5%).

The estimation of the model parameters with the SEM algorithm and the interpretation of the results is conducted in Section 1.5.

1.3 Methodology

In the next two sections, the new methodology is introduced. In Section 1.3.1, we propose an SEM algorithm for parameter estimations in linear and linear mixed models with an interval-censored response variable. In Section 1.3.2, we present a parametric bootstrap for the estimation of the standard errors of the fixed effects that accounts for the interval censoring of the dependent variable.

1.3.1 Parameter estimation

The linear mixed model is a generalization of the linear model that allows for additional random deviations (effects) besides the random error term. In the classical linear mixed model, the dependent variable is observed on a continuous scale. Following Laird and Ware (1983), the

model can be written in matrix notation as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v} + \mathbf{e}, \tag{1.1}$$

where $\mathbf{y}$ is a $N \times 1$ column vector of the dependent variable, N is the sample size, $\mathbf{X}$ is a $N \times p$ matrix where p is equal to the number of predictors, $\boldsymbol{\beta}$ is a column vector of the fixed effects regression parameters of size $p \times 1$, $\mathbf{Z}$ is the $N \times q$ design matrix with q random effects, $\mathbf{v}$ is a $q \times 1$ vector of random effects and $\mathbf{e}$ is the residual vector of size $N \times 1$. The distribution of the random effects is given by

$$\mathbf{v} \sim N(\mathbf{0}, \mathbf{G}), \qquad \text{where } \mathbf{G} = \begin{bmatrix} \sigma_0^2 & \sigma_{01} & \dots & \sigma_{0q} \\ \sigma_{10} & \sigma_1^2 & \dots & \sigma_{1q} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{q0} & \sigma_{q1} & \dots & \sigma_q^2 \end{bmatrix},$$

and the distribution of the residuals is given by $\mathbf{e} \sim N(\mathbf{0}, \mathbf{R})$ with $\mathbf{R} = \mathbf{I}_N \sigma_e^2$ where $\mathbf{I}_N$ is the identity matrix and σ_e^2 is the residual variance. Furthermore, the random effects $\mathbf{v}$ and the residuals $\mathbf{e}$ are assumed to be independent. The fixed linear predictor, as in standard OLS regressions, is given by $\mathbf{X}\boldsymbol{\beta}$. The random part of the model is given by $\mathbf{Z}\mathbf{v} + \mathbf{e}$. The vector of random effects $\mathbf{v}$ is not directly estimated (it can be predicted). Instead, the variance components of $\mathbf{G}$ as well as the residual variance σ_e^2 are estimated. The design matrices $\mathbf{X}$ and $\mathbf{Z}$ enable the flexible modeling of a variety of linear models, e.g., block designs or different hierarchical designs. In order to control for correlation within clusters, random intercepts and/or random slopes can be included into the model, e.g., to control for students within the same schools or for people from different nationalities. The variance covariance matrix $\mathbf{G}$ makes it possible to incorporate different correlation structures between the random slope and intercept. Furthermore, through a general formulation of $\mathbf{R}$, it can be controlled for heteroscedastic and correlated residuals. For a more detailed introduction of mixed models, see Searle et al. (1992); Verbeke and Molenberghs (2000); Pinheiro and Bates (2000); Raudenbush and Bryk (2002); Demidenko (2004); McCulloch et al. (2008); Snijders and Bosker (2011). When the dependent variable is measured on a continuous scale, estimation is usually facilitated by ML or REML methods (Lindstrom and Bates, 1990). However, when the dependent variable is interval censored, standard estimation methods cannot be applied without adjustment.

Consider that the only observed information concerning the dependent variable $\mathbf{y}$ is that it falls into a certain interval on a continuous scale. Thus, the continuous scale is divided into n_k intervals of varying width, where the k-th interval is given by (A_{k-1}, A_k), with A_{k-1} being the lower and A_k the upper bound of each interval. The variable k $(1 \leq k \leq n_k)$ indicates into which of the n_k intervals an observation falls, with $\mathbf{k}$ being a $N \times 1$ column vector $\mathbf{k} = (k_1, k_2, \dots, k_N)^T$. Depending on the application, the outer intervals might be open-ended, i.e., $A_0 = -\infty$ and $A_{n_k} = +\infty$ are possible. This kind of censoring leads to a loss of information, because the distribution of the sample data within each interval remains unobserved. Therefore, we start by reconstructing the distribution of the unobserved $\mathbf{y}$, using

the observed group identifier $\mathbf{k}$ and the linear relationship stated in Model (1.1). To reconstruct the unknown distribution $f(\mathbf{y}|\mathbf{X}, \mathbf{Z}, \mathbf{v}, \mathbf{k}, \boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\boldsymbol{\beta}, \mathbf{R}, \mathbf{G})$, the Bayes theorem (Bayes, 1763) is applied. It follows that

$$\begin{aligned} f(\mathbf{y}|\mathbf{X}, \mathbf{Z}, \mathbf{v}, \mathbf{k}, \boldsymbol{\theta}) &= \frac{f(\mathbf{k}|\mathbf{y}, \mathbf{X}, \mathbf{Z}, \mathbf{v}, \boldsymbol{\theta}) f(\mathbf{y}|\mathbf{X}, \mathbf{Z}, \mathbf{v}, \boldsymbol{\theta})}{f(\mathbf{k}|\mathbf{X}, \mathbf{Z}, \mathbf{v}, \boldsymbol{\theta})} \\ &\propto f(\mathbf{k}|\mathbf{y}, \mathbf{X}, \mathbf{Z}, \mathbf{v}, \boldsymbol{\theta}) f(\mathbf{y}|\mathbf{X}, \mathbf{Z}, \mathbf{v}, \boldsymbol{\theta}). \end{aligned}$$

Additionally $f(\mathbf{k}|\mathbf{y}, \mathbf{X}, \mathbf{Z}, \mathbf{v}, \boldsymbol{\theta}) = f(\mathbf{k}|\mathbf{y})$ because the conditional distribution of $\mathbf{k}$ only depends on $\mathbf{y}$. It is given by $f(\mathbf{k}|\mathbf{y}) = \mathbf{r}$ with $\mathbf{r}$ being a $N \times 1$ column vector $\mathbf{r} = (r_1, r_2, \ldots, r_N)^T$ with

$$r_i = \begin{cases} 1 & \text{if } A_{k-1} \leq y_i \leq A_k, \\ 0 & \text{else,} \end{cases}$$

for $i = 1, \ldots, N$. The conditional distribution of $\mathbf{y}$ equals

$$f(\mathbf{y}|\mathbf{X}, \mathbf{Z}, \mathbf{v}, \boldsymbol{\theta}) \sim \quad N(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v}, \mathbf{R}).$$

This follows directly from the priorly-stated assumptions of the linear mixed model. The unknown model parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}, \mathbf{R}, \mathbf{G})$ are estimated using pseudo samples $\tilde{\mathbf{y}}$ (maximization-step) of the unknown $\mathbf{y}$ that are iteratively drawn from the conditional distribution $f(\mathbf{y}|\mathbf{X}, \mathbf{Z}, \mathbf{v}, \mathbf{k}, \boldsymbol{\theta})$ (stochastic-step). The computational details of the SEM algorithm are given below.

Computational details

To fit Model (1.1) with an interval-censored response, the parameter vector $\hat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\beta}}, \widehat{\mathbf{R}}, \widehat{\mathbf{G}})$ is estimated and pseudo samples of $\mathbf{y}$ are iteratively drawn by the following algorithm. The pseudo samples are drawn from the following conditional distribution:

$$f(\mathbf{y}|\mathbf{X}, \mathbf{Z}, \mathbf{v}, \mathbf{k}, \boldsymbol{\theta}) \propto \mathbf{I}(A_{\mathbf{k}-1} \leq \mathbf{y} \leq A_{\mathbf{k}}) \times N(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v}, \mathbf{R}),$$

where $\mathbf{I}(\cdot)$ denotes the indicator function. So, for observations with explanatory variables $\mathbf{X}$, the corresponding $\tilde{\mathbf{y}}$ is randomly drawn from $N(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v}, \mathbf{R})$ conditional on the given interval $(A_{\mathbf{k}-1} \leq \mathbf{y} \leq A_{\mathbf{k}})$. The conditional distribution $f(\mathbf{y}|\mathbf{X}, \mathbf{Z}, \mathbf{v}, \mathbf{k}, \boldsymbol{\theta})$ has the form of a two-sided truncated normal distribution. Once $\hat{\boldsymbol{\theta}}$ is estimated its probability density function is given by

$$\hat{f}(\mathbf{y}|\mathbf{X}, \mathbf{Z}, \hat{\mathbf{v}}, \mathbf{k}, \hat{\boldsymbol{\theta}}) = \frac{\phi\left(\frac{\mathbf{y} - \hat{\boldsymbol{\mu}}}{\hat{\sigma}_e}\right)}{\hat{\sigma}_e\left(\Phi\left(\frac{A_{\mathbf{k}} - \hat{\boldsymbol{\mu}}}{\hat{\sigma}_e}\right) - \Phi\left(\frac{A_{\mathbf{k}-1} - \hat{\boldsymbol{\mu}}}{\hat{\sigma}_e}\right)\right)},$$

with $\hat{\boldsymbol{\mu}} = \mathbf{X}\widehat{\boldsymbol{\beta}} + \mathbf{Z}\hat{\mathbf{v}}$, where $\phi(\cdot)$ is the probability density function of the standard normal distribution and $\Phi(\cdot)$ is its cumulative distribution function. By definition, $\Phi\left(\frac{A_{\mathbf{k}} - \hat{\boldsymbol{\mu}}}{\hat{\sigma}_e}\right) = 1$ if $A_{\mathbf{k}} = +\infty$ and $\Phi\left(\frac{A_{\mathbf{k}-1} - \hat{\boldsymbol{\mu}}}{\hat{\sigma}_e}\right) = 0$ if $A_{\mathbf{k}-1} = -\infty$. The explicit steps of the algorithm are given

by:

1. Estimate $\hat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\beta}}, \widehat{\mathbf{R}}, \widehat{\mathbf{G}})$ of Model (1.1) using the midpoints of the intervals as a substitute for the unknown $\mathbf{y}$. The joint density of $\mathbf{y}$ and $\mathbf{v}$ is equal to $f(\mathbf{y}, \mathbf{v}) = f(\mathbf{y}|\mathbf{v})f(\mathbf{v})$. The parameters are estimated by maximizing the joint density with respect to $\boldsymbol{\beta}$ and $\mathbf{v}$ simultaneously using REML (Thompson, 1962).

2. **Stochastic step:** For $i = 1, \ldots, N$, sample from the conditional distribution $f(\mathbf{y}|\mathbf{X}, \mathbf{Z}, \mathbf{v}, \mathbf{k}, \boldsymbol{\theta})$ by drawing randomly from $N(\mathbf{X}\widehat{\boldsymbol{\beta}} + \mathbf{Z}\hat{\mathbf{v}}, \widehat{\mathbf{R}})$ within the given interval $(A_{\mathbf{k}-1} \leq \mathbf{y} \leq A_{\mathbf{k}})$ (a two-sided truncated normal distribution) obtaining $(\tilde{\mathbf{y}}, \mathbf{X}, \mathbf{Z})$. The drawn pseudo $\tilde{\mathbf{y}}$ are used as a replacement for the unobserved $\mathbf{y}$.

3. **Maximization step:** Re-estimate the vector $\hat{\boldsymbol{\theta}}$ of Model (1.1) using the pseudo samples $(\tilde{\mathbf{y}}, \mathbf{X}, \mathbf{Z})$ obtained in Step 2. Parameter estimation is carried out by REML, as in Step 1.

4. Iterate Steps 2-3 $B^{(SEM)} + M^{(SEM)}$ times, with $B^{(SEM)}$ burn-in iterations and $M^{(SEM)}$ additional iterations.

5. Discard the burn-in iterations and estimate $\hat{\boldsymbol{\theta}}$ by averaging the obtained $M^{(SEM)}$ estimates.

In the presence of open-ended intervals $A_0 = -\infty$ and $A_{n_k} = +\infty$, the midpoints M_1 and M_{n_k} for the open-ended intervals in iteration Step 1 are computed as follows:

$$M_1 = (A_1 - \overline{D})/2,$$
$$M_{n_k} = (A_{n_k-1} + \overline{D})/2,$$

where,

$$\overline{D} = \frac{1}{(n_k - 2)} \sum_{k=2}^{n_k-1} |A_{k-1} - A_k|.$$

Simulation results show that the choice of the midpoints for the open-ended intervals, has no impact on the estimation results. This is due to the fact that these midpoints only serve as a proxy for the first iteration step. In the case of open-ended intervals it is drawn from a one-sided truncated normal distribution after the first iteration step.

As for linear models, the normality assumption of the residuals is also crucial for linear mixed models in order to obtain estimation results that allow for valid inference. This holds true for models with a continuous dependent variable as well as for models with an interval-censored dependent variable. To obtain normality, a logarithmic transformation is commonly applied to the dependent variable $\log(\mathbf{y})$ in the linear regression context, especially for wage or income equations. When the dependent variable is interval censored and the SEM algorithm is used, the logarithmic transformation can be applied by transforming the interval bounds before iteration Step 1. Thus, the n_k intervals, where the k-th interval is given by (A_{k-1}, A_k) are simply transformed by taking the logarithm $(\log(A_{k-1}), \log(A_k))$. If any interval bound is

negative the intervals have to be shifted to the positive region in order to ensure all interval bounds are non-negative before applying the log transformation. Lower or upper open-ended intervals, e.g., $A_0 = -\infty$ or $A_{n_k} = +\infty$ do not need to be transformed, they remain open-ended since pseudo $\tilde{\mathbf{y}}$ are drawn from a one-sided truncated normal distribution as previously described. After the transformation of the interval bounds, the SEM algorithm is applied to the transformed intervals as stated before.

The proposed algorithm estimates the parameters of a linear mixed model with an interval-censored dependent variable. Another popular and often remarkably simple method (Meng and Rubin, 1991) for estimating parameters using the ML method in models with incomplete (unobserved or missing) data is the EM algorithm introduced by Dempster et al. (1977). Based on the EM algorithm, Stewart (1983) introduced an ML estimator for linear regression models for which convergence is guaranteed (Burridge, 1981). However, the EM algorithm is hard to implement whenever the expectation in the expectation step has a complex form (Yang et al., 2016). Furthermore, the EM algorithm might need many iterations to convergence (Ruud, 1991). The SEM algorithm, introduced as an extension of the classical EM algorithm in Celeux and Dieboldt (1985) and Celeux et al. (1996), simply replaces the expectation step of the EM algorithm with a stochastic approximation (the stochastic step). Based on Celeux and Dieboldt (1985) and Celeux et al. (1996), our proposed approach can be seen as an SEM algorithm, where Step 2 of the algorithm is the stochastic step and Step 3 is the maximization step. A similar SEM algorithm has already been applied for kernel density estimation on aggregated data in Groß et al. (2017) and Walter and Weimer (2018). The SEM algorithm has the advantage of providing more information about the data (Diebolt and Ip, 1996), since in contrast to the EM algorithm, the SEM algorithm maximizes the complete data log-likelihood in the maximization step (Celeux et al., 1996). In contrast to classical EM algorithms, the SEM algorithm does not underestimate the variance of the unknown $\mathbf{y}$. Furthermore, the SEM algorithm can easily be implemented in parallel computing environments to minimize computational time (Meng and Rubin, 1991).

The described theory can also be applied to the estimation of linear models with an interval-censored dependent variable without additional random parameters. In this case, the conditional distribution $f(\mathbf{y}|\mathbf{X}, \mathbf{Z}, \mathbf{v}, \boldsymbol{\theta})$ simplifies to $f(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma_e^2) \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma_e^2)$, because there are no random effects in the model. Hence, in the algorithm it is again iteratively drawn from a two-sided truncated normal distribution.

In this paper, the SEM algorithm is formulated for the case that all unobserved continuous observations are interval censored by the same interval bounds. However, the algorithm can also be extended to situations in which every observation has its own unique interval bounds. This can lead to overlapping intervals and to gaps between different intervals. However, this does not impede the use of the SEM algorithm, because a properly adjusted SEM algorithm would simply draw from the unique intervals. Also, situations in which only some observations are censored and others are observed on a continuous scale can be handled by the proposed SEM algorithm. In this scenario, the SEM algorithm only draws pseudo samples for the interval-censored observations and the continuous observations stay constant during the

iterations of the SEM algorithm.

1.3.2 Estimation of standard errors

In this section, a parametric bootstrap for the estimation of the standard errors of the fixed effects that accounts for the additional variability coming from the interval-censored response variable is introduced. In linear mixed models, the standard errors of the fixed effects are commonly estimated by the inverse of the Fisher information matrix (Pinheiro and Bates, 2000; Verbeke and Molenberghs, 2000). When the SEM algorithm is used to estimate the fixed effects because the dependent variable is interval censored, a linear mixed model is fitted to a new set of pseudo samples $(\tilde{\mathbf{y}}, \mathbf{X}, \mathbf{Z})$ in each of the $B^{(SEM)} + M^{(SEM)}$ maximization steps (see Section 1.3.1). However, estimating the standard errors of the final $\widehat{\boldsymbol{\beta}}$ by simply averaging the standard errors obtained by the inverse of the Fisher information matrix from each iteration step would lead to erroneous results. Since simple averaging neglects the additional variation of $\widehat{\boldsymbol{\beta}}$ between the $B^{(SEM)} + M^{(SEM)}$ iteration steps (the between variance), this approach leads to underestimated standard errors.

A way to successfully consider the additional variation due to the interval-censored response variable is the application of a suitable bootstrap method. Bootstrapping is a resampling method that enables the estimation of standard errors and confidence intervals when no explicit formula is available, or applicable, as in the case of the SEM algorithm. Bootstrapping was introduced by Efron (1979) and further developed by Efron and Stein (1981); Efron (1982); Efron and Tibshirani (1986, 1993). Existing bootstrap methods can be divided into two main categories: non-parametric and parametric. Non-parametric bootstraps replace the unknown original distribution by the empirical distribution of the sample (Ette, 1997), while parametric bootstraps reconstruct the unknown original distribution from data that is generated by a parametric model (Wu, 1986; Davison and Hinkley, 1997; Wehrens et al., 2000). To estimate the standard errors in the SEM context, we propose the use of a parametric bootstrap. This bootstrap approach shows promising results in the literature (Wang et al., 2006; Thai et al., 2013) as well as in the conducted simulation study in Section 1.4. The bootstrap is further extended to account for the additional uncertainty that comes from the interval-censored dependent variable. For this, each continuous bootstrap sample is interval censored according to the original intervals. The parametric bootstrap in the SEM context is described by the following iteration steps:

1. Run the SEM algorithm to obtain $\hat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\beta}}, \widehat{\mathbf{R}}, \widehat{\mathbf{G}})$.

2. Generate a bootstrap sample $\mathbf{y}^* = \mathbf{X}\widehat{\boldsymbol{\beta}} + \mathbf{Z}\mathbf{v}^* + \mathbf{e}^*$, by randomly sampling from $\mathbf{v}^* \sim N(\mathbf{0}, \widehat{\mathbf{G}})$ and $\mathbf{e}^* \sim N(\mathbf{0}, \widehat{\mathbf{R}})$.

3. Divide the continuous bootstrap sample of $\mathbf{y}^*$ into n_k intervals, where the k-th interval is given by (A_{k-1}, A_k). This step is necessary to account for the additional uncertainty coming from the interval-censored dependent variable.

4. Run the SEM algorithm and obtain the bootstrap parameter estimates $\widehat{\boldsymbol{\beta}}_b^*$.

5. Iterate Steps 2-4 $b = 1, \ldots, B$ times.

The bootstrap standard errors are given by

$$SE(\widehat{\boldsymbol{\beta}}) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\widehat{\boldsymbol{\beta}}_b^* - \overline{\boldsymbol{\beta}})^2},$$

where

$$\overline{\boldsymbol{\beta}} = \frac{1}{B}\sum_{b=1}^{B}\widehat{\boldsymbol{\beta}}_b^*.$$

Any bootstrap confidence interval can be estimated by using the percentiles of the bootstrap distribution $\left(\widehat{\boldsymbol{\beta}}^*_{(\alpha/2)}, \widehat{\boldsymbol{\beta}}^*_{(1-\alpha/2)}\right)$, where $\widehat{\boldsymbol{\beta}}^*_{(1-\alpha/2)}$ is equal to the $1-\alpha/2$ percentile of the bootstrapped coefficients $\widehat{\boldsymbol{\beta}}_b^*$ (Davison and Hinkley, 1997; Carlin and Louis, 2000). In the linear regression context, without random parameters, the standard errors can be estimated with a standard non-parametric bootstrap. For linear and linear mixed regression, the SEM algorithm, the non-parametric and the parametric bootstrap are available in the R package `smicd` from the Comprehensive R Archive Network (Walter, 2018).

1.4 Simulation study

In this section, a Monte Carlo simulation study is conducted to asses the performance of the proposed SEM algorithm. The aim of the simulation study is to evaluate the estimated fixed effects obtained by the SEM algorithm and its bootstrapped standard errors under different settings.

The data in the simulation study is generated by three different models. The models in Setting (A) and (C) closely follow the simulation study in Geraci and Bottai (2014); Tzavidis et al. (2016) and Borgoni et al. (2018), while the model in Setting (B) generates a skewed dependent variable that mimics a typical income distribution. The generated continuous data of the dependent variable is interval censored in order to apply the SEM algorithm. The goal of Setting (A) is to evaluate the effect of a different number of intervals on the estimation results of the fixed effects and its standard errors. Therefore, the dependent variable is censored to six, 12, and 24 intervals. The number of intervals are chosen with regards to other censuses. While the German Microcensus collects data on 24 intervals, the census from New Zealand collects income censored to 16 intervals (Statistics New Zealand, 2013) and the Australian census collects data that is censored to only 12 intervals (Australian Bureau of Statistics, 2011). With six intervals we chose an even more extreme scenario to present the performance of the SEM algorithm in such situations. It is expected that naive approaches like the midpoint regression perform substantially worse in such scenarios in comparison to the SEM algorithm. The distribution of the interval-censored dependent variable is given in Appendix 1.7 in Tables 1.6, 1.7 and 1.8. In Setting (B) and (C), the dependent variable is censored to 12 intervals since this number represents a realistic, not too extreme, censoring scheme. The distribution of the dependent variable in Setting (B) is presented in Figure 1.2 and in Table 1.9 in Appendix 1.7. It can be seen that

the interval width increases with increasing y_{ij}, comparable to the German Microcensus. The

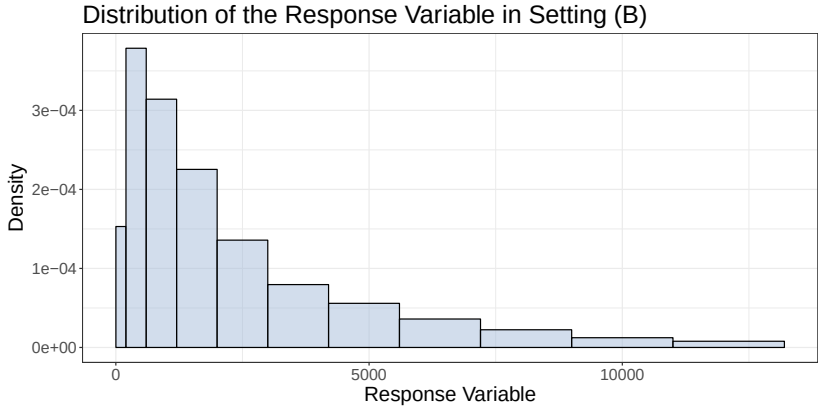


Figure 1.2: Interval-censored distribution of y_{ij} from Setting (B). Since the upper bound of the upper interval is $+\infty$ it is omitted for visualization purposes.

goal of Setting (B) is to study the performance of the SEM algorithm in a scenario in which the linear mixed model assumptions are not fulfilled, e.g., the residuals are not normally distributed. This is often observed for income or wage equations. The violated model assumptions make the use of a transformation such as the logarithmic transformation necessary. Therefore, the SEM algorithm is applied to the log-transformed interval bounds. Setting (C) is chosen to evaluate the SEM algorithm under a more complex linear mixed model. Hence, the model in Setting (C) includes a random slope and a random intercept. The distribution of the interval-censored dependent variable is given in Table 1.10 in Appendix 1.7. The data is generated under the following three models:

- **Setting (A)**

$$y_{ij} = 100 + 2x_{ij} + v_j + e_{ij}, \quad i = 1, \dots, n_j, \quad j = 1, \dots, 100$$

- **Setting (B)**

$$y_{ij} = exp(10 - x_{ij} + v_j + e_{ij}), \quad i = 1, \dots, n_j, \quad j = 1, \dots, 100$$

- **Setting (C)**

$$y_{ij} = 100 + (2 + z_j)x_{ij} + v_j + e_{ij}, \quad i = 1, \dots, n_j, \quad j = 1, \dots, 100$$

In all three settings, the group-specific sample size n_j is kept constant over the Monte Carlo replications and varies between five and 20. This leads to a total sample size of $n_{total} = \sum_{j=1}^{100} n_j = 1259$. In Setting (A) and (C), the independent variable x_{ij} is uniformly distributed $U[0, 20]$, the random intercept $v_j \sim N(0, 3)$ and $e_{ij} \sim N(0, 5)$ with v_j and e_{ij} being independent. The random slope parameter z_j in Setting (C) is $z_j \sim N(0, 2)$ and the correlation between v_j and z_j is set to 0.40. In Setting (B) $x_{ij} \sim N(0, 0.5)$, the random intercept $v_j \sim N(0, 0.16)$

and $e_{ij} \sim N(0, 0.8)$ with v_j and e_{ij} being independent.

In all three settings, the model parameters are estimated by different estimation methods. First, the SEM algorithm, abbreviated by SEM, is applied for the parameter estimation with $B^{(SEM)} = 40$ burn-in and $M^{(SEM)} = 200$ additional iterations. Convergence is checked visually for randomly chosen simulation runs for all three settings by plotting the parameter estimates for each iteration step of the SEM algorithm. In all of the checked simulation runs convergence of the SEM algorithm is achieved. The issue of checking convergence as a practitioner is discussed in more detail in Section 1.5. The performance of the SEM algorithm is compared to other estimation methods that can be applied when the dependent variable is interval censored. The first competing method assigns each unobserved y_{ij} to its corresponding interval midpoint and simply estimates the regression parameters based on these midpoints using REML. This approach is abbreviated by MID. The second competing estimator (INT) is the method proposed by Stewart (1983) and implemented in the R package `intReg` from Toomet (2015). The proposed EM algorithm attains the maximum likelihood parameter estimates when the dependent variable is interval censored. Since the method is only available for models without random effects, it ignores the random effects dependent structure. Hence, the method INT does not account for the clustering of the data. Furthermore, the estimates of the linear mixed model (LME) with the uncensored continuous dependent variable are used as a reference model. LME can be seen as the gold standard because it uses the full information of the dependent variable (the observed sample y_{ij}). In Setting (B), all methods are applied to the log-transformed dependent variable or to the log-transformed intervals, respectively, to assure that the normality assumption of the residuals is fulfilled.

The performance of the discussed methods is evaluated by estimating the relative bias (RB) and the relative efficiency (EFF) of the fixed effects. For each setting $M = 500$ Monte Carlo samples are generated independently by the described models. The RB is given by

$$RB(\widehat{\boldsymbol{\beta}}) = \frac{1}{M} \sum_{m=1}^{M} \left(\frac{\widehat{\boldsymbol{\beta}}^{(m)} - \boldsymbol{\beta}}{\boldsymbol{\beta}} \right) \times 100,$$

where $\widehat{\boldsymbol{\beta}}^{(m)}$ is the estimated fixed effect of iteration step m and $\boldsymbol{\beta}$ is the corresponding true value. The relative efficiency EFF is defined as

$$EFF(\widehat{\boldsymbol{\beta}}) = \frac{s^2_{model}(\widehat{\boldsymbol{\beta}})}{s^2_{LME}(\widehat{\boldsymbol{\beta}})},$$

where $s^2(\widehat{\boldsymbol{\beta}}) = M^{-1} \sum_{m=1}^{M} (\widehat{\boldsymbol{\beta}}^{(m)} - \overline{\boldsymbol{\beta}})^2$ and $\overline{\boldsymbol{\beta}} = M^{-1} \sum_{m=1}^{M} \widehat{\boldsymbol{\beta}}^{(m)}$. The estimates of LME are used as a reference model in the denominator because they are based on the continuous dependent variable (the observed y_{ij}). In the numerator, the estimates of the proposed SEM algorithm and the estimates of the competing methods INT and MID are plugged in.

The number of bootstrap samples for the estimation of the standard errors of the fixed effects is set to $B = 500$. The performance of the bootstrapped standard errors is evaluated by reporting averages over simulations of the empirical (Monte Carlo) standard errors

$SE(\widehat{\boldsymbol{\beta}}) = \sqrt{M^{-1}\sum_{m=1}^{M}(\widehat{\boldsymbol{\beta}}^{(m)} - \overline{\boldsymbol{\beta}})^2}$ and the estimated standard errors of the fixed effects.

The simulation results for the fixed effects are given in Table 1.2 and for the standard errors in Table 1.3. The results are discussed in detail in the following subsections.

Table 1.2: Estimation results of the RB, the EFF and the point estimate for the fixed effects averaged over the 500 Monte Carlo samples for the different settings and the 4 different estimation methods.

	$\hat{\beta}_0$			$\hat{\beta}_1$		
	RB	EFF	$\hat{\beta}_0$	RB	EFF	$\hat{\beta}_1$
Setting (A) - 6 intervals						
LME	0.0072	1.0000	100.0072	-0.0419	1.0000	1.9992
SEM	0.0368	1.5402	100.0368	-0.1889	2.2434	1.9962
MID	-18.4089	44.3905	81.5911	72.9730	69.7612	3.4595
INT	0.0088	1.8337	100.0088	-0.0787	2.7488	1.9984
Setting (A) - 12 intervals						
LME	0.0072	1.0000	100.0072	-0.0419	1.0000	1.9992
SEM	0.0078	1.1142	100.0078	-0.0556	1.3006	1.9989
MID	-5.0185	20.0462	94.9815	19.1119	33.0699	2.3822
INT	0.0025	1.3922	100.0025	-0.0605	1.7834	1.9988
Setting (A) - 24 intervals						
LME	0.0072	1.0000	100.0072	-0.0419	1.0000	1.9992
SEM	0.0048	1.0339	100.0048	-0.0413	1.0816	1.9992
MID	-1.7530	8.6745	98.2470	6.5016	13.5934	2.1300
INT	0.0019	1.3107	100.0019	-0.0567	1.5463	1.9989
Setting (B) - 12 intervals						
LME Log	-0.1269	1.0000	9.9873	-0.4975	1.0000	-0.9950
SEM Log	-0.1181	1.0727	9.9882	-0.4504	1.0931	-0.9955
MID Log	1.8681	1.3730	10.1868	10.6633	1.6880	-1.1066
INT Log	-0.0749	1.4420	9.9925	-0.2422	1.5100	-0.9976
Setting (C) - 12 intervals						
LME	0.0017	1.0000	100.0017	0.5229	1.0000	2.0105
SEM	0.0192	1.3264	100.0192	0.3549	1.0260	2.0071
MID	-0.4420	3.8319	99.5580	0.4657	1.1755	2.0093
INT	-0.0786	5.1123	99.9214	0.9418	1.3299	2.0188

Setting (A)

Setting (A) serves to analyze the effect of the number of intervals the dependent variable is censored to on the estimation results of the fixed effects. Estimating the model parameters with the observed continuous dependent variable (LME) yields, as expected, unbiased estimation results (see Table 1.2). Furthermore, applying the proposed SEM algorithm and INT in order

Table 1.3: Estimation results of the empirical and the estimated standard error averaged over the 500 Monte Carlo samples using the SEM algorithm for all settings.

	$\hat{\beta}_0$		$\hat{\beta}_1$	
	Empirical standard error	Estimated standard error	Empirical standard error	Estimated standard error
Setting (A) - 6 intervals	0.2669	0.2548	0.0171	0.0167
Setting (A) - 12 intervals	0.2270	0.2288	0.0130	0.0129
Setting (A) - 24 intervals	0.2187	0.2206	0.0118	0.0117
Setting (B) - 12 intervals	0.1374	0.1359	0.0521	0.0510
Setting (C) - 12 intervals	0.2656	0.2471	0.1040	0.1035

to estimate the model parameters with an interval-censored dependent variable gives unbiased estimation results for all three censoring schemes of Setting (A). These results are anticipated because theoretically both methods give unbiased estimation results whenever the model assumptions are fulfilled. However, the SEM algorithm is more efficient than INT for all three censoring schemes. This is expected because INT neglects the unobserved heterogeneity coming from the clustered data. As the number of intervals increases, the efficiency of the SEM algorithm also increases because more information is available for the estimation of the parameters. For the 24-interval scenario, the SEM algorithm is almost as efficient as LME. The MID method gives heavily biased and inefficient results for all three censoring scenarios. This is due to the fact that the MID method does not account for the unobserved distribution of the data within each interval. The results from Setting (A) demonstrate the ability of the SEM algorithm to efficiently account for the clustered structure of the data under different censoring scenarios.

In order to evaluate the bootstrapped standard errors of the fixed effects, the estimated standard errors are compared to the empirical standard errors. From Table 1.3, it can be observed that the parametric bootstrap offers a good approximation of the fixed effects standard errors for all censoring schemes. As expected, the standard errors increase when the number of intervals decreases due to the additional uncertainty coming from the fewer interval bounds that are used in the estimation process of the SEM algorithm.

Setting (B)

Setting (B) is set up to evaluate the SEM algorithm under the logarithmic transformation. The findings are comparable to Setting (A). The methods LME, SEM, and INT provide unbiased estimation results for the fixed effects. However, since INT neglects the clustered structure of the data, the SEM algorithm that accounts for the clustering is more efficient. Even though the dependent variable is interval-censored to only 12 intervals, the SEM algorithm is almost as efficient as LME. Again, using the midpoints of the intervals as proxy for the unobserved distribution of the data within each interval yields biased and inefficient results.

The evaluation of the standard errors (see Table 1.3) provides evidence that the proposed parametric bootstrap can also be applied under transformation.

Setting (C)

Finally, Setting (C) serves to evaluate the performance of the SEM algorithm under a more complex linear mixed model (random slope and random intercept). The estimation results for the fixed effects are unbiased for LME, SEM, and INT. The superiority of the SEM algorithm is expressed in the large efficiency advantage compared to INT. This result is expected because the interval regression neglects the random slope and intercept from the true data-generating process. The midpoint regression only exhibits a small bias in this setting. However, this is just an artefact of the censoring scheme that favors, in this particular set up, the midpoint regression. This result does not provide evidence that the midpoint regression yields reliable fixed effects parameter estimates for linear mixed models. In fact, in Cameron (1987) it is shown that the midpoint regression gives biased estimation results.

As for Setting (A) and (B), the parametric bootstrap provides a good approximation for the standard errors of the fixed effects (see Table 1.3).

Overall, the simulation results provide empirical evidence that the SEM algorithm gives unbiased estimation results for different data-generating processes and various censoring schemes. Furthermore, the SEM algorithm is more efficient than the competing methods because it accounts for the clustered structure of the data. The proposed parametric bootstrap for the estimation of the standard errors of the fixed effects accounts well for the additional uncertainty coming from the interval-censored dependent variable.

In the next section, the SEM algorithm is applied to data from the German Microcensus.

1.5 Application: German Microcensus

In this section, the SEM algorithm is applied to model income with data from the German Microcensus. The data and the estimation problem are described in detail in Section 1.2. As presented in Section 1.2, the logarithm of the interval-censored net income variable is modeled by a linear mixed regression model. The explanatory variables education, job, sex, age, region, and the variable specified as random intercept (nationality) are also described in Section 1.2.

In order to estimate the model parameters, the SEM algorithm is applied with 40 burn-in and 200 additional iterations. To assure convergence of the SEM algorithm the convergence of all parameters is visually checked. Exemplary, the convergence plots for the regression parameter female and age are given in Figure 1.3. In these plots, the estimated coefficient (y-axis) is plotted for each iteration step (x-axis) of the SEM algorithm. From the plot, it can be seen that the parameters have converged.

The number of bootstrap samples for the estimation of the standard errors is set to 500, similar to the number of bootstrap samples in the simulation study conducted in Section 1.4. The estimated linear mixed model is presented in Table 1.4. The marginal R-square of the model is 0.41 and the conditional R-square is 0.46, thus the random intercept increases the model fit by 0.05 (Nakagawa and Schielzeth, 2013; Johnson, 2014). The intraclass correlation coefficient (ICC) that measures the correlation among people with the same nationality is 0.23. Thus, the clustered structure of the data should not be neglected and the use of a linear mixed

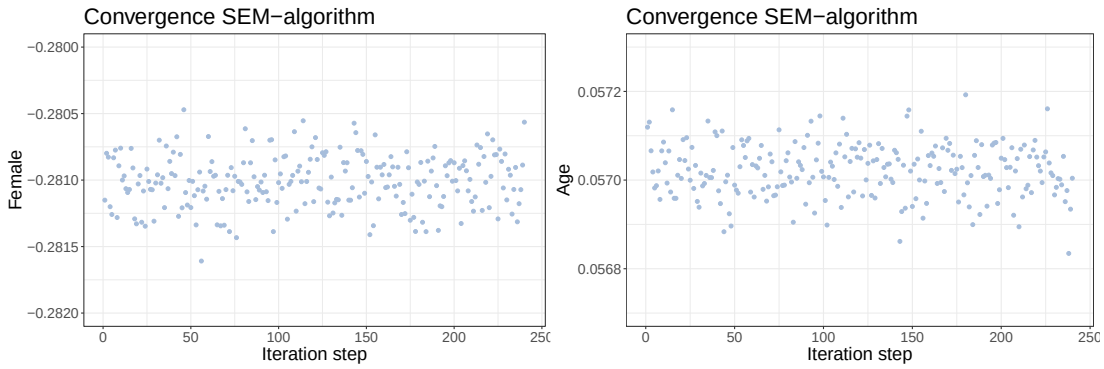


Figure 1.3: Convergence plots of the estimated fixed effects.

model is justified. All of the above measures are estimated as averages over the iteration runs of the SEM algorithm. With regard to the gender pay gap, it can be seen that being female rather than male lowers income, on average, by 28% holding all other regressors constant. Not surprisingly, by looking at education it can be noted that, on average, a higher educational achievement increases income. Also, people with a management job position have, on average, the highest income among all job categories. As commonly described in the literature, there is a significant East West income gap in Germany (Blum et al., 2010). People from the East (formerly German Democratic Republic GDR) have, on average, a 26% lower income than people from the West (base category). The age of a person has, on average, a positive effect on income, but with decreasing marginal effects (negative coefficient for age squared). This result is plausible because people reduce working hours or work in part-time jobs more often when they get older. The bootstrapped 95% confidence intervals indicate that all estimated fixed effects have a significant impact on income.

The presented application shows how the proposed methodology enables researchers to analyze interval-censored income data from the German Microcensus with linear mixed regression models. While the analysis presents some initial interesting insights into the explanation of income, the income equation can be further developed and improved by experts from the field of wage and income modeling. The analysis was conducted with the R package `smicd` (Walter, 2018).

Table 1.4: Estimation results obtained by the SEM algorithm for the linear mixed model fitted to German Microcensus data.

Fixed Effects	Estimates	CI LB$^+$	CI UB$^+$
(Intercept)	6.35443	6.13609	6.58860
Male (Base category)			
Female	-0.28101	-0.28438	-0.27760
ISCED 1 (Base category)			
ISCED 2	0.05860	0.04654	0.06966
ISCED 3c	0.36520	0.34621	0.38485
ISCED 3a	0.01348	0.00040	0.02743
ISCED 3b	0.19238	0.18114	0.20212
ISCED 4a, b	0.29086	0.27901	0.30291
ISCED 5b	0.33089	0.31910	0.34264
ISCED 5a	0.51163	0.49962	0.52200
ISCED 6	0.83952	0.82386	0.85510
West (Base category)			
East	-0.26320	-0.26772	-0.25898
Berlin	-0.17841	-0.18631	-0.17129
Age	0.05702	0.05605	0.05787
Age squared	-0.00057	-0.00058	-0.00056
Manager (Base category)			
Professionals	-0.15501	-0.16233	-0.14794
Technicians and Associate Professionals	-0.26021	-0.26645	-0.25374
Clerical Support Workers	-0.30240	-0.31010	-0.29525
Services and Sales Workers	-0.46191	-0.46928	-0.45430
Skilled Agricultural, Forestry and Fishery Workers	-0.56141	-0.57770	-0.54567
Craft and Related Trades Workers	-0.40545	-0.41212	-0.39827
Plant and Machine Operators and Assemblers	-0.44183	-0.44932	-0.43413
Elementary Occupations	-0.61204	-0.62105	-0.60279
Armed Forces Occupations	-0.06635	-0.08708	-0.04585
Random Effects	Variance	SD	ICC
Nationality (Intercept)	0.01498	0.12239	
Residuals	0.16055	0.40068	0.23393
Marginal R-squared: 0.410	Conditional R-squared: 0.460		

$^+$CI = 95% confidence interval, LB = lower bound, UB = upper bound

1.6 Discussion

In surveys or censuses, income data is often only observed as an interval-censored variable due to confidentiality constraints or in order to decrease item non-response. This is also the case for the largest survey in Europe, the German Microcensus. In this paper, statistical methodology is proposed that enables the estimation of linear and linear mixed models with an interval-censored dependent variable. The proposed methodology is motivated by EM algorithms that are often used to estimate model parameters with ML theory from incomplete data (in our case the interval-censored dependent variable). The introduced SEM algorithm is a further development of the EM algorithm that replaces the unobserved (interval-censored) dependent variable in each iteration step with a continuous pseudo sample. From these continuous pseudo samples, the linear mixed model parameters are estimated. The estimation of the standard errors of the fixed effects is facilitated by a parametric bootstrap. The bootstrap accounts for the additional uncertainty coming from the interval-censored dependent variable. The methodology works for linear mixed models with multiple hierarchical levels and complex correlation structures. It can also be simplified to be applied to linear models. The methodology is evaluated by detailed model-based simulations. The simulation results underline the superiority of the SEM algorithm compared to other available estimation methods. It is implemented in the R package `smicd` from the Comprehensive R Archive Network (Walter, 2018).

The SEM algorithm is then used to model interval-censored income data with a random intercept model. The analysis is based on the SUF of the German Microcensus with 311,659 observations. The estimation results demonstrate the effect of different explanatory variables on income. In order to control for different nationalities, a random intercept is included in the model.

Further research will focus on analytical standard errors for the fixed effects since the proposed parametric bootstrap is very computationally intensive. Additionally, a convergence rule could be implemented to stop the SEM algorithm after a sufficient number of iterations and thus save computing time.

1.7 Appendix

Table 1.5: German Microcensus, 24 intervals: Distribution of personal net income.

Interval	Number of observation
(1,150]	180
(150,300]	341
(300,500]	2133
(500,700]	4553
(700,900]	8053
(900,1100]	14115
(1100,1300]	21793
(1300,1500]	27133
(1500,1700]	30368
(1700,2000]	43299
(2000,2300]	40033
(2300,2600]	29411
(2600,2900]	17516
(2900,3200]	16987
(3200,3600]	15150
(3600,4000]	10203
(4000,4500]	10084
(4500,5000]	5417
(5000,5500]	3628
(5500,6000]	2610
(6000,7500]	3298
(7500,10000]	2834
(10000,18000]	1802
(18000,$+\infty$)	718

Table 1.6: Setting (A), 6 intervals: Distribution of one arbitrary sample data set.

Interval	Number of observation
(1,104]	126
(104,112]	260
(112,120]	236
(120,128]	252
(128,136]	256
(136,$+\infty$)	129

Table 1.7: Setting (A), 12 intervals: Distribution of one arbitrary sample data set.

Interval	Number of observation
(1,100]	32
(100,104]	94
(104,108]	126
(108,112]	134
(112,116]	133
(116,120]	103
(120,124]	120
(124,128]	132
(128,132]	136
(132,136]	120
(136,140]	88
(140,$+\infty$)	41

Table 1.8: Setting (A), 24 intervals: Distribution of one arbitrary sample data set.

Interval	Number of observation
(1,98]	15
(98,100]	17
(100,102]	40
(102,104]	54
(104,106]	63
(106,108]	63
(108,110]	63
(110,112]	71
(112,114]	66
(114,116]	67
(116,118]	53
(118,120]	50
(120,122]	54
(122,124]	66
(124,126]	63
(126,128]	69
(128,130]	74
(130,132]	62
(132,134]	64
(134,136]	56
(136,138]	45
(138,140]	43
(140,142]	33
(142,$+\infty$)	8

Table 1.9: Setting (B), 12 intervals: Distribution of one arbitrary sample data set.

Interval	Number of observation
(1,200]	37
(200,600]	184
(600,1200]	229
(1200,2000]	219
(2000,3000]	165
(3000,4200]	116
(4200,5600]	95
(5600,7200]	70
(7200,9000]	49
(9000,11000]	30
(11000,13200]	21
(13200,$+\infty$)	44

Table 1.10: Setting (C), 12 intervals: Distribution of one arbitrary sample data set.

Interval	Number of observation
(1,75]	10
(75,87.5]	20
(87.5,100]	134
(100,112]	338
(112,125]	265
(125,138]	189
(138,150]	136
(150,162]	95
(162,175]	37
(175,188]	24
(188,200]	6
(200,$+\infty$)	5

Table 1.11: Distribution of the variable Sex.

Sex	Percentage
Male	74.8
Female	25.2

Table 1.12: Distribution of the variable Education.

Education	Percentage
ISCED 1 – no general or vocational certificate or school certificate obtained after no more than 7 years of school attendance)	1.7
ISCED 2 – secondary education without professional degree or secondary degree with completed semi-skilled training, internship or year of pre-vocational training or no general graduation, but with semi-skilled training, internship or pre-vocational training	7.3
ISCED 3c – preparatory service for intermediate service in public administration	0.7
ISCED 3a – qualification for university or university of applied science	2.2
ISCED 3b – apprenticeship or vocational qualifying degree at a full-time vocational school, annual school of health care	45.4
ISCED 4a, b – qualification for university or university of applied science and apprenticeship, vocational qualifying degree at a full-time vocational school, annual school of health care	7.3
ISCED 5b – master craftsman, technician, or equivalent technical college degree, 2 or 3 years medical school, university of cooperative education degree, or specialized or engineering school of the GDR graduation or public administration college degree	14.0
ISCED 5a – university of applied science, university	19.1
ISCED 6 – doctorate	2.2

Table 1.13: Distribution of the variable Region.

Region	Percentage
West	80.8
East	15.4
Berlin	3.8

Table 1.14: Distribution of the variable Age.

Statistical Measure	Age
Min.	16
1st Qu.	37
Median	45
Mean	44
3rd Qu.	52
Max.	93

Table 1.15: Distribution of the variable Job.

Job field	Percentage
Managers	6.6
Professionals	14.7
Technicians and Associate Professionals	25.8
Clerical Support Workers	11.1
Services and Sales Workers	10.2
Skilled Agricultural, Forestry and Fishery Workers	1.0
Craft and Related Trades Workers	14.7
Plant and Machine Operators and Assemblers	10.5
Elementary Occupations	4.8
Armed Forces Occupations	0.5

Table 1.16: Distribution of the variable Nationality.

Country of the nationality	Percentage
Germany	91.47
Bosnia and Herzegovina	0.24
Bulgaria	0.10
France	0.17
Croatia	0.34
Greece	0.43
Italy	0.89
Macedonia	0.13
Netherlands	0.21
Kosovo	0.34
Austria	0.24
Poland	0.79
Portugal	0.17
Romania	0.26
Russia	0.56
Spain	0.15
Turkey	2.31
Hungary	0.11
Ukraine	0.15
Morocco	0.09
South America	0.09
United States of America	0.17
Afghanistan	0.05
Vietnam	0.13
Iraq	0.06
Iran	0.10
Kazakhstan	0.15
Thailand	0.03
China	0.06

Part II

Direct Estimation and Prediction of Statistical Indicators with Interval-Censored Data

Chapter 2

Estimating Poverty and Inequality Indicators using Interval-Censored Income Data from the German Microcensus

2.1 Introduction

In its Global Risks Report 2017, the World Economic Forum proclaims rising income and wealth disparity as the number one trend in determing global developments, governing the risks of, among others, profound social instability and unemployment (World Economic Forum, 2017). Germany has also faced an increase in income inequality since its reunification in 1990 (Fuchs-Schündeln et al., 2010; Bönke et al., 2014). Yet, the question of how poverty and inequality is defined and can accurately be measured or diagnosed in a society remains debatable, see for example Hagenaars and Vos (1988) and Lok-Dessallien (1999). A common way to measure poverty and inequality is the estimation of statistical poverty and inequality indicators. However, for several reasons computing them in practice is not a trivial task. Since income information is not easily accessible governments or statistical offices need to conduct surveys or censuses to garner information about personal or household income. One main difficulty is that, in most societies, income is considered a private topic. In the survey literature, questions about the aspects of income are referred to as "sensitive question", therefore item non-response is high for these questions (Hagenaars and Vos, 1988; Moore and Welniak, 2000). To counter this, many censuses, such as the German (Statistisches Bundesamt, 2017), the Australian (Australian Bureau of Statistics, 2011), the Colombian (Departamento Administrativo Nacional De Estadística, 2005) and the census from New Zealand (Statistics New Zealand, 2013), do not ask for the exact income of their citizens. They ask only for the income interval a person or household belongs to, thereby creating a sense of anonymity. The so obtained income data is not metric but rather interval censored (or grouped). This makes the use of standard formulas for the estimation of poverty and inequality indicators impossible because they rely on metric

data.

To clarify the terminology, depending on the author, the term grouped or censored income data can have different statistical meanings. Some authors such as Milanovic (2003) andMinoiu and Reddy (2008) use the term grouped data to refer to quantile means and Chotikapanich et al. (2007) consider population shares and class means. We use the term interval-censored (or grouped) data, to refer to data that has the form of a frequency table, as in Hall and Wand (1996) or Chen (2017). This type of data is obtained by the aforementioned censuses.

A common parametric approach for density estimation from interval-censored data is the use of the multinomial distribution, see for example Reed and Wu (2008) and Kleiber (2008). From the estimated parametric density any poverty and inequality indicator can be calculated. Chen (2017) proposes a generalized approach to multinomial maximum likelihood estimation for several types of grouped data, showing its consistency and supplying complementary simulation results.

With respect to inequality indicators, Kakwani and Podder (2008) argue against the parametric estimation of the income density from grouped data due to its lack of precision and present a method that can be utilized to estimate the Lorenz curve directly from the interval-censored data in order to compute inequality indicators.

While many authors agree with the Kakwani and Podder (2008) critique on the estimation of parametric distributions, they resolve these issues by instead using non-parametric estimators to model income instead. The popularity of these estimators comes from the fact that they do not require any prior assumptions about the theoretical distribution or its family. Although most authors do not directly address the topic of interval-censored or grouped data, there is much literature about rounded data, which is easily obtained from interval-censored data by substituting the intervals with their centers. Hall (1982), Scott and Sheather (1985), and Hall and Wand (1996) study the effects of rounded and interval-censored data on standard, non-parametric kernel density estimation (KDE). In contrast to uncensored data, they derive that the mean integrated squared error of the KDE for rounded data depends on the smoothness of the used kernel function. Moreover, they find that censoring affects the bias rather than the variance of the estimate. Additionally, Hall and Wand (1996) present minimum grid sizes for KDE which are needed to achieve a given degree of accuracy. Grid size corresponds to the amount of points and therefore to the amount of intervals when the interval centers are used on which the density is estimated.

Wang and Wertelecki (2013) point out that standard KDE leads to increasingly spiky density estimates at rounded points with a growing sample size. KDE becomes smoother when larger bandwidths are used, thus an oversmoothed bandwidth selection was proposed by Wand and Jones (1995) and implemented in the R package `KernSmooth` (Wand, 2015). Nevertheless, Wang and Wertelecki (2013) argue that this mostly leads to flatter estimates that underestimate the true density. Alternatively, they propose a bootstrap-type kernel density estimator and show in a simulation study that the estimator provides better accuracy than the standard KDE and the over-smoothed KDE.

Groß et al. (2017) melt the principle of stochastic expectation-maximization algorithms

(Nielsen, 2000) with KDE to create a new density estimation alogrithm for rounded two-dimensional data. Its superiority compared to a standard KDE is made apparent in a simulation study (Groß and Rendtel, 2016). Their algorithm can be seen as a generalization of the Wang and Wertelecki (2013) estimator.

Although a correctly estimated density leads to correctly estimated poverty and inequality indicators Lenau and Münnich (2016) focus their analysis on the impact of different estimation methods on the direct performance of the estimated statistical indicators. They evaluate three different estimation methods: Non-parametric splines, estimating the generalized beta distribution of the second kind (GB2), and linear interpolation. Linear interpolation is the method used by the German statistical offices to estimate indicators from interval-censored data. This approach is similar to assuming a uniform distribution within each interval. They conclude that the performance of the different methods depends highly on the censoring schemes and none of the methods showed adequate results in terms of bias and variance for all analyzed censoring schemes.

To overcome the disadvantage of the different estimation methods, we propose a non-parametric KDE algorithm that is based on the algorithm of Groß et al. (2017). The KDE algorithm enables the estimation of poverty and inequality indicators from interval-censored data under different censoring schemes. In order to obtain representative results, the KDE algorithm can incorporate survey weights. The standard errors of the statistical indicators are estimated by a non-parametric bootstrap.

The paper is structured as follows. In Section 2.2, the KDE algorithm and the proposed non-parametric bootstrap are introduced. In Section 2.3, the properties of the KDE algorithm and the bootstrap are evaluated using Monte Carlo simulation studies under different interval-censoring schemes and different theoretical distributions. In Section 2.4, the algorithm is used to estimate regional poverty and inequality indicators from the German Microcensus. A final discussion of the major results, their implications, and an outlook is given in Section 2.5.

2.2 Methodology

In order to estimate poverty and inequality indicators, we propose a novel KDE algorithm to generate metric pseudo samples from the observed interval-censored data. By using the pseudo samples, poverty and inequality indicators can be estimated applying standard formulas. In the next two subsections, the novel KDE algorithm is introduced and a non-parametric bootstrap is proposed for the variance estimation of the statistical indicators.

2.2.1 Kernel density estimation from interval-censored data

Kernel density estimation is one of the most established non-parametric density estimation techniques in the literature and was first introduced by Rosenblatt (1956) and Parzen (1962). It is applied to estimate a continuous density from a random variable with density $f(x)$ directly from its independent and identically distributed observations without making any prior assumptions about its distributional family. Let $X = \{X_1, X_2, \ldots, X_n\}$ denote a sample of

size n. For $i = 1, \ldots, n$ the KDE is defined as

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} k\left(\frac{x - X_i}{h}\right),$$

where $k(\cdot)$ is a kernel function and the bandwidth is denoted by $h > 0$. For the shape and performance of the estimator, the choice of the bandwidth h is essential. The larger the h, the smoother the estimated density, but also the more information about details, such as local extrema, may be lost (Zambom and Dias, 2012). Hence, bandwidth selection methods are widely discussed in the literature with the two main categories being plug-in and cross-validation (Jones et al., 1996; Loader, 1999; Henderson and Parmeter, 2015). The basic idea of the first is to minimize the asymptotic mean integrated squared error whilst substituting the unknown density in the optimization with a pilot estimate, whereas the second method is a more data-driven approach, for example, utilizing leave-one-out cross-validation.

In the presented *Naive* KDE, it is assumed that observations are taken directly from the continuous distribution that is to be estimated. Often, however, collecting continuous data is not possible due to various restrictions in practice, such as, for example, confidentiality concerns. In these situations we are left with interval-cesnsored data, where only the interval information is observed. Thus, only the lower A_{K-1} and upper A_K interval bounds (A_{K-1}, A_K) of X is observed and its continuous value remains unknown. The continous scale is divided into n_K intervals. The variable K $(1 \leq K \leq n_K)$ indicates which of the intervals an observation $K = \{K_1, K_2, \ldots, K_n\}$ falls into. Open-ended intervals, thus $A_0 = -\infty$ or $A_{n_K} = +\infty$ have to be replaced by a finite number (see Section 2.3.4). Applying KDE to the interval midpoints of the interval-censored data falsely allocates too much probability mass to the midpoints and too little to the unobserved X_i. This leads to spiky estimates, unless the bandwidth is chosen to be very large (Wang and Wertelecki, 2013). Increasing the bandwidth cannot be considered as a solution to this problem because this causes additional loss of information about the underlying true distribution. The Wang and Wertelecki (2013) simulation study further found standard KDE to be very sensitive to sample size when interval censoring is ignored. Furthermore, Hall (1982) and Hall and Wand (1996) showed that, in contrast to uncensored data, the asymptotic performance of KDE for interval-censored data depends on the smoothness of the kernel function in use.

These findings underline the necessity of using a more sophisticated non-parametric approach for density estimation from interval-censored data. Wang and Wertelecki (2013) introduce a bootstrap-type KDE based on a measurement error model and confirmed its superiority over the *Naive* estimator with simulations. Groß et al. (2017) then generalized and extended the approach based on the stochastic expectation-maximization (SEM) algorithm and iterative bootstrapping. Their newly proposed density estimator, abbreviated to GRSST, outperforms *Naive* KDE and a measurement error-based estimator by Delaigle (2007), especially for stronger interval censoring. Since the GRSST estimator was formulated for two-dimensional data with equal-sized interval censoring, we reformulate the approach. The reformulated KDE algorithm enables the density estimation for one-dimensional data with unequally sized censor-

ing. During the algorithm pseudo samples of the unobserved X_i are generated from which the density and any statistical indicator can be estimated. Hence, for the estimation of poverty and inequality indicators the unobserved distribution of the interval-censored X is reconstructed. This is done with the use of the known interval information K. From Bayes' theorem it follows that the conditional distribution of X given K is:

$$\pi(X|K) \propto \pi(K|X)\pi(X),$$

where $\pi(K|X)$ is defined as a product of Dirac distributions $\pi(K|X) = \prod_{i=1}^{n} \pi(K_i|X_i)$ with

$$\pi(K_i|X_i) = \begin{cases} 1 & \text{if } A_{K-1} \leq X_i \leq A_K, \\ 0 & \text{otherwise,} \end{cases}$$

for $i = 1, \ldots, n$. Using this formulation pseudo samples (imputations) of the unknown X_i are drawn that enable the estimation of any statistical indicator. Since $\pi(X) = \prod_{i=1}^{n} f(X_i)$ is initially unknown, an initializing estimate $\hat{f}_h(x)$ that is based on the interval midpoints, serves as a proxy. After that, the pseudo samples drawn from $\pi(X|K)$ are used to re-estimate $\pi(X)$. The following section focuses on the exact implementation of the proposed algorithm and discusses similarities to the popular EM algorithm by Dempster et al. (1977) and the SEM algorithm by Celeux and Dieboldt (1985) and Celeux et al. (1996).

Estimation and Computational Details

As in Groß et al. (2017) to fit the model pseudosamples of X_i are drawn from the conditional distribution

$$\pi(X_i|K_i) \propto \mathbf{I}(A_{K-1} \leq X_i \leq A_K) f(X_i),$$

where $\mathbf{I}(\cdot)$ denotes the indicator function. The conditional distribution of X_i given K_i is the product of a uniform distribution and density $f(x)$. As the density $f(x)$ is unknown it is replaced by $\hat{f}_h(x)$, an estimate that is obtained by the prior defined kernel density estimator. Hence, X_i is iteratively drawn from the known interval (A_{K-1}, A_K) with the current density estimate $\hat{f}_h(x)$ used as sampling weight. The steps of the iterative algorithm are described below.

Step 1: Use the midpoints of the intervals as pseudo $\tilde{X}_i$ for the unknown X_i. Obtain a pilot estimate of $\hat{f}_h(x)$, by applying KDE. Choose a sufficiently large bandwidth h, such that no rounding spikes occur.

Step 2: Evaluate $\hat{f}_h(x)$ on an equal-spaced fine grid $G = \{g_1, \ldots, g_j\}$ with j grid points $g_1, \ldots, g_j$. The width of the grid is denoted by δ_g. It is given by,

$$\delta_g = \frac{|A_0 - A_{n_K}|}{j - 1},$$

and the grid is defined as,

$$G = \{g_1 = A_0, g_2 = A_0 + \delta_g, g_3 = A_0 + 2\delta_g, \ldots, g_{j-1} = A_0 + (j-2)\delta_g, g_j = A_{n_K}\}.$$

Step 3: Sample from $\pi(X|K)$ by drawing a pseudo sample $\tilde{X}_i$ randomly from $\{G_K = g_j | g_j \in (A_{K-1}, A_K)\}$ with sampling weights $\hat{f}_h(\tilde{X}_i)$ for $K = 1, \ldots, n_K$. The sample size within each interval is given by the number of observations within each interval.

Step 4: Estimate any statistical indicator of interest $\hat{I}$ using the pseudo $\tilde{X}_i$.

Step 5: Recompute $\hat{f}_h(x)$, using the pseudo samples $\tilde{X}_i$ obtained in iteration Step 3.

Step 6: Repeat Steps 2-5, with $B_{(KDE)}$ burn-in and $S_{(KDE)}$ additional iterations.

Step 7: Discard the $B_{(KDE)}$ burn-in iterations and estimate $\hat{I}$ by averaging the obtained $S_{(KDE)}$ estimates.

The KDE algorithm estimates the distribution of an interval-censored variable by only using the interval information. An algorithm that is widely used for models that depend on latent variables (in our case the unobserved interval-censored X) is the EM algorithm (Dempster et al., 1977). In the EM algorithm the expectation of $X|K$ is obtained analytically. However, in the context of kernel density estimation this does not work because all values inside an interval would be concentrated at one point, the expectation. In a SEM algorithm, the analytical E-step from the EM algorithm is replaced by the drawing of pseudo samples (Celeux and Dieboldt, 1985; Celeux et al., 1996). In case of the KDE algorithm, it is drawn from the distribution of $\pi(X|K)$. Hence, the proposed KDE algorithm has similarities to a SEM algorithm. In its common form, the EM and SEM algorithm are used for maximum likelihood (ML) estimation with unobserved data. McLachlan and Krishnan (2008) proposed a generalization of the SEM algorithm that can be used with surrogates for the ML estimation. In the KDE algorithm the maximization of the asymptotic mean integrated squared error is used as such a surrogate. More information on the similarities between the KDE, the EM, and SEM algorithm and the GRSST estimator – on which the KDE algorithm is based – are given in (Groß et al., 2017).

2.2.2 Variance estimation

This section introduces a method for the variance estimation of the statistical indicators that are estimated by the KDE algorithm. A common way to estimate the variance, if X is observed on a continuous scale, is linearization. Taylor linearization (Tepping, 1968; Woodruff, 1971; Wolter, 1985; Tille, 2001) is a well-known and commonly applied method for the estimation of variance for non-linear indicators, such as ratios or correlations. However, the method cannot be applied for the variance estimation of all non-linear indicators. For the variance estimation of mathematically more complex indicators, e.g., the Gini coefficient, Deville (1999) introduced the generalized linearization method. The generalized linearization method is also used

by Eurostat for the variance estimation of complex indicators (Osier, 2009). Nevertheless, linearization cannot be applied when the variable of interest is observed as an interval-censored variable (Lenau and Münnich, 2016). To still produce variance estimates, resampling methods, such as bootstrapping can be applied (Münnich, 2008). Bootstrapping methods approximate the variance of an estimated indicator, in cases where the variance cannot be stated as closed-form solution (Bruch et al., 2011). Therefore, the bootstrap introduced by Efron (1979) and Shao and Tu (1995) is used for the variance estimation of the indicators estimated by the KDE algorithm. Also, any confidence interval can be estimated by using the quantiles from the bootstrap results (Rao and Wu, 1988; Rao et al., 1992; Pretson, 2008). The use of the bootstrapping allows us to avoid theoretical calculations. However, the potential disadvantage is a long computational time. The non-parametric bootstrap is based on the assumption that the drawn sample is representative of the population. Therefore, the empirical distribution function $\hat{F}$ is a non-parametric estimate of the population distribution F. The desired poverty indicator of interest $\hat{I}$, is the empirical estimate of the true parameter. The bootstrap standard errors are calculated as follows:

Step 1: Draw with replacement a bootstrap sample of the interval-censored $X_i^{(b)}$ of size n from the sample data set.

Step 2: Apply the KDE algorithm to the bootstrap interval-censored sample $X_i^{(b)}$ for the estimation of any indicator $\hat{I}^{(b)}$ of interest.

Iterate Steps 1-2, $b = 1, \ldots, B$ times and estimate the standard error $se(\hat{I}) = \sqrt{\frac{\sum_{b=1}^{B}(\hat{I}^{(b)} - \overline{I}^{(b)})^2}{B}}$ with $\overline{I} = \frac{1}{B}\sum_{b=1}^{B}\hat{I}^{(b)}$.

2.3 Simulation results

This section presents extensive model-based simulation results in order to evaluate the performance of the KDE algorithm in the context of estimating poverty and inequality from interval-censored income data. The simulation study is set up with the following specifications. From a theoretical distribution $M = 500$ samples of simulated income data are drawn. The drawn samples are censored to specific intervals. The sample size for each sample is $n = 10000$. The KDE algorithm is evaluated for large samples because interval-censored income data is common for censuses which, in general, have very large sample sizes. For instance, in the application in Section 2.4, German Microcensus data is used which has a sample size of $n = 454852$. From the simulated interval-censored income data different poverty and inequality indicators are estimated. The formulas are presented for metric data because the KDE algorithm generates metric data from interval-censored data that is used to estimate the statistical indicators. Consider $X = (X_1, \ldots, X_n)$ with $X_1 \leq \ldots \leq X_n$ and let $w = (w_1, \ldots, w_n)$ be the corresponding sampling weights. The weighted mean and the weighted quantiles (10%, 25%, 50%,

75%, 90%) are given by

$$\hat{I}_{\text{Mean}} = \frac{\sum_{i=1}^{n} w_i X_i}{\sum_{i=1}^{n} w_i}, \tag{2.1}$$

$$\hat{I}_{\text{Q}_{(p)}} = \begin{cases} \frac{1}{2}\left(X_i + X_{i+1}\right) & \text{if } \sum_{j=1}^{i} w_j = p \sum_{j=1}^{n} w_j, \\ X_{i+1} & \text{if } \sum_{j=1}^{i} w_j \leq p \sum_{j=1}^{n} w_j \leq \sum_{j=1}^{i+1} w_j, \end{cases} \tag{2.2}$$

where p denotes the quantile $p \in (0, 1)$. In the simulation study sampling weights are not included, because they are not needed to evaluate the performance of the KDE algorithm. Therefore, $w_i = 1 \forall i$ in the simulation study. However, in the application in Section 2.4 weights are included for representative inference. The weighted poverty measures Headcount Ratio (HCR) and Poverty Gap (PGap) (Foster et al., 1984) are given by

$$\hat{I}_{\text{HCR}} = \frac{1}{\sum_{i=1}^{n} w_i} \sum_{i=1}^{n} w_i \mathbf{I}(X_i \leq z), \tag{2.3}$$

$$\hat{I}_{\text{PGap}} = \frac{1}{\sum_{i=1}^{n} w_i} \sum_{i=1}^{n} w_i \left(\frac{z - X_i}{z}\right) \mathbf{I}(X_i \leq z), \tag{2.4}$$

where $\mathbf{I}(\cdot)$ denotes the indicator function. The HCR and PGap include a threshold z that is known as the poverty line. For the simulation a relative poverty line, defined as 60% of the median of the simulated income variable is chosen. This corresponds to the EU definition (Eurostat, 2014). The HCR is a measure of the percentage of observations (individuals or households) below the poverty line, whereas the PGap measures the average distance of those observations from the poverty line. Inequality is commonly measured by the Gini coefficient (Gini, 1912) and the quintile share ratio (QSR). The weighted indicators are estimated by

$$\hat{I}_{\text{Gini}} = \left[\frac{2 \sum_{i=1}^{n} \left(w_i x_i \sum_{j=1}^{i} w_j \right) - \sum_{i=1}^{n} w_i^2 X_i}{\sum_{i=1}^{n} w_i \sum_{i=1}^{n} w_i X_i} - 1 \right], \tag{2.5}$$

$$\hat{I}_{\text{QSR}} = \frac{\sum_{i=1}^{n} \mathbf{I}(X_i \geq \hat{Q}_{0.8}) w_i X_i}{\sum_{i=1}^{n} \mathbf{I}(X_i \leq \hat{Q}_{0.2}) w_i X_i}. \tag{2.6}$$

The range of the Gini coefficient lies between 0 and 1. The higher its value, the higher the inequality. If the Gini coefficient is equal to 0 there is perfect equality in the data, whereas a Gini coefficient of 1 indicates perfect inequality. The QSR is the ratio of observations richer than 20% of the richest observations to the 20% of the poorest observations. Higher values of the QSR indicate higher inequality.

The indicators are estimated by the proposed KDE algorithm. The number of burn-in iterations of the algorithm is set to $B_{(KDE)} = 80$, the number of additional iterations $S_{(KDE)} = 400$. Our experiences running several simulations show that 480 iterations are usually enough to ensure convergence. Nevertheless, we check the convergence plots from randomly chosen simulation runs to assure that the indicators in the presented simulations converge. The issue of convergence is discussed in more detail in Section 2.4. The number of grid points is set to

$j = 4000$. In general, a higher number of grid points leads to more precise estimation results, because the number of grid points determines how many unique values the pseudo samples of the interval-censored variable can consist of. However, the estimation time increases with the increasing number of grid points. In the simulation, the number of grid points is chosen such that a further increase in the number of grid points does not lead to better estimation results. The presented poverty and inequality indicators are not only estimated by the KDE algorithm (KDE). For comparison, the indicators are also estimated by linear interpolation. This method is used by the Federal Office of Statistics in Germany for the estimation of poverty and inequality indicators from the interval-censored income variable of the German Microcensus (Information und Technik (NRW), 2009). This approach gives the same results as assuming a uniform distribution within the income classes (Uni). Furthermore, the statistical indicators are estimated by using the midpoints (Mid) of the intervals as a proxy for the unobserved values within the income interval. Finally, the statistical indicators are also estimated with the continuous uncensored data (True). The results of the point estimates are evaluated by the relative bias (rB),

$$rB(\hat{I}) = \frac{1}{M}\sum_{m=1}^{M}\left(\frac{\hat{I}_m - I}{I}\right) \times 100,$$

and the empirical standard errors (se.emp),

$$se.emp(\hat{I}) = \sqrt{\frac{1}{M}\sum_{m=1}^{M}(\hat{I}_m - \overline{I})^2},$$

with

$$\overline{I} = \frac{1}{M}\sum_{m=1}^{M}\hat{I}_m.$$

The proposed non-parametric bootstrap for the estimation of the standard errors is evaluated by comparing the estimated standard errors to the empirical standard errors. The bootstrap is run with $B = 100$. This number shows it is sufficient to obtain valid approximations of the standard errors.

The simulation study is divided into four subsections. In Section 2.3.1, the influence of different numbers of intervals on the performance of the KDE algorithm is evaluated. In Section 2.3.2, different true distributions are evaluated and, in Section 2.3.3, the effect of equal vs. ascending interval width is studied. Section 2.3.4 summarizes the final results and discusses the issue of how to handle open-ended intervals.

2.3.1 Different interval-censoring scenarios

In this section, the influence of the number of intervals on the performance of the KDE algorithm is studied. As theoretical distribution the four-parameter GB2 distribution that is often used to model income is specified such that the GB2 distribution well approximates the empirical German income distribution (Graf and Nedyalkova, 2014). The chosen parameters are

given in Table 2.3. The drawn samples are interval censored using three different censoring scenarios. In Scenario 1, the data is censored to 24 intervals as in the German Microcensus (Statistisches Bundesamt, 2017) that is used in the application in Section 2.4. The interval widths are chosen such that the interval-censored theoretical distribution follows the empirical distribution of the household income in the German Microcensus. This is visualized in Figure 2.1 in the upper two panels. The lower two panels show the GB2 distribution censored to 16 intervals (Scenario 2) and eight intervals (Scenario 3). The performance of the algorithm with the lower number of classes is studied because censuses from other countries censor the income variable to fewer than 24 intervals. For example, in the census from New Zealand the income variable is censored to 16 intervals (Statistics New Zealand, 2013), in the Australian census the data is censored to 12 intervals (Australian Bureau of Statistics, 2011), and in the Colombian census the income variable is censored to only nine intervals (Departamento Administrativo Nacional De Estadística, 2005).

The results of the point estimates are given in Table 2.1. Using the continuous uncensored data for the estimation of the poverty and inequality indicators leads to unbiased results. This is not surprising as the sample size ($n = 10000$) is very large. Using only the interval information, the KDE algorithm outperforms the other approaches (Mid and Uni) in all three scenarios. The out-performance is especially stronger for indicators that rely on the whole shape of the distribution (Gini coefficient, mean), for the more extreme quantiles (10% quantile and 90 % quantile), and for indicators that rely on more extreme quantiles (QSR). As the number of intervals decreases, the performance of the KDE algorithm worsens. Nevertheless, the bias is still under 1% for all indicators, except for the QSR, PGap and the Gini coefficient. The QSR shows a bias of -1.1%, the PGap a bias of 2.3%, and the Gini coefficient shows a bias of -1.9% in the eight-interval scenario.

The estimated indicators using the other approaches (Mid and Uni) exhibit far larger biases as the number of intervals decreases. For example, in the eight-interval scenario the PGap has a bias of 22% and 20% and the Gini coefficient of 14% and 24% for the estimation approaches Uni and Mid, respectively. This shows the superiority of the KDE algorithm.

The precision of the KDE algorithm, measured by the empirical standard error (se.emp), is for all three scenarios close to the estimation results using the uncensored data. This is the case because the estimated indicators rely on the metric pseudo samples from the KDE algorithm. However, the pseudo samples can – in rare circumstances – include very extreme values that lead to a higher variance when statistical indicators are estimated that rely on the whole distribution. This is, for example, the case for the mean in the 24-interval scenario. The KDE algorithm almost loses no precision for a lower number of intervals. The methods Uni and Mid lead to less precise estimation results, especially with fewer intervals. For some of the estimated quantiles the empirical standard error of the Mid approach is 0. This is due to the fact that the Mid approach estimates the indicators on the midpoints of the intervals. This leads to only 24, 16 or eight unique values, respectively. With a sample size of ($n = 10000$) the estimated quantiles are likely to fall on the same midpoint for each of the 500 Monte Carlo iterations. If the estimated quantile is constant over all Monte Carlo iterations, the empirical

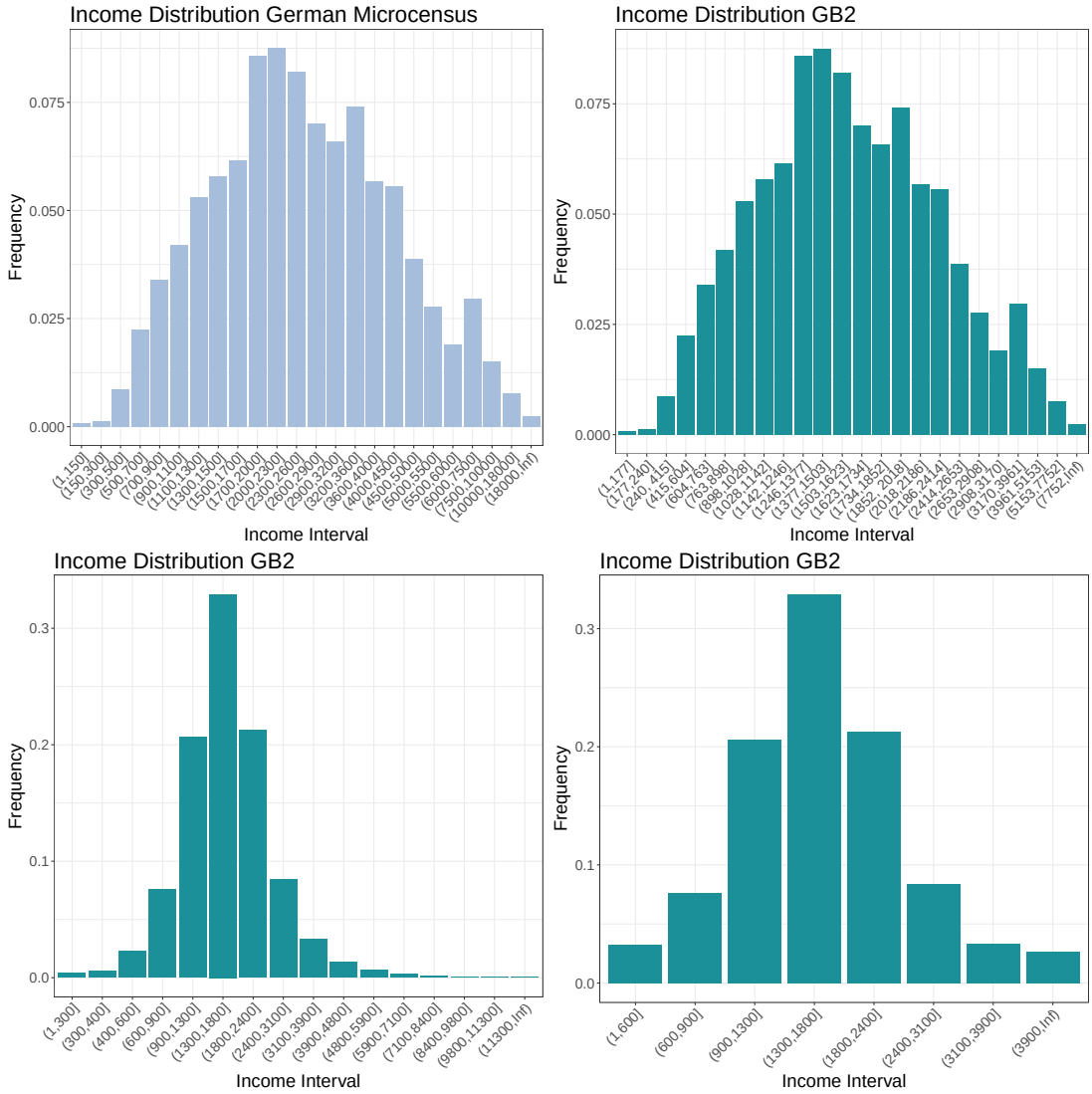


Figure 2.1: Interval-censored income distribution of the German Microcensus (upper left) and theoretical GB2 distribution. The GB2 distribution is censored to 24 (upper right), 16 (lower left) and 8 intervals (lower right).

standard error is 0.

In Table 2.2, the proposed bootstrap for the estimation of the standard errors is evaluated for the three different censoring scenarios. The standard errors estimated by the non-parametric bootstrap (se.est) offer a good approximation of the empirical standard errors (se.emp). This underlines the reliability of the proposed bootstrap method.

2.3.2 Different true distributions

While the previous section evaluates the performance of the KDE algorithm using different censoring schemes, this section focuses on the evaluation of the performance using different theoretical distributions. A large number of theoretical distributions are suggested in the literature for modeling income distributions (McDonald and Ransom, 1979; McDonald, 1984; Mc-

Table 2.1: Relative bias (rB) and the empirical standard error (se.emp) for the different estimation methods estimated for a selection of statistical indicators.

		$Q_{0.1}$	$Q_{0.25}$	Median	$Q_{0.75}$	$Q_{0.9}$	Mean	HCR	QSR	PGap	Gini
Quality Measure	Estimation Method										
		GB2: 24 intervals									
rB	True	0.053	0.036	0.008	-0.003	0.017	0.023	-0.087	-0.005	-0.163	-0.005
	KDE	-0.102	-0.059	-0.033	-0.045	0.121	0.002	-0.141	0.720	0.181	-0.036
	Uni	-0.366	-0.086	0.065	0.080	0.171	1.104	1.087	3.751	2.628	3.374
	Mid	-4.654	0.003	-0.313	1.501	1.848	2.218	-11.962	35.517	1.529	6.161
se.emp	True	87.600	72.172	71.259	109.180	222.019	95.973	0.003	0.049	0.001	0.003
	KDE	84.944	68.284	69.756	112.048	227.883	121.231	0.003	0.067	0.001	0.004
	Uni	96.181	69.987	70.633	119.183	240.357	111.912	0.003	0.060	0.001	0.003
	Mid	83.717	0.000	0.000	738.583	1092.148	137.517	0.003	0.351	0.001	0.005
		GB2: 16 intervals									
rB	True	-0.007	0.012	0.022	0.021	0.014	-0.020	-0.030	-0.077	0.109	-0.102
	KDE	0.323	-0.021	0.260	0.190	-0.051	-0.018	0.478	0.699	0.034	-0.401
	Uni	-0.991	-1.832	0.823	3.492	3.543	1.154	4.522	5.113	7.699	3.691
	Mid	-14.210	-8.097	-1.200	3.499	3.098	1.536	-12.619	92.185	6.194	0.835
se.emp	True	90.029	72.505	78.428	113.178	232.863	101.242	0.003	0.048	0.001	0.003
	KDE	88.476	72.731	73.944	119.657	229.199	101.652	0.003	0.049	0.001	0.003
	Uni	120.142	84.036	81.005	131.425	248.381	110.794	0.003	0.055	0.001	0.003
	Mid	221.137	0.000	0.000	0.000	0.000	121.311	0.003	0.321	0.001	0.004
		GB2: 8 intervals									
rB	True	0.076	0.006	-0.016	0.021	0.017	-0.006	-0.103	-0.051	-0.131	-0.037
	KDE	0.106	-0.173	0.252	0.145	-0.141	-0.685	0.119	-1.151	2.329	-1.871
	Uni	-0.980	-1.850	0.820	3.519	3.587	4.190	4.323	17.586	21.758	13.522
	Mid	-13.972	-8.012	-1.155	3.582	3.092	10.187	-12.555	164.261	20.273	24.256
se.emp	True	92.276	75.720	71.976	111.044	240.443	100.286	0.003	0.050	0.001	0.003
	KDE	88.373	74.822	70.126	113.115	231.700	126.809	0.003	0.071	0.001	0.004
	Uni	120.998	86.888	73.876	128.360	253.586	132.150	0.003	0.075	0.001	0.004
	Mid	220.916	0.000	0.000	0.000	0.000	183.278	0.003	0.481	0.001	0.005

Table 2.2: Empirical and estimated standard error for the selected statistical indicators.

		$Q_{0.1}$	$Q_{0.25}$	Median	$Q_{0.75}$	$Q_{0.9}$	Mean	HCR	QSR	PGap	Gini
Measure	Estimation Method										
		GB2: 24 intervals									
se.emp	KDE	84.944	68.284	69.756	112.048	227.883	121.231	0.003	0.067	0.001	0.004
se.est		84.945	71.525	72.437	110.804	234.200	120.855	0.003	0.067	0.001	0.004
		GB2: 16 intervals									
se.emp	KDE	88.476	72.731	73.944	119.657	229.199	101.652	0.003	0.049	0.001	0.003
se.est		87.972	70.564	68.708	110.969	224.122	96.000	0.003	0.050	0.001	0.003
		GB2: 8 intervals									
se.emp	KDE	88.373	74.822	70.126	113.115	231.700	126.809	0.003	0.071	0.001	0.004
se.est		85.036	71.131	68.217	109.751	229.160	132.415	0.003	0.076	0.001	0.005

Donald and Xu, 1995; Bandourian et al., 2003; Kleiber and Kotz, 2003). According to McDonald (1984), McDonald and Xu (1995), Bordley et al. (1997), McDonald and Ransom (2008) the GB2 distribution is well-suited for modelling income and it is superior to other parametric distributions (Kleiber and Kotz, 2003; Dastrup et al., 2007; Jenkins, 2009). Nevertheless, two special cases of the GB2 distribution are used for evaluations in order to illustrate the flexibility of the KDE algorithm: the Dagum (Dagum, 1977) distribution and the Singh-Maddala (Singh and Maddala, 1976) distribution. The choice of parameters follows Bandourian et al. (2002) (see Table 2.3) in order to approximate empirical income distributions. The data is censored to eight intervals and the interval width is chosen such that the relative frequency within each interval is similar to the eight-interval GB2 scenario from the previous section (see Figure 2.2 and 2.1). The eight-interval scenario is chosen to evaluate the KDE algorithm under extreme scenarios. By keeping the relative frequencies equal within each interval the effect of different distributions (GB2, Dagum, and Singh-Maddala) on the estimation results is isolatedly evaluated.

Table 2.3: Distributions for the Model-based simulation.

Distribution	Parameter			
GB2	7.481	16351	0.4	0.468
Dagum	4.413	94075	0.337	
Singh-Maddala	1.771	500000	25.12	

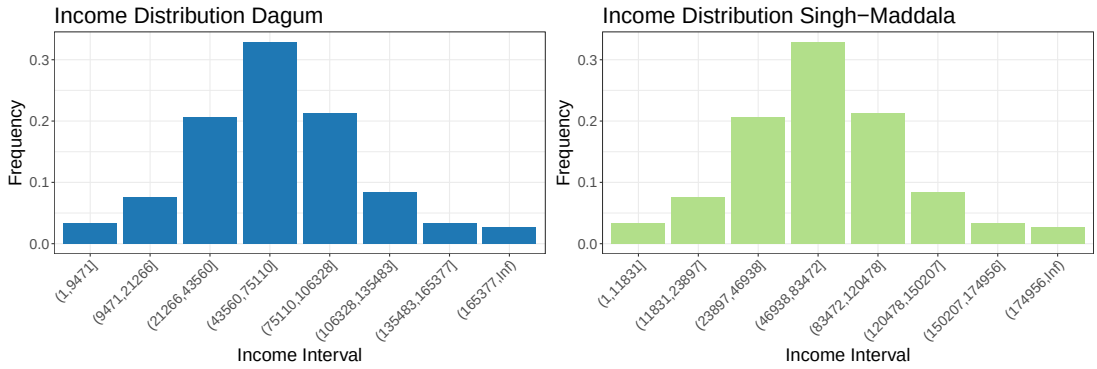


Figure 2.2: Dagum and Singh-Maddala distribution censored to 8 intervals.

The estimation results of the point estimates are given in Table 2.4 (Dagum and Singh-Maddala) and Table 2.1 (GB2). As expected, using the uncensored data leads to unbiased estimation results. Also, the KDE algorithm that only uses the interval information yields unbiased results for all indicators under the different scenarios. Hence, the performance of the KDE algorithm is not impaired by the underlying theoretical distribution. The benchmark methods (Uni and Mid) give heavily biased estimation results, especially for indicators that depend on the whole distribution. For example, the QSR has a bias of 16.5% (Uni) and 210% (Mid) for the Dagum scenario and 18.5% (Uni) and 200% (Mid) for the Singh-Maddala scenario. These simulation results disqualify both estimation methods for use in practical applications. Regarding the precision, the conclusions from the previous section are transferable.

Table 2.4: Relative bias (rB) and the empirical standard error (se.emp) for the different estimation methods estimated for a selection of statistical indicators.

		$Q_{0.1}$	$Q_{0.25}$	Median	$Q_{0.75}$	$Q_{0.9}$	Mean	HCR	QSR	PGap	Gini
Quality Measure	Estimation Method	**Dagum: 8 intervals**									
rB	True	0.041	-0.014	0.020	0.003	0.005	0.015	0.032	0.072	0.036	0.028
	KDE	0.192	0.088	-0.146	0.225	0.038	-0.396	-0.126	-0.770	-0.084	-0.851
	Uni	-0.977	-1.719	0.675	3.150	2.883	5.454	2.579	16.532	4.163	9.840
	Mid	-23.304	-12.787	-2.552	3.227	2.420	12.042	29.230	209.641	-2.171	16.251
se.emp	True	399.449	437.440	455.249	584.052	988.153	442.182	0.004	0.128	0.002	0.003
	KDE	382.632	422.677	440.771	567.565	964.208	479.943	0.004	0.135	0.002	0.003
	Uni	459.406	461.163	456.904	645.016	1052.903	613.491	0.004	0.171	0.002	0.004
	Mid	0.000	0.000	0.000	0.000	0.000	826.842	0.005	1.024	0.002	0.005
		Singh-Maddala: 8 intervals									
rB	True	-0.070	0.001	0.035	0.014	-0.015	0.003	0.023	0.017	0.041	-0.006
	KDE	0.270	0.014	0.042	-0.039	-0.031	0.093	-0.039	0.714	0.085	0.213
	Uni	-1.031	-1.210	1.652	2.963	2.039	6.269	1.800	18.504	4.321	11.024
	Mid	-21.083	-11.797	-1.789	3.039	1.636	12.618	27.516	199.584	-1.651	17.009
se.emp	True	416.957	486.609	555.653	731.369	1049.186	443.818	0.004	0.099	0.002	0.002
	KDE	389.926	447.684	546.007	698.835	998.289	462.384	0.004	0.106	0.002	0.002
	Uni	467.696	502.097	547.127	784.601	1072.791	598.248	0.004	0.145	0.002	0.003
	Mid	784.213	0.000	0.000	0.000	0.000	784.707	0.005	0.930	0.002	0.004

As given in Table 2.5, the estimated standard errors offer a good approximation of the empirical standard errors for the different scenarios.

2.3.3 Equal and ascending interval width

While the German (Statistisches Bundesamt, 2017), the Australian (Australian Bureau of Statistics, 2011), the Colombian (Departamento Administrativo Nacional De Estadística, 2005), and the census from New Zealand (Statistics New Zealand, 2013) use ascending class width, previous research shows that the performance of alternative estimation methods depends on the interval width (Lenau and Münnich, 2016). More precisely, performance depends on whether the data is censored to intervals of equal width or ascending width. Therefore, the GB2 distribution from Table 2.3 is now censored to eight intervals with equal class width (except the last interval, which has an open-ended upper interval bound). In all previous simulation scenarios ascending interval width is used. Figure 2.3 shows the censored GB2 distribution. The theoretical distribution is kept fixed in order to evaluate the influence of the censoring on the performance.

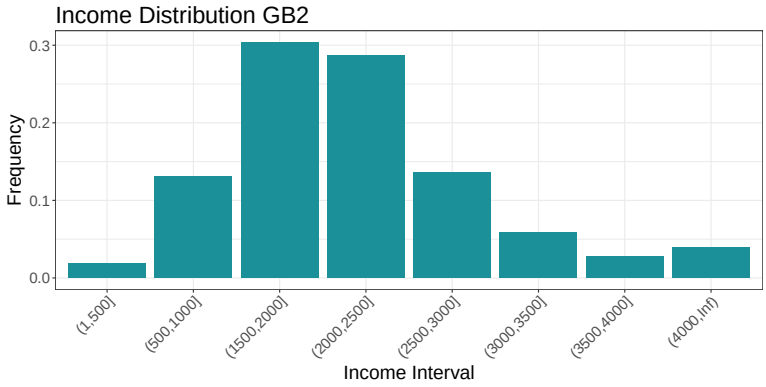


Figure 2.3: GB2 distribution censored to equally sized intervals (except the last – open-ended – interval).

The results of the point estimates are given in Table 2.6. As before, using the uncensored data leads to unbiased estimates. The estimates obtained by the KDE algorithm are unbiased except for the QSR, PGap, and Gini coefficient. These estimates exhibit a very small bias of -1.7%, 1.4% and -2.2%. However, the results are comparable to the estimation results from the GB2 scenario with eight intervals with ascending interval width. Hence, the KDE algorithm does not seem to be affected by the censoring scheme. The benchmark indicators Uni and Mid show, as before, large biases especially for indicators that rely on the whole shape of the distribution. With regard to precision, the results and interpretation are the same as before.

The proposed bootstrap also gives valid results with equal-sized intervals (see Table 2.7).

2.3.4 Conclusion and final remarks

The simulation results show that the KDE algorithm outperforms other approaches (Uni and Mid) in terms of bias in all scenarios. The KDE method gives unbiased results under differ-

Table 2.5: Empirical and estimated standard error for the selected statistical indicators.

		$Q_{0.1}$	$Q_{0.25}$	Median	$Q_{0.75}$	$Q_{0.9}$	Mean	HCR	QSR	PGap	Gini
Measure	Estimation Method	**Dagum: 8 intervals**									
se.emp	KDE	382.632	422.677	440.771	567.565	964.208	479.943	0.004	0.135	0.002	0.003
se.est		385.340	420.523	445.765	573.573	953.225	468.896	0.004	0.134	0.002	0.003
		Singh-Maddala: 8 intervals									
se.emp	KDE	389.926	447.684	546.007	698.835	998.289	462.384	0.004	0.106	0.002	0.002
se.est		386.539	430.594	523.137	691.090	983.671	460.726	0.004	0.110	0.002	0.002

Table 2.6: Relative bias (rB) and the empirical standard error (se.emp) for the different estimation methods estimated for a selection of statistical indicators.

		$Q_{0.1}$	$Q_{0.25}$	Median	$Q_{0.75}$	$Q_{0.9}$	Mean	HCR	QSR	PGap	Gini
Quality Measure	Estimation Method	**GB2: 8 intervals (equally sized)**									
rB	True	0.079	0.035	0.013	-0.026	-0.092	-0.024	-0.127	-0.144	-0.248	-0.110
	KDE	-0.005	-0.422	0.238	-0.066	0.050	-0.840	0.290	-1.706	1.370	-2.181
	Uni	-7.074	-2.388	0.909	0.560	1.704	4.648	7.351	21.365	30.052	16.251
	Mid	-14.151	4.640	11.598	10.730	3.174	12.498	19.720	73.226	28.594	30.467
se.emp	True	88.841	75.061	72.038	111.139	233.943	95.398	0.003	0.051	0.001	0.003
	KDE	86.255	70.621	76.142	109.506	223.898	128.012	0.003	0.076	0.001	0.005
	Uni	116.469	70.955	88.391	156.503	260.393	130.810	0.003	0.076	0.001	0.004
	Mid	0.000	0.000	0.000	544.426	0.000	180.793	0.004	0.281	0.001	0.005

Table 2.7: Empirical and estimated standard error for the selected statistical indicators.

		$Q_{0.1}$	$Q_{0.25}$	Median	$Q_{0.75}$	$Q_{0.9}$	Mean	HCR	QSR	PGap	Gini
Measure	Estimation Method	**GB2: 8 intervals (equally sized)**									
se.emp	KDE	86.255	70.621	76.142	109.506	223.898	128.012	0.003	0.076	0.001	0.005
se.est		84.456	67.507	75.134	108.778	224.587	138.326	0.003	0.079	0.001	0.005

ent censoring schemes and for different underlying theoretical distributions. The relative bias increases slightly whenever the number of intervals decreases. However, also in very extreme censoring scenarios (with only eight intervals), the results are very precise. The relative bias is under 1% for almost all indicators. The KDE method shows comparable results in terms of precision to the direct estimation of the indicators from the continuous uncensored data. Additionally, it is superior to other approaches (Mid and Uni) that show worse precision for most indicators. Due to its easy usage, its ability to adapt to different underlying theoretical distributions and different censoring schemes and its precision practitioners should prefer the KDE algorithm to other approaches.

The KDE algorithm cannot handle open-ended intervals. As mentioned before, lower bounds equal to $-\infty$ or upper bounds equal to $+\infty$ have to be replaced by a finite number. The chosen value effects the performance of the KDE algorithm. However, not all poverty and inequality indicators depend on the outer intervals. Indicators that depend on the outer intervals are indicators that depend, by their definition, on the whole distribution e.g., the mean or the Gini coefficient. These indicators are always influenced by the way in which open-ended intervals are handled, whereas other indicators, such as the median, are only affected if they fall into one of the open-ended outer intervals. The replacement value used for open-ended upper and lower intervals also has an impact on the performance of the methods Uni and Mid. To make simulation results from the different estimation methods comparable to each other, we replace $+\infty$ of the upper interval with a value of three times the value of the lower bound. For instance, if the interval is $(4000, +\infty)$ we replace the upper bound with $4000 * 3 = 12000$, resulting in the interval $(4000, 12000]$ which is used by the KDE algorithm. In an application the practitioner should choose the interval bounds for open-ended intervals with caution, with regard to content and to the censoring scheme. However, our experiences running several simulations indicate that a value of three times the value of the lower bound serves as a good approximation when working with interval-censored income data.

2.4 Estimating poverty and inequality indicators from the German Microcensus

In this section, the KDE algorithm is applied to the problem of estimating poverty and inequality indicators from interval-censored German Microcensus data. The relevance of poverty and inequality estimation becomes apparent when considering the rich amount of literature available on this topic. Germany's increasing inequality has sparked the interest of many scholars and governmental institutions. Known for stable wages in the 70s and 80s (Abraham and Houseman, 1995), Germany has faced growing income inequality since its reunification in 1990 (Fuchs-Schündeln et al., 2010; Bönke et al., 2014).

Most of these studies consider or focus on survey data such as the Socio-Economic Panel (SOEP) or the Income, Receipts, Expenditure survey (in German: Einkommens- und Verbrauchsstichprobe) (EVS). In contrast to the Microcensus, participation is voluntary and participants are asked for their exact income (not interval censored), which enables the estimation

of poverty and inequality indicators using standard formulas. However, since the German Microcensus is by far the biggest survey in Germany it would be favorable to use its data for the estimation of poverty and inequality. The proposed KDE algorithm makes the valid and precise estimation of complex poverty and inequality indicator from interval-censored data possible. This allows researchers and practitioners to use the German Microcensus for the further and more in-depth investigation of the increasing income inequality in Germany. The following application presents estimation results from cross-sectional data for the year 2012. To investigate the spatial distribution of inequality, the different indicators are estimated for the 16 federal states.

2.4.1 Data and preparation

The German Microcensus is a representative household survey conducted by the Federal Statistical Office of Germany. About 1% of the German population is chosen randomly by a specified survey design and is asked about the living conditions. The Microcensus was first conducted in 1957 and provides data regarding the structure and the economic and social status of the population. Over the years the Microcensus has become one of the most important data sources regarding aspects such as partnership, family, labor market, and eduction. For the estimation of poverty and inequality the variable household net income is used. For the analysis the Scientific-Use-File (SUF), a 70% sample of the Microcensus is used (Statistisches Bundesamt, 2017). After data cleaning, we are left with a sample size of $n_{Germany} = 454852$. Since interests also lie in the spatial distribution of poverty and inequality the statistical indicators are estimated for each federal state separately and for Germany. The sample size for each federal state and its location is given in Table 2.8 and Figure 2.7 in Appendix 2.6. The sample sizes are very large for each federal state even for Bremen, the state with the smallest sample size $n_{Bremen} = 3356$. Thus, there are enough observations to directly (without covariates) estimate the statistical indicators with small standard errors. As previously mentioned, the variable household income is interval censored to 24 intervals. The distribution is visualized in Figure 2.1 in the upper-left panel. To make the household income comparable between households of different sizes, the OECD household weights are used to estimate equivalized household income. Each household's interval bound is divided by its corresponding OECD weight. For instance, a household within interval $(1300, 1500]$ and with an OECD weight of 1.5 has equivalence interval bounds of $(867, 1000]$.

2.4.2 Estimation and results

In order to estimate the poverty and inequality indicators, the KDE algorithm is applied to the equivalenced interval bounds. The open-ended interval is handled as described in Section 2.3.4. Furthermore, for representative results the extrapolation factors of the Microcensus are used for the estimation of the weighted statistical indicators (formulas are given in Equation (2.1)-(2.6)). Therefore, the KDE algorithm draws iteratively new metric pseudo samples plus the corresponding extrapolation weight from the equivalenced interval-censored household in-

come. As in the simulations, the number of burn-in iterations is $B_{(KDE)} = 80$, the number of additional iterations is $S_{(KDE)} = 400$ and the number of grid points $j = 4000$. The number of $B_{(KDE)}$ and $S_{(KDE)}$ is sufficiently large as is seen in the convergence plot in Figure 2.4. Both indicators converge after 480 iterations. While indicators that are dependent on the whole distribution converge slower (e.g., the Gini coefficient), indicators that do not depend on the whole distribution (e.g., the HCR) converge faster. Also, all other indicators are checked for convergence, but only two plots are shown exemplarily. The standard errors of the weighted indicators are estimated by the described non-parametric bootstrap as proposed by Alfons and Templ (2013). Differently than described in Section 2.2.2 not only the interval censored observations are drawn but also the corresponding weights. The number of bootstrap samples is set to $B = 100$ as in the simulation study.

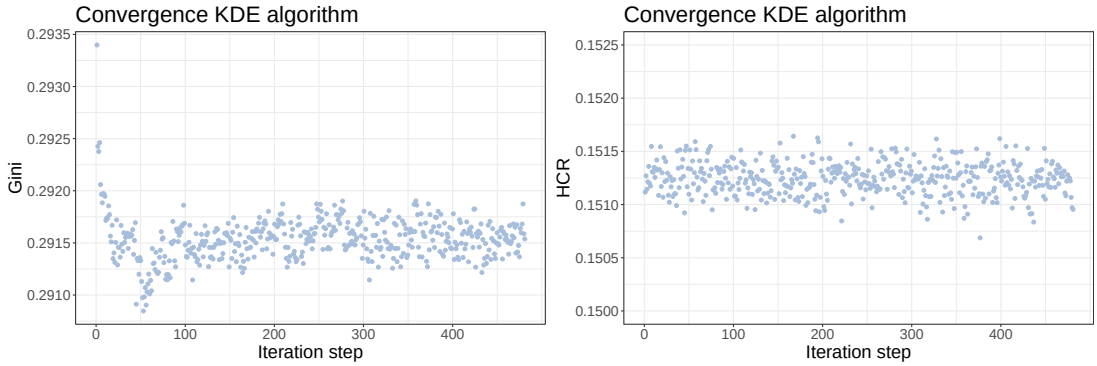
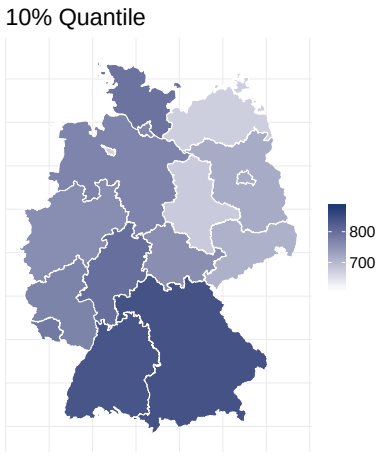


Figure 2.4: Convergence of the KDE algorithm for the Gini coefficient and the HCR.

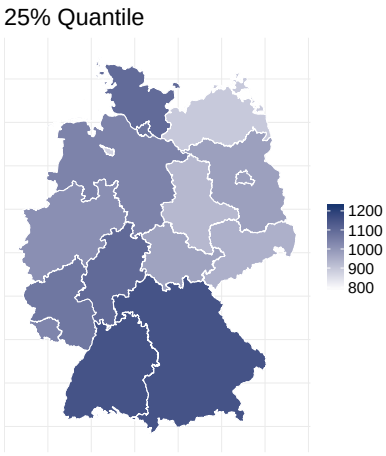
The estimated indicators are presented in Figure 2.5 and 2.6 and the exact values and the estimated standard errors are given in Appendix 2.6 in Table 2.9. The estimated HCR $= 0.15$, the Gini $= 0.29$ and the QSR $= 4.31$. These results are comparable to the results from the EVS. The EVS reports the following values: HCR $= 0.16$, the Gini $= 0.27$ and the QSR $= 4.1$ (Statistisches Bundesamt, 2018c). Owing to the large sample size, valid estimates for smaller geographical areas can be estimated to evaluate the regional distribution of poverty and inequality in Germany. The quantiles and the mean indicate that the East (formerly German Democratic Republic DDR) is poorer than the West. This result is commonly known in Germany and is not very surprising. Nevertheless, Brandenburg and Berlin have higher incomes than the rest of East Germany (Mecklenburg-Vorpommern, Saxony, Saxony-Anhalt and Thuringia). Also Bremen, a federal state in the West, shows low income for the 10% and 25% quantile in comparison to the rest of West Germany, while for the higher quantiles Bremen shows similar results to the rest of Germany. The poorest states with a median of 1,211.29 Euro and 1,247.05 Euro are Mecklenburg-Vorpommern and Saxony-Anhalt and the richest ones with a median of 1,580.43 Euro and 1,580.35 Euro are Baden-Württemberg and Bavaria. For the estimation of the HCR and PGap, a regional poverty line (60% of the median) is used. The HCR indicates that in the East fewer people live under the regional poverty line than in the West. Also, the people living under the poverty line live closer to it in the East, as shown by the PGap. When looking at the QSR and the Gini coefficient, the East-West trend is less striving. Nevertheless,

the states in the East have lesser income inequality. The most unequal states with a Gini coefficient of 0.32 and 0.31 are Hamburg and Bremen and the most equal ones with a Gini coefficient of 0.25 and 0.25 are Saxony and Thuringia. The estimated standard errors of the indicators on state areas are quite small. Therefore, estimating precise indicators for smaller geographical areas would probably also be possible, in order to get an even closer look at the geographical distribution of poverty and inequality.

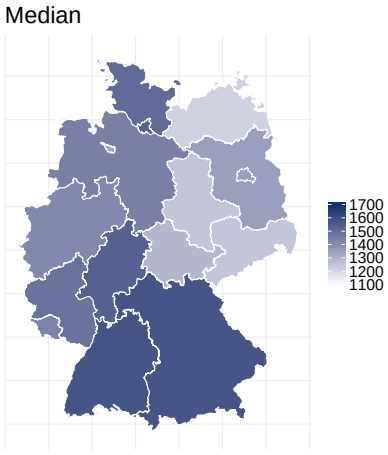
The application impressively demonstrates how the KDE algorithm enables the estimation of poverty and inequality indicators from interval-censored data. The precise estimations obtained by the KDE algorithm enable statisticians and statistical offices to report a variety of poverty and inequality indicators using the German Microcensus. The regional estimates will help to identify regions with lower income and higher inequality to target political activities more accurately for those in need.



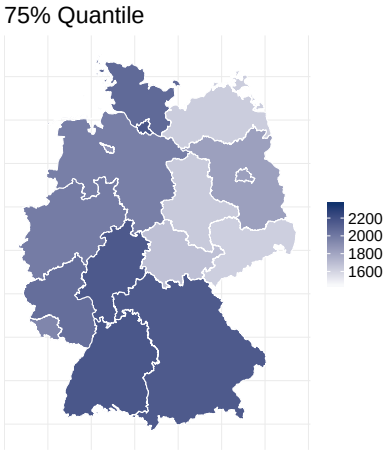
(a) Regional distribution of the 10% quantile.



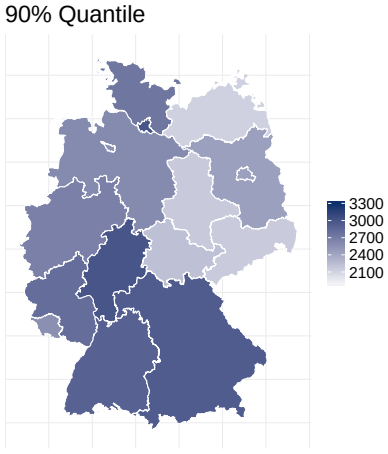
(b) Regional distribution of the 25% quantile.



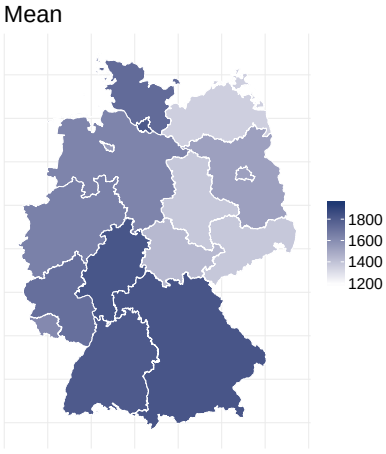
(c) Regional distribution of the median.



(d) Regional distribution of the 75% quantile.

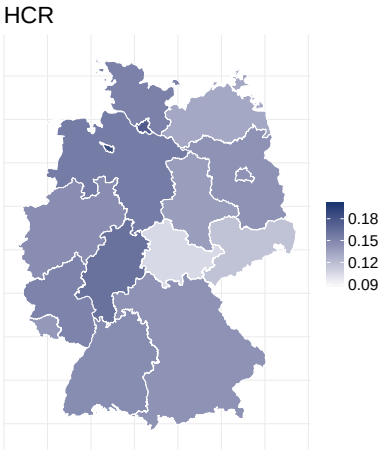


(e) Regional distribution of the 90% quantile.

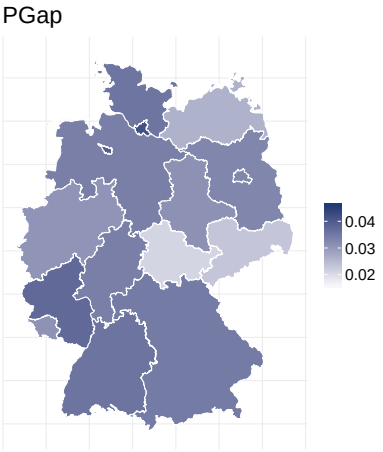


(f) Regional distribution of the mean.

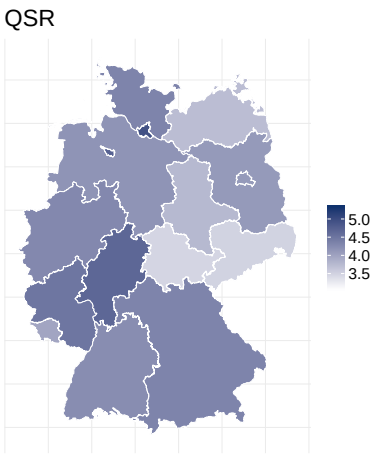
Figure 2.5: Regional distribution of different statistical indicators in Germany.



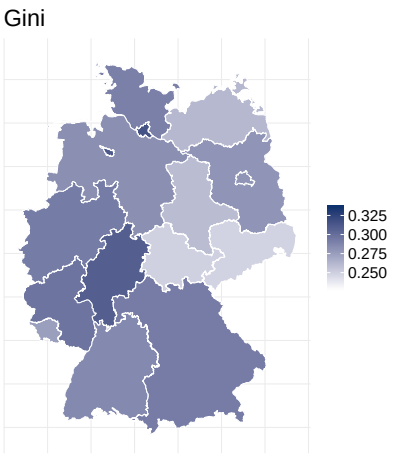
(a) Regional distribution of the HCR.



(b) Regional distribution of the PGap.



(c) Regional distribution of the QSR.



(d) Regional distribution of the Gini coefficient.

Figure 2.6: Regional distribution of different statistical indicators in Germany.

2.5 Discussion and outlook

In numerous censuses e.g., the German Microcensus or the Australian census, the variable household or personal income is not observed on a continuous scale, but is rather censored to specific intervals. This is due to confidentiality constraints or to reduce item non-response. Estimating poverty and inequality indicators from these kinds of data requires more sophisticated statistical methods. As an estimation method we propose an iterative KDE algorithm that enables the precise estimation of statistical indicators from interval-censored data. The proposed KDE algorithm has similarities to SEM algorithms that are commonly used for the estimation of models that depend on latent unobserved variables (in our case the interval-censored income). However, instead of maximizing the likelihood as is common for SEM algorithms, the asymptotic mean integrated squared error of the KDE is maximized. For the estimation of the standard errors of the statistical indicators a non-parametric bootstrap is proposed. The KDE algorithm and the bootstrap work for different censoring scenarios and different underlying true distributions. The methodology is available in the R package `smicd` from the Comprehensive R Archive Network (Walter, 2018). Our simulation results demonstrate that the estimated poverty and inequality indicators outperform other estimation techniques (linear interpolation or the use of the midpoints of the intervals) in terms of bias. Also, the standard errors of the estimates are close to the standard errors from the estimates that were obtained with the uncensored data, supporting the precision of the algorithm. Furthermore, the KDE algorithm has the advantage of adapting to different interval-censored theoretical distributions. Therefore, it is universally applicable for the estimation of poverty and inequality indicators from interval-censored income data. We demonstrate the usefulness by estimating regional poverty and inequality indicators from the German Microcensus. To get representative results the algorithm is extended to take OECD equivalence weights and survey weights into account. The estimated regional indicators are plotted on maps that visualize the magnitude of poverty and inequality in Germany. With the help of the KDE algorithm statistical indicators can be precisely estimated from interval-censored data in order to tackle the increasing problem of rising poverty and inequality in societies all over the world.

Further research should focus on convergence criteria that make the manual choice of the number of iteration obsolete.

Acknowledgements

We thank Timo Schmid and Marcus Groß for discussions and helpful comments on this paper.

2.6 Appendix

Table 2.8: Sample size for Germany and each of the 16 federal states.

State	Sample size	Number in Map
Germany	454852	
Schleswig-Holstein	15302	1
Hamburg	8630	2
Lower Saxony	45828	3
Bremen	3356	4
North Rhine-Westphalia	90778	5
Hesse	35730	6
Rhineland-Palatinate	21229	7
Baden-Württemberg	58685	8
Bavaria	75244	9
Saarland	5688	10
Berlin	19311	11
Brandenburg	15400	12
Mecklenburg-Vorpommern	8706	13
Saxony	24609	14
Saxony-Anhalt	13495	15
Thuringia	12861	16

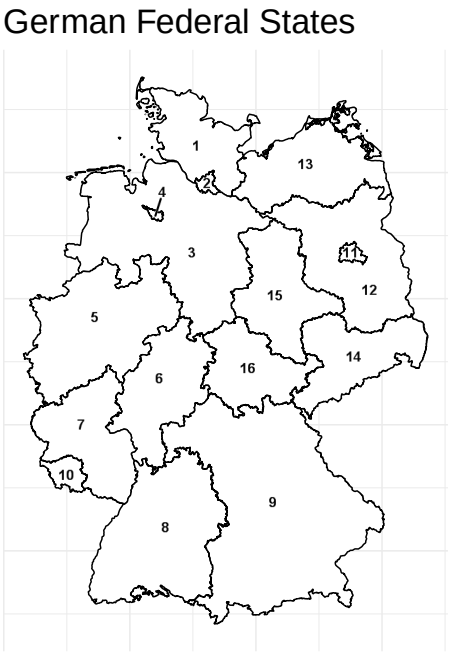


Figure 2.7: German Federal States, the names of the corresponding numbers are given in Table 2.8.

Table 2.9: Estimated statistical indicators for Germany and the 16 federal states. Standard errors are given in parentheses.

	Quant0.1	Quant0.25	Median	Quant0.75	Quant0.9	Mean	HCR	QSR	PGap	Gini
Germany	770.16	1040.23	1445.53	1998.96	2714.63	1675.88	0.15	4.31	0.03	0.29
	(0.00)	(3.28)	(4.93)	(2.59)	(3.69)	(1.85)	(0.00)	(0.01)	(0.00)	(0.00)
Schleswig-Holstein	794.21	1092.99	1512.96	2071.11	2743.79	1736.25	0.15	4.33	0.04	0.29
	(6.04)	(5.83)	(7.39)	(7.80)	(20.01)	(9.35)	(0.00)	(0.05)	(0.00)	(0.00)
Hamburg	765.68	1069.24	1540.20	2166.83	3002.79	1815.45	0.17	4.92	0.04	0.32
	(7.16)	(9.21)	(10.47)	(13.44)	(29.75)	(14.14)	(0.00)	(0.09)	(0.00)	(0.00)
Lower Saxony	770.08	1040.25	1445.04	1970.84	2603.04	1636.36	0.16	4.16	0.03	0.28
	(4.10)	(2.53)	(5.44)	(7.11)	(13.00)	(5.79)	(0.00)	(0.04)	(0.00)	(0.00)
Bremen	665.44	876.91	1328.23	1879.66	2564.10	1540.03	0.18	4.72	0.04	0.31
	(10.82)	(9.93)	(16.76)	(25.28)	(47.95)	(19.19)	(0.01)	(0.12)	(0.00)	(0.01)
North Rhine-Westphalia	756.85	1013.41	1418.50	1985.64	2674.27	1649.22	0.15	4.29	0.03	0.29
	(0.72)	(0.01)	(3.01)	(4.70)	(9.05)	(3.69)	(0.00)	(0.02)	(0.00)	(0.00)
Hesse	798.23	1094.75	1540.36	2149.61	2997.03	1825.06	0.16	4.66	0.03	0.31
	(4.56)	(4.88)	(6.03)	(7.49)	(14.23)	(7.54)	(0.00)	(0.04)	(0.00)	(0.00)
Rhineland-Palatinate	770.65	1067.23	1485.95	2052.43	2810.09	1720.30	0.15	4.49	0.04	0.30
	(5.39)	(5.36)	(6.36)	(8.97)	(17.17)	(8.40)	(0.00)	(0.06)	(0.00)	(0.00)
Baden-Württemberg	837.76	1148.33	1580.43	2160.84	2900.98	1806.40	0.15	4.24	0.04	0.29
	(2.23)	(4.10)	(4.08)	(6.62)	(10.50)	(5.52)	(0.00)	(0.03)	(0.00)	(0.00)
Bavaria	841.94	1148.28	1580.35	2147.52	2944.04	1826.62	0.14	4.33	0.03	0.29
	(5.37)	(4.62)	(4.99)	(5.19)	(8.81)	(4.75)	(0.00)	(0.03)	(0.00)	(0.00)
Saarland	784.70	1035.41	1434.20	1938.86	2559.91	1615.95	0.14	3.98	0.03	0.27
	(8.77)	(9.73)	(10.55)	(14.70)	(32.54)	(13.40)	(0.01)	(0.07)	(0.00)	(0.00)
Berlin	730.96	912.14	1328.50	1867.05	2552.45	1547.47	0.15	4.15	0.02	0.29
	(5.38)	(5.19)	(8.41)	(11.20)	(15.87)	(8.77)	(0.01)	(0.05)	(0.00)	(0.00)
Brandenburg	716.80	979.24	1351.02	1823.73	2446.72	1528.25	0.14	4.10	0.03	0.28
	(5.93)	(6.78)	(6.43)	(10.03)	(17.36)	(8.08)	(0.00)	(0.05)	(0.00)	(0.00)
Mecklenburg-Vorpommern	671.30	895.52	1211.29	1629.61	2120.56	1355.61	0.13	3.74	0.03	0.26
	(4.92)	(5.77)	(7.51)	(9.78)	(20.77)	(11.96)	(0.00)	(0.08)	(0.00)	(0.01)
Saxony	709.57	945.88	1247.40	1622.64	2155.08	1383.20	0.12	3.52	0.02	0.25
	(4.62)	(4.00)	(4.31)	(5.48)	(10.93)	(5.09)	(0.00)	(0.03)	(0.00)	(0.00)
Saxony-Anhalt	675.70	928.47	1247.05	1643.78	2161.33	1382.23	0.14	3.78	0.03	0.26
	(5.25)	(5.85)	(5.90)	(7.68)	(17.09)	(6.24)	(0.00)	(0.05)	(0.00)	(0.00)
Thuringia	755.17	973.23	1283.80	1683.44	2226.40	1435.52	0.11	3.50	0.02	0.25
	(5.32)	(4.23)	(5.50)	(7.11)	(16.00)	(7.45)	(0.00)	(0.04)	(0.00)	(0.00)

Chapter 3

Small Area Estimation with Interval-Censored Income Data

3.1 Introduction

Recent applications of small area estimation (SAE) methodologies have been concerned with the estimation of area-specific income indicators, for example the median income, the head count ratio and the quintile share ratio (Rao and Molina, 2015; Rojas-Perilla et al., 2017; Tzavidis et al., 2018). Popular SAE methods that have been used in this context include the so-called World Bank method (Elbers et al., 2003) and the empirical best predictor (EBP) method (Molina and Rao, 2010). In these papers, small area estimation is based on the use of a unit-level nested error regression (random effects) model estimated with income as a response variable that is measured on a continuous scale.

It is tempting for survey designers to reduce survey related costs by collecting information on income using income bands as opposed to detailed income information (Micklewright and Schnepf, 2010). Collecting data in bands may also help with reducing respondent burden, item non-response and micro-data disclosure risk. On the other hand, it is also reasonable to expect that collecting interval-censored data may result in a loss of information compared to collecting on a continuous scale. The impact of this loss of information on the quality of official statistics estimates is of particular importance. Interval-censored household income data are collected as part of the German Microcensus (Statistisches Bundesamt, 2017). In the UK, the Office for National Statistics experimented with the collection of interval-censored income data in the lead up to the 2001 census (Collins and White, 1996). In this paper the terms grouped data, banded data, and interval-censored data are used interchangeably.

Although regression methods for grouped data have been studied in the econometric literature (Hsiao, 1983) to the best of our knowledge this is not the case with random effects models. Small area estimation is therefore challenging when the analyst only has access to a grouped response variable. The present paper proposes an extension of the EBP method when the response variable is banded. The methodology works by reversing the process of banding, leading to an outcome measured on a continuous scale. Estimation of the parameters

of the unit-level nested error regression model is implemented via a stochastic expectation-maximization (SEM) algorithm (Celeux and Dieboldt, 1985). The estimated model parameters are then used for small area prediction. The proposed methodology also allows for the use of data-driven transformations when the error term of the model lack normality. Following González-Manteiga et al. (2008), the estimation of the mean squared error (MSE) of the small area estimates is facilitated by a parametric bootstrap, which incorporates the additional uncertainty due to interval censoring of the response variable assuming that the censoring mechanism is known. The proposed method assumes that there is no measurement error in reporting the band associated with the latent continuous variable. In this paper we develop the methodology under a 2-level nested error regression model but an extension to 3-level structures – incorporating possible cluster effects – along the lines of the methodology proposed by Marhuenda et al. (2017) is feasible. Finally, as it is the case with the EBP method or the World Bank method, we assume access to micro-data for the model covariates from census or administrative data. The proposed methodology makes possible the use of SAE methods with interval-censored outcomes and therefore it enables survey organizations to consider collecting data in this form.

The paper is organized as follows. Section 3.2 introduces the SEM algorithm that is used for the estimation of the regression parameters of a 2-level nested error regression model. In Section 3.3, the EBP method with interval-censored data is presented. In Section 3.4 extensive model-based simulations are carried out. In Section 3.5, the new statistical methodology is used to estimate poverty and inequality indicators from interval-censored data from Mexico. Finally, the main results are summarized and discussed in Section 3.6.

3.2 The nested error linear regression model with an interval-censored response variable

Consider a finite population U of size N, divided into D areas/domains. The terms areas and domains are used interchangeably in this paper. The population size of each of the D-domains $U_1, U_2, \ldots, U_D$ is given by $N_1, N_2, \ldots, N_D$. Let us for now assume that the response variable denoted by y_{ij} is measured on a continuous scale, where $j = 1, 2, \ldots, n_i$ denotes the jth unit belonging to the ith domain, with $i = 1, 2, \ldots, D$. The vector x is defined as $x_{ij}^T = (x_{1ij}, \ldots, x_{pij})$, where p denotes the number of explanatory variables. A nested error linear regression model is used for modeling the relationship between the variable of interest and auxiliary information with the unexplained variation being captured by the random effect term, u_i and the residuals e_{ij}. In the simplest case a 2-level nested error regression model as defined in Battese et al. (1988) is given by

$$\begin{aligned} y_{ij} &= x_{ij}^T\beta + u_i + e_{ij}, \qquad (j = 1, \ldots, n_i), \qquad (i = 1, \ldots, D), \\ u_i &\overset{iid}{\sim} N(0, \sigma_u^2), \\ e_{ij} &\overset{iid}{\sim} N(0, \sigma_e^2), \\ y_{ij}|x_{ij}, u_i &\sim N(x_{ij}^T\beta + u_i, \sigma_e^2). \end{aligned} \tag{3.1}$$

In the case of interval-censored data, y_{ij} is unobserved and the only observed information concerning the dependent variable is, that it falls within an interval. The continuous scale is divided into K intervals, where the k-th interval is given by (A_{k-1}, A_k). The variable k_{ij} $(1 \leq k_{ij} \leq K)$ indicates in which of the intervals the dependent variable falls into. The first and K-th interval are allowed to be open ended, therefore $A_0 = -\infty$ and $A_K = +\infty$ are possible. Situations in which either or none of the outer intervals are open ended can also be handled by the proposed methodology. Furthermore, the interval length is allowed to be arbitrary and can vary between intervals. Since the underlying distribution of y_{ij} is unknown, the aim is to reconstruct the conditional distribution $f(y_{ij}|x_{ij}, k_{ij}, u_i, \theta)$, where $\theta = (\beta, \sigma_e^2, \sigma_u^2)$ are the unknown model parameters, β is a $p \times 1$ vector of regressors and the error terms u_i and e_{ij} are assumed to be independent and normally distributed. Estimation methods such as maximum likelihood (ML) or restricted maximum likelihood (REML) are utilized for estimating θ when y_{ij} is observed on a continuous scale (Lindstrom and Bates, 1990). However, when the response variable is interval censored, formulating the likelihood is more challenging. In this section an SEM algorithm for fitting the model is proposed and data-driven transformations are also considered for handling potential departures from the model assumptions.

Different approaches for dealing with interval-censored response variables in regression modeling that assume independent observations have been proposed in the literature. A naive approach uses ordinary least squares on the midpoints of the intervals. While this approach is easy to implement (Thompson and Nelson, 2003), it has two major drawbacks. The uncertainty associated with the value of each observation within each interval is not accounted for and dealing with open ended intervals is not easy. Simulation results demonstrate, that the grouping coarseness affects the quality of the estimates hence necessitating more advanced estimation methods (Cameron, 1987). Nevertheless, the naive approach can provide results of acceptable quality if the grouping is very fine (Fryer and Pethybridge, 1972). An alternative approach is to view the outcome as ordinal and use an ordered probit or logit regression model (McCullagh, 1980). However, since the predicted values are then expressed in terms of the probability of belonging in each interval, these models cannot be used for predicting the unknown value of the response variable on the continuous scale.

To overcome these drawbacks, linear regression models for left-censored (Tobin, 1958), right-censored (Rosett and Nelson, 1975) and grouped (or interval-censored) (Stewart, 1983) dependent variables have been proposed. Stewart (1983) proposes an expectation-maximization (EM) algorithm for estimating the model parameters of a linear regression model with grouped response variable.

For the nested error regression model we were unable to find relevant literature. Therefore, in this paper we propose a SEM algorithm (Celeux and Dieboldt, 1985; Celeux et al., 1996) for estimating the parameters of the nested error regression model when the outcome is interval censored. A similar SEM algorithm is proposed in Groß et al. (2017) for kernel density estimation on aggregated data.

To reconstruct the unknown distribution $f(y_{ij}|x_{ij}, k_{ij}, u_i, \theta)$ we use the Bayes theorem and

express the target distribution as follows:

$$f(y_{ij}|x_{ij}, k_{ij}, u_i, \theta) = \frac{f(k_{ij}|y_{ij}, x_{ij}, u_i, \theta) f(y_{ij}|x_{ij}, u_i, \theta)}{f(k_{ij}|x_{ij}, u_i, \theta)}$$
$$\propto f(k_{ij}|y_{ij}, x_{ij}, u_i, \theta) f(y_{ij}|x_{ij}, u_i, \theta).$$

Since $f(k_{ij}|y_{ij}, x_{ij}, u_i, \theta) = f(k_{ij}|y_{ij})$, the conditional distribution of k_{ij} is given by

$$f(k_{ij}|y_{ij}) = \begin{cases} 1 & \text{if } A_{k-1} \leq y_{ij} \leq A_k, \\ 0 & \text{else,} \end{cases}$$

and under (3.1),

$$f(y_{ij}|x_{ij}, u_i, \theta) \sim \quad N(x_{ij}^T \beta + u_i, \sigma_e^2).$$

Because y_{ij} is unobserved, one approach to fitting the model defined above is to use the SEM algorithm. The algorithm works by replacing the unobserved data in the complete data likelihood by generating pseudo samples given the observed data and the current values of θ (S-step) and then maximizes the complete data likelihood for updating θ (M-step). The updated vector of model parameters is used for generating pseudo samples $\tilde{y}_{ij}$ of the unknown y_{ij} from the conditional distribution $f(y_{ij}|x_{ij}, k_{ij}, u_i, \theta)$. The iterations stop after $B + M$ steps.

3.2.1 The SEM algorithm

Assuming θ is known, pseudo samples, $\tilde{y}_{ij}$, are drawn from the following conditional distribution

$$f(y_{ij}|x_{ij}, k_{ij}, u_i, \theta) \propto \mathbf{I}(A_{k-1} \leq y_{ij} \leq A_k) \times N(x_{ij}^T \beta + u_i, \sigma_e^2),$$

where $\mathbf{I}(\cdot)$ denotes the indicator function. The conditional distribution of y_{ij} has the form of a two sided truncated normal distribution given by

$$f(y_{ij}|x_{ij}, k_{ij}, u_i, \theta) = \frac{\phi\left(\frac{y_{ij} - \mu_{ij}}{\sigma_e}\right)}{\sigma_e \left(\Phi\left(\frac{A_k - \mu_{ij}}{\sigma_e}\right) - \Phi\left(\frac{A_{k-1} - \mu_{ij}}{\sigma_e}\right)\right)},$$

with $\mu_{ij} = x_{ij}^T \beta + u_i$, $\phi(\cdot)$ is the probability density function of the standard normal distribution and $\Phi(\cdot)$ is its cumulative distribution function. By definition $\Phi\left(\frac{A_k - \mu_{ij}}{\sigma_e}\right) = 1$ if $A_k = +\infty$ and $\Phi\left(\frac{A_{k-1} - \mu_{ij}}{\sigma_e}\right) = 0$ if $A_{k-1} = -\infty$. For each observation with explanatory variables x_{ij} the corresponding $\tilde{y}_{ij}$ is randomly drawn from $N(x_{ij}^T \beta + u_i, \sigma_e^2)$ within the given interval $(A_{k-1} \leq y_{ij} \leq A_k)$. This is the S-step of the SEM algorithm. The M-step comprises fitting the nested error regression model using the newly generated $(\tilde{y}_{ij}, x_{ij})$. The steps of the SEM algorithm are as follows:

1. Estimate $\hat{\theta} = (\hat{\beta}, \hat{\sigma}_e^2, \hat{\sigma}_u^2)$ from (3.1) using the midpoints of the intervals as a substitute for the unknown y_{ij}. The parameters are estimated using restricted maximum likelihood

(REML).

2. **S-step:** For $j = 1, \ldots, n_i$ and $i = 1, \ldots, D$ sample from the conditional distribution $f(y_{ij}|x_{ij}, k_{ij}, u_i, \theta)$ by drawing randomly from $N(x_{ij}^T\hat{\beta} + \hat{u}_i, \hat{\sigma}_e^2)$ within the given interval $(A_{k-1} \leq y_{ij} \leq A_k)$ obtaining $(\tilde{y}_{ij}, x_{ij})$. The drawn pseudo $\tilde{y}_{ij}$ are used as replacement for the unknown y_{ij}.

3. **M-step:** Re-estimate the vector $\hat{\theta}$ using (3.1) and the pseudo samples $(\tilde{y}_{ij}, x_{ij})$ from Step 2. The parameters are estimated using REML as in Step 1.

4. Iterate Steps 2-3 $B + M$ times, with B burn-in iterations and M additional iterations.

5. Discard the burn-in iterations and estimate $\hat{\theta}$ by averaging the derived M estimates.

For open ended intervals $A_0 = -\infty$ and $A_K = +\infty$, the midpoints M_1 and M_K in Step 1 are computed as follows:

$$M_1 = (A_1 - \overline{D})/2,$$
$$M_K = (A_{K-1} + \overline{D})/2,$$

where

$$\overline{D} = \frac{1}{(K-2)} \sum_{k=2}^{K-1} |A_{k-1} - A_k|.$$

Empirical results show that the procedure for handling open ended intervals in the first iteration step has little impact on prediction.

The proposed SEM algorithm makes repeated use of a two sided truncated normal distribution, by drawing from $N(x_{ij}^T\hat{\beta} + \hat{u}_i, \hat{\sigma}_e^2)$ within the given interval $(A_{k-1} \leq y_{ij} \leq, A_k)$. Therefore, the performance of the SEM algorithm relies on the Gaussian assumptions of the error terms being met. To ensure that this is the case, the SEM algorithm is extended to allow for the use of transformations.

3.2.2 The SEM algorithm under transformations

Transformations of the outcome can be used for ensuring that the model assumptions are met. Broadly speaking one can use non-adaptive or adaptive transformations. For the application in this paper that models income-type data, the logarithmic transformation is probably the one most commonly used. While the logarithmic transformation is easy to use, there is no guarantee that it will provide the best transformation for the target distribution. This is particularly crucial in this paper since the validity of the normality assumption of the residuals cannot be tested due to the fact that the response variable is interval censored. Therefore, using adaptive (data-driven) transformations, instead of fixed transformations, is preferable. In addition, the logarithmic transformation can be obtained as a special case of a family of adaptive transformations. In this paper we focus on the use of the Box-Cox transformation (Box and Cox, 1964; Draper and Cox, 1969) and its extension under the nested error regression model (Gurka et al.,

2006). The Box-Cox transformation is given by

$$y_{ij}(\lambda) = \begin{cases} \frac{(y_{ij}+s)^{\lambda}-1}{\lambda} & \text{if } \lambda \neq 0, \\ \ln(y_{ij}+s) & \text{if } \lambda = 0, \end{cases}$$

where s is a fixed shift parameter that assures $y_{ij}+s > 0$. The Box-Cox transformation depends on the transformation parameter λ that is used for transforming the data $T_{\lambda}(y_{ij}) = y_{ij}(\lambda)$. The aim is to find the value of λ given the data such that the assumptions about the error terms of the nested error regression model are met (Gurka et al., 2006). The implementation of data-driven transformations within the SEM algorithm is computationally intensive because the transformation parameter λ has to be estimated in each iteration step. The algorithm is structured into two parts. In Part 1 the SEM algorithm is used for finding the optimal transformation parameter, $\hat{\lambda}^{(F)}$. In Part 2 the SEM algorithm is implemented with the estimated $\hat{\lambda}^{(F)}$ from Part 1. The detailed steps of the SEM algorithm under transformations are given below.

Part 1

1. Define a grid g of possible values of λ. Using each value in the grid in turn, implement the steps below.

2. Use the scaled version of the Box-Cox transformation, as defined in Rojas-Perilla et al. (2017), to transform the midpoints of each interval (A_{k-1}, A_k) and fit the nested error regression model (3.1). Repeat the same process for each value of λ in g and select the value of $\hat{\lambda}$ that maximizes the restricted maximum likelihood (Bartlett, 1937).

3. Using the selected value of $\hat{\lambda}$ from the previous step fit the nested error regression model (3.1) to obtain $\hat{\theta} = (\hat{\beta}, \hat{\sigma}_e^2, \hat{\sigma}_u^2)$.

4. Generate a new pseudo sample as a proxy for the unobserved $y_{ij}(\hat{\lambda})$. To do this, for $j = 1, \ldots, n_i$ and $i = 1, \ldots, D$ sample from the conditional distribution $f(y_{ij}(\lambda)|x_{ij}, k_{ij}, u_i)$ by drawing from $N(x_{ij}^T\hat{\beta} + \hat{u}_i, \hat{\sigma}_e^2)$ within the given interval $(A_{k-1}(\hat{\lambda}) \leq y_{ij}(\hat{\lambda}) \leq A_k(\hat{\lambda}))$ to obtain $(\tilde{y}_{ij}(\hat{\lambda}), x_{ij})$. Back-transform $\tilde{y}_{ij}(\hat{\lambda})$ to the original scale $\tilde{y}_{ij}$ using the selected $\hat{\lambda}$ from Step 2.

5. Go to Step 1 and select a new optimal $\hat{\lambda}$ this time using the newly generated $\tilde{y}_{ij}$ from the previous step in Step 2 of the algorithm instead of the interval midpoints.

6. Iterate Steps 1-5 $B + M$ times, with B burn-in iterations and M additional iterations.

7. Discard the burn-in iterations and estimate the final $\hat{\lambda}^{(F)}$ by averaging the M estimates of $\hat{\lambda}$.

Part 2

8. Use $\hat{\lambda}^{(F)}$ from Part 1 and the the Box-Cox transformation to transform the midpoints of each interval (A_{k-1}, A_k).

9. Fit the nested error regression model (3.1) using the transformed midpoints to obtain $\hat{\theta} = (\hat{\beta}, \hat{\sigma}_e^2, \hat{\sigma}_u^2)$.

10. **S-step:** Generate a new pseudo sample as a proxy for the unobserved continuous outcome. To do this, for $j = 1, \ldots, n_i$ and $i = 1, \ldots, D$ sample from the conditional distribution $f(y_{ij}(\lambda)|x_{ij}, k_{ij}, u_i)$ by drawing from $N(x_{ij}^T\hat{\beta} + \hat{u}_i, \hat{\sigma}_e^2)$ within the given interval $(A_{k-1}(\hat{\lambda}^{(F)}) \leq y_{ij}(\hat{\lambda}^{(F)}) \leq, A_k(\hat{\lambda}^{(F)}))$ to obtain $(\tilde{y}_{ij}(\hat{\lambda}^{(F)}), x_{ij})$.

11. **M-step:** Re-estimate the vector $\hat{\theta}$ using (3.1) and the pseudo samples $(\tilde{y}_{ij}(\hat{\lambda}^{(F)}), x_{ij})$ from the previous step.

12. Iterate Steps 10-11 $B + M$ times, with B burn-in iterations and M additional iterations.

13. Discard the B burn in iterations and estimate $\hat{\theta}$ by averaging the derived M estimates.

Figure 3.1 illustrates why in the case of using transformations it is important to structure the SEM algorithm in two parts, i.e., finding the optimal λ first and then using the optimal λ, estimating β. The left panel of Figure 3.1 plots the estimated λ for each iteration step of the algorithm for estimating its convergence. The right panel of Figure 3.1 plots $\hat{\lambda}$ against any $\hat{\beta}$ for each iteration step of Part 1. From that plot it is clear that by simply running Part 1 and averaging the M estimates of $\hat{\beta}$ and $\hat{\lambda}$ the averaged estimates would not correspond. This is the case because the relationship between $\hat{\lambda}$ and $\hat{\beta}$ is non-linear. Therefore, the SEM algorithm has to be divided into two parts. In Part 1 the final $\hat{\lambda}^{(F)}$ is estimated and in Part 2 this estimate is used for estimating the parameters of the nested error regression model on the transformed scale.

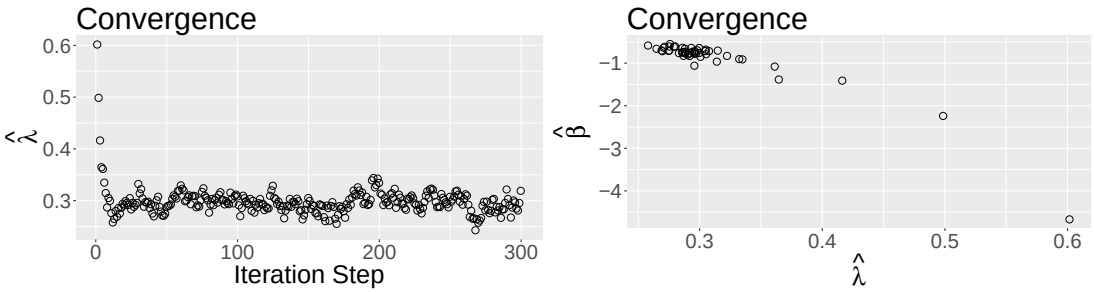


Figure 3.1: Convergence of $\hat{\lambda}$ and any $\hat{\beta}$ using the SEM algorithm.

3.3 Small area empirical best prediction with interval-censored data

In this section we present methods for small area prediction when the response variable is interval censored. The application is on estimating income-type indicators. In addition to the notation we introduced in the previous section, we denote sampled units in area i by s_i and the non-sampled units by r_i. For each area i the sample size is n_i with $n = \sum_{i=1}^{D} n_i$ and the population vector y_i for area i comprises sampled and non-sampled units $y_i^T = (y_{is}^T, y_{ir}^T)$. The target of inference are small area parameters that include linear and non-linear indicators which

can be expressed as functions of an income variable for example, average and median equivalized income, the head count ratio (at risk of poverty indicator), the quantile share ratio index, and the Gini coefficient. Since in this paper we assume the availability of unit-level survey and census/administrative data, two methods for estimating non-linear indicators are available, the World Bank method (Elbers et al., 2003) and the popular EBP method (Molina and Rao, 2010). Although our focus is on the use of the EBP method, the proposed methodology can be applied to the World Bank method too. The EBP method makes use of unit-level nested error regression model and is summarized below. The response variable is income that is only available in the survey. The explanatory variables, used for modeling the income variable, are available both in the survey and in the census data sets. After the model is fitted using the survey data, the estimated model parameters are combined with census micro-data to form unit-level synthetic census predictions of the income variable. These synthetic values are then used for estimating the target parameters. Census predictions are generated by using the conditional predictive distribution of the out-of-sample data given the sample data. The starting point is the following unit-level nested error regression model,

$$y_{ij} = x_{ij}^T\beta + u_i + e_{ij}, \quad u_i \overset{iid}{\sim} N(0, \sigma_u^2), \quad e_{ij} \overset{iid}{\sim} N(0, \sigma_e^2),$$

where u_i denotes the domain (area) random effect. Assuming normality for the unit-level error and the domain random effects, the conditional distribution of the out-of-sample data given the sample data is also normal. Predictions for the entire population of area i are generated from the following model,

$$\begin{aligned} &y_{ij}^* = x_{ij}^T\beta + \tilde{u}_i + u_i^* + e_{ij}^*, \\ &u_i^* \overset{iid}{\sim} N(0, \sigma_u^2 \times (1-\gamma_i)), \quad e_{ij}^* \overset{iid}{\sim} N(0, \sigma_e^2), \quad \gamma_i = \frac{\sigma_u^2}{\sigma_u^2 + \frac{\sigma_e^2}{n_i}}, \end{aligned} \tag{3.2}$$

where $\tilde{u}_i = E(u_i|y_{is})$ is the conditional expectation of u_i given the sample data y_{is}. Implementation of Equation (3.2) requires replacing the unknown quantities $\beta, \sigma_u, \sigma_e$, with estimates and simulating L synthetic populations of the income variable, y_{ij}^*. Linear and non-linear indicators are computed in each domain i for each replication and the estimates are averaged over the number of Monte Carlo simulations L. This number is usually set equal to $L = 50$ or $L = 100$ but higher numbers are also possible.

In the presence of an interval-censored income variable the EBP approach needs to be modified. In the first step the model parameters, $\hat{\theta} = (\hat{\beta}, \hat{\sigma}_u^2, \hat{\sigma}_e^2)$, are estimated using the SEM algorithm outlined in Section 3.2. Having estimated $\hat{\theta}$, the remaining steps of the Monte Carlo algorithm used to implement the EBP approach are as follows:

1. Use the sample data and the SEM algorithm to estimate $\hat{\theta} = (\hat{\beta}, \hat{\sigma}_u^2, \hat{\sigma}_e^2)$ and $\hat{\gamma}_i = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \frac{\hat{\sigma}_e^2}{n_i}}$.

2. For $l = 1, \ldots, L$:

   (a) Generate a synthetic population under the nested error regression model $\hat{y}_{ij}^{*(l)} =$

$x_{ij}^T\hat{\beta} + \tilde{u}_i + u_i^{*(l)} + e_{ij}^{*(l)}$, where x_{ij} are population micro-data for unit j in area i, $u_i^{*(l)} \overset{iid}{\sim} N(0, \hat{\sigma}_u^2(1-\hat{\gamma}_i))$, $e_{ij}^{*(l)} \overset{iid}{\sim} N(0, \hat{\sigma}_e^2)$ and $\tilde{u}_i$ is given by $\tilde{u}_i = E(u_i|y_{is})$.

(b) In each area, estimate the target parameter $\hat{I}_i^{(l)}$ using $\hat{y}_{ij}^{*(l)}$.

3. The target parameter is estimated by averaging over the L Monte Carlo estimates $\hat{I}_i^{(l)}$ in each area,

$$\hat{I}_i^{EBP} = \frac{1}{L}\sum_{l=1}^{L}\hat{I}_i^{(l)}.$$

It is likely that when modeling an income variable the normality assumptions of the nested error regression model may not hold. In this case a suitable transformation is needed and the SEM algorithm is implemented using the results in Section 3.2.2. After estimates of $\hat{\theta} = (\hat{\beta}, \hat{\sigma}_u^2, \hat{\sigma}_e^2)$ and $\hat{\lambda}^{(F)}$ have been obtained, the Monte Carlo steps of the EBP method are implemented as follows:

1. Use the sample data and the SEM algorithm to estimate $\hat{\theta} = (\hat{\beta}, \hat{\sigma}_u^2, \hat{\sigma}_e^2)$, $\hat{\lambda}^{(F)}$ and $\hat{\gamma}_i = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \frac{\hat{\sigma}_e^2}{n_i}}$.

2. For $l = 1, \ldots, L$:

   (a) Generate a synthetic population under the nested error regression model $\hat{y}_{ij}^{*(l)}(\hat{\lambda}^{(F)}) = x_{ij}^T\hat{\beta} + \tilde{u}_i + u_i^{*(l)} + e_{ij}^{*(l)}$, where x_{ij} are population micro-data for unit j in area i, $u_i^{*(l)} \overset{iid}{\sim} N(0, \hat{\sigma}_u^2(1-\hat{\gamma}_i))$, $e_{ij}^{*(l)} \overset{iid}{\sim} N(0, \hat{\sigma}_e^2)$ and $\tilde{u}_i$ is given by $\tilde{u}_i = E(u_i|y_{is})$.

   (b) Back-transform to the original scale $\hat{y}_{ij}^{*(l)} = T^{-1}\left(\hat{y}_{ij}^{*(l)}(\hat{\lambda}^{(F)})\right)$.

   (c) In each area, estimate the target parameter $\hat{I}_i^{(l)}$ using $\hat{y}_{ij}^{*(l)}$.

3. The target parameter is estimated by averaging over the L Monte Carlo estimates $\hat{I}_i^{(l)}$ in each area,

$$\hat{I}_i^{EBP} = \frac{1}{L}\sum_{l=1}^{L}\hat{I}_i^{(l)}.$$

For non-sampled areas we cannot estimate an area random effect, hence $\tilde{u}_i$ is not available. In this case Step 2(a) above is modified such that synthetic values of the outcome are generated as follows, $\hat{y}_{ij}^{*(l)}(\hat{\lambda}^{(F)}) = x_{ij}^T\hat{\beta} + u_i^{*(l)} + e_{ij}^{*(l)}$, where the error terms are drawn from $u_i^{*(l)} \overset{iid}{\sim} N(0, \hat{\sigma}_u^2)$ and $e_{ij}^{*(l)} \overset{iid}{\sim} N(0, \hat{\sigma}_e^2)$. The same applies to the case where we are working with the untransformed response variable.

3.3.1 Mean squared error estimation

MSE estimation is a crucial step in small area estimation. Complications arise due to the complexity of non-linear indicators which make the development of analytic MSE estimators difficult. For the EBP Molina and Rao (2010) propose a parametric bootstrap MSE estimator

under the nested error regression model. A parametric bootstrap is also used when working with an interval-censored outcome. However, there are two additional sources of variability we need to account for. One is the uncertainty due to the estimation of the transformation parameter and the second is the uncertainty resulting from working with limited information due to interval censoring. The bootstrap MSE assumes that the mechanism used to interval censor the response variable is known. Denoting by b the bootstrap iteration, the bootstrap MSE estimator below is presented in the more general case where a transformation of the response variable is used.

1. (a) Using the sample estimates, $\hat{\beta}, \hat{\sigma}_u^2, \hat{\sigma}_e^2, \hat{\lambda}^{(F)}$, at convergence of the SEM algorithm, generate $u_i^{*(b)} \overset{iid}{\sim} N(0, \hat{\sigma}_u^2)$ and $e_{ij}^{*(b)} \overset{iid}{\sim} N(0, \hat{\sigma}_e^2)$ and a bootstrap superpopulation $\hat{y}_{ij}^{*(b)}(\hat{\lambda}^{(F)}) = x_{ij}^T\hat{\beta} + u_i^{*(b)} + e_{ij}^{*(b)}$, where x_{ij} are population micro-data for unit j in area i.

   (b) Back-transform $\hat{y}_{ij}^{*(b)} = T^{-1}\left(\hat{y}_{ij}^{*(b)}(\hat{\lambda}^{(F)})\right)$ to the original scale and compute the population value of the target parameter in area i and bootstrap iterations b, $I_{i,b}$.

   (c) Select a bootstrap sample using a simple random sampling with replacement from each area that respects the area-specific sample sizes of the original sample.

   (d) Using the known censoring mechanism and the bootstrap sample data, create the interval-censored response variable.

   (e) Use the SEM algorithm with the current bootstrap sample for deriving EBP estimates of the target parameters. In this case where a transformation is used this consists of using Parts 1 and 2 from Section 3.2.2 and the EBP algorithm under a transformation described in Section 3.3.

   (f) Obtain EBP estimates of the target parameter in area i and bootstrap iteration b, $\hat{I}_{i,b}^{EBP}$.

2. Using a total of B bootstrap samples, the MSE estimator is computed as follows:

$$\widehat{\text{MSE}}\left(\hat{I}_i^{EBP}\right) = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{I}_{i,b}^{EBP} - I_{i,b}\right)^2.$$

3.4 Model-based simulations

This section presents model-based simulation results for assessing the performance of the proposed methodology. In particular, we assess the performance of EBP point estimators and of corresponding MSE estimators. As poverty indicators we estimate the head count ratio (HCR) and the poverty gap (PGAP) as defined in Foster et al. (1984) (income deprivation) and the

mean of household income in each area i. The indicators are defined as follows:

$$\text{HCR}_i = \frac{1}{n_i}\sum_{j=1}^{n_i}\mathbf{I}(y_{ij} \leq z),$$

$$\text{PGAP}_i = \frac{1}{n_i}\sum_{j=1}^{n_i}\Big(\frac{z - y_{ij}}{z}\Big)\mathbf{I}(y_{ij} \leq z),$$

$$\text{Mean}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij},$$

where y_{ij} denotes the outcome variable, $\mathbf{I}(\cdot)$ denotes the indicator function and z is the poverty threshold. In the simulations and application in this paper z is set to 60% of the median of income, as defined by (Eurostat, 2014).

Two population models outlined in Table 3.1 are used for generating the simulated data. The normal scenario (in Section 3.4.1) is used for evaluating the performance of the EBP approach under interval censoring of the response variable when the model assumptions are met. The log-scale scenario (in Section 3.4.2) attempts to mimic the distribution of an income variable we might work with in practice. In this case the Gaussian assumptions of the model are not met and the use of transformations is necessary. For both population models we also assess the estimation of the transformation parameter λ. This can be achieved since the theoretical values of λ under the normal and log-normal scenarios are known. In addition, we also assess the properties of the proposed bootstrap MSE estimator.

For the normality-based scenario we use two different interval-censoring mechanisms, referred to as normal scenario 1 and normal scenario 2 (see Tables 3.8 and 3.9 in Appendix 3.7). This allows us to explore the impact of the number of groups on the performance of the small area estimators which is of interest for survey practitioners. Under normal scenario 1 the number of groups used for splitting the income variable is 14. Under normal scenario 2 we use a more extreme censoring mechanism with only seven groups. The reason for deciding to use this number of groups in the simulation studies follows international practice. For example, the census in Colombia groups income in nine intervals, in Australia 10 intervals are used and in New Zealand 14 (Departamento Administrativo Nacional De Estadística, 2005; Australian Bureau of Statistics, 2011; Statistics New Zealand, 2013).

In each Monte Carlo run a finite population U of size $N = 10000$ is generated and is partitioned into $D = 50$ areas $U_1, U_2, \ldots, U_D$ each with size $N_i = 200$. From the finite population we select a sample using an unbalanced design with area-specific sample sizes n_i ranging between $8 \leq n_i \leq 29$. The total sample size is $n = 921$. In total we run 200 Monte Carlo iterations with the number of Monte Carlo iterations for implementing the EBP set to $L = 200$ and the number of bootstrap iterations for MSE estimation set to $B = 200$.

Different small area estimators are compared. In particular, we compare the EBP under the model that assumes that the continuous response variable is available (abbreviated below by LME) to the EBP when only the interval-censored variable is available and the use of the SEM algorithm is necessary (abbreviated below by SEM). For both model-based scenarios we

Table 3.1: Model-based simulation scenarios.

Scenario	Model	x_{ij}	z_{ij}	μ_i	u_i	e_{ij}
Normal	$4500 - 400x_{ij} + u_i + e_{ij}$	$N(\mu_i, 3)$	-	$U(-3, 3)$	$N(0, 500^2)$	$N(0, 1000^2)$
Log-scale	$exp(10 - x_{ij} - 0.5z_{ij}u_i + e_{ij})$	$N(\mu_i, 2)$	$N(0, 1)$	$U(-3, 3)$	$N(0, 0.4^2)$	$N(0, 0.8^2)$

further compare the standard EBP when a Box-Cox transformation is used (LME Box-Cox) to the EBP-SEM approach when a Box-Cox transformation is used (abbreviated below by SEM Box-Cox). This allows us to examine how well the parameter of the Box-Cox transformation λ is estimated when we only have access to the interval-censored response. For assessing the use of a fixed transformation, the standard EBP as well as the EBP with grouped data is used with a logarithmic transformation (LME Log, SEM Log). The SEM algorithm uses 40 burn-in iterations and 200 additional iterations. This number is sufficient in the simulation study to ensure convergence. The convergence of the SEM algorithm is graphically checked by plotting the parameter estimates at each iteration step for randomly chosen simulation runs.

The performance of point estimates is assessed by computing the area-specific root mean squared error (RMSE) of the target parameter $\hat{I}_i$ (Hyndman and Koehler, 2006). Tables are used to report the average, over areas, of the RMSE. The area-specific values of the RMSE are computed as follows:

$$RMSE\big(\hat{I}_i^{EBP}\big) = \left[\frac{1}{M}\sum_{m=1}^{M}\big(\hat{I}_i^{EBP(m)} - I_i^{(m)}\big)^2\right]^{1/2}, \tag{3.3}$$

where M is the total number of Monte Carlo iterations, m denotes the Monte Carlo iteration, $\hat{I}_i^{EBP}$ is the estimated indicator using one of the above mentioned methods and I_i is the true population value.

The proposed MSE estimator is visually evaluated, by plotting the estimated root MSE defined as $\text{Est.RMSE}_i := \sqrt{\widehat{\text{MSE}}\left(\hat{I}_i^{EBP}\right)}$ and the empirical root MSE, called Emp.RMSE_i as defined in (3.3) for each area i. Furthermore, for each area i the relative bias and the relative RMSE of the Est.RMSE_i are estimated as follows:

$$rel.RMSE\big(\text{Est.RMSE}_i\big) = \left[\left(\frac{\text{Est.RMSE}_i - \text{Emp.RMSE}_i}{\text{Emp.RMSE}_i}\right)^2\right]^{1/2} \times 100,$$
$$rel.Bias\big(\text{Est.RMSE}_i\big) = \left(\frac{\text{Est.RMSE}_i - \text{Emp.RMSE}_i}{\text{Emp.RMSE}_i}\right) \times 100.$$

The results of the different scenarios are presented in the next two sections.

3.4.1 Results: Normality-based scenarios

Table 3.2 presents the results for normal scenario 1 (14 intervals) and normal scenario 2 (seven intervals) using the SEM method, the SEM Box-Cox method, and the LME and LME Box-Cox

methods. The results show that the performance of the EBPs using the SEM algorithm is close to the performance of the EBPs when the continuous outcome is fully available. As expected, when using the fully available continuous outcome the EBP estimates are more efficient (lower RMSE) than the SEM-based estimates. However, despite working with the interval-censored outcome the increase in RMSE (reduction in efficiency) is not dramatic which demonstrates that the SEM algorithm works well. In line with the theory, the results also show that as the number of classes used to discretise the continuous outcome reduces (from 14 to seven groups), the RMSE of the SEM-based estimates increases. This is reasonable as in this case the information available reduces. Nevertheless, even in the case of scenario 2 we would argue that the performance of the SEM-based estimates is reasonable. Our view is based on the fact that seven groups present a rather extreme scenario.

Table 3.2: Performance of the estimated EBPs in terms of RMSE over areas.

	Indicator:	Mean		HCR		PGAP	
		Median	Mean	Median	Mean	Median	Mean
Normal scenario 1 (14 intervals)							
RMSE	LME	201.482	212.450	0.033	0.035	0.014	0.015
	LME Box-Cox	201.675	212.466	0.033	0.035	0.014	0.016
	SEM	203.783	217.075	0.034	0.036	0.014	0.016
	SEM Box-Cox	204.604	217.335	0.034	0.036	0.014	0.017
Normal scenario 2 (7 intervals)							
RMSE	LME	200.725	212.405	0.033	0.035	0.014	0.015
	LME Box-Cox	201.422	212.502	0.033	0.035	0.014	0.016
	SEM	216.780	225.692	0.035	0.038	0.015	0.017
	SEM Box-Cox	215.324	225.897	0.035	0.037	0.016	0.018
Log-scale scenario (7 intervals)							
RMSE	LME Log	994.586	988.374	0.063	0.065	0.039	0.040
	LME Box-Cox	995.068	992.021	0.063	0.065	0.040	0.040
	SEM Log	1046.724	1030.190	0.066	0.068	0.041	0.042
	SEM Box-Cox	1043.407	1040.646	0.066	0.068	0.040	0.042

In Figures 3.2 and 3.3 the estimated density of the population y_{ij} values are plotted against the estimated density of $\hat{y}_{ij}^{*(l)}$ using the different estimation methods from one arbitrarily chosen simulation run. The plots confirm the previous conclusions. The density estimated with the SEM methods is close to the population density and estimates become less accurate as the number of intervals used in censoring the response variable decreases (see Figures 3.2 and 3.3).

The performance of the estimates using the SEM and SEM Box-Cox methods is very similar. In the case of the normal-based scenarios this is expected since the data driven transformation parameter, λ, is estimated to be close to one, which is equivalent to using no transformation. This is confirmed by looking at the estimation of λ given in Table 3.3. Hence, the Box-Cox transformation adapts well to the shape of the data distribution, even though only the interval information is used estimating λ. The estimation accuracy for λ also depends on the number of intervals used for censoring the response variable, as seen in Table 3.3 for normal scenario 1 and normal scenario 2.

The MSE results for the different indicators are summarized in Table 3.4. Table 3.4 shows the relative bias and the relative RMSE of the estimated RMSE. Overall, the relative bias and

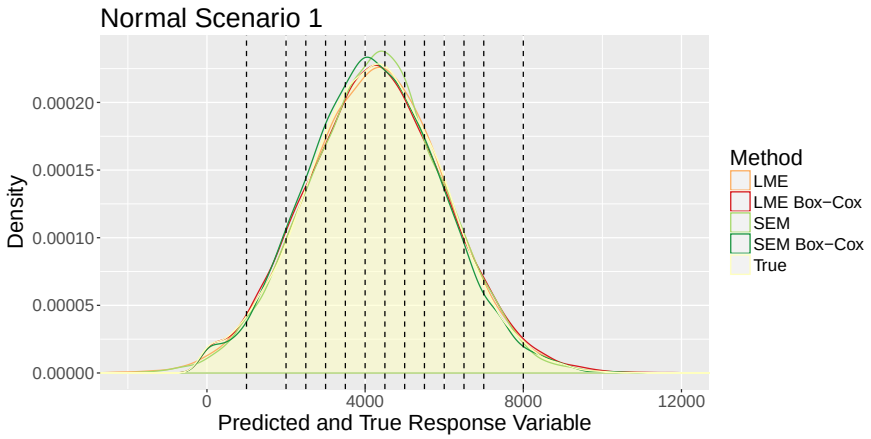


Figure 3.2: Normal scenario 1 (14 intervals): Estimate of the true population density and estimate of the predicted population density $\hat{y}_{ij}^{*(l)}$ from a randomly chosen run l.

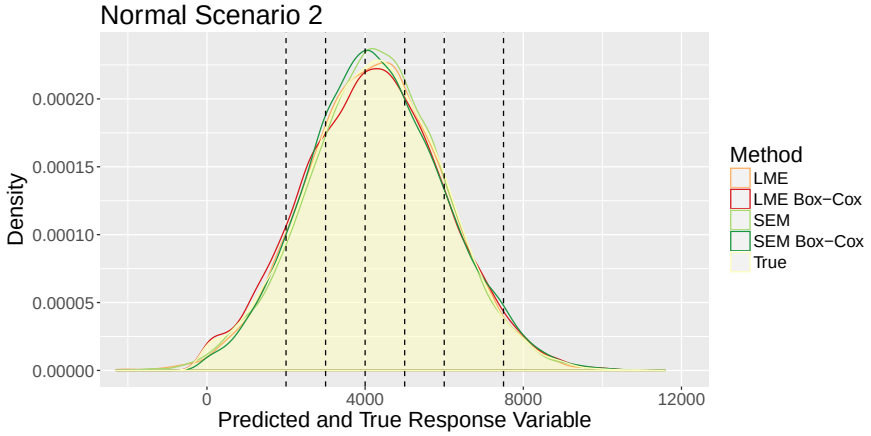


Figure 3.3: Normal scenario 2 (7 intervals): Estimate of the true population density and estimate of the predicted population density $\hat{y}_{ij}^{*(l)}$ from a randomly chosen run l.

relative RMSE are relatively low. In particular, for most scenarios and target parameters the relative bias is below 10% and for a few scenarios somewhat above 10%. The relative RMSE also shows that the bootstrap estimator is stable. From Table 3.4 we cannot evaluate how well the estimated RMSE tracks the empirical (Monte Carlo) RMSE. Therefore, Figure 3.4 shows the estimated and empirical RMSE over the domains when estimating the HCR using the SEM Box-Cox method. We conclude that the estimated RMSE tracks the empirical RMSE well.

3.4.2 Results: Log-scale scenario

In this section we present results when the assumptions of the nested error regression model are not met. This is the case for the log-scale scenario. Its generation mechanism is described in Table 3.1. For this scenario the response variable is grouped in seven intervals hence, a fairly extreme censoring mechanism is evaluated. The distribution of the response variable using one arbitrarily chosen Monte Carlo population can be seen in Table 3.10 in Appendix 3.7. Four estimation methods are compared, namely the SEM Box-Cox method, the SEM Log

Table 3.3: Estimates of the transformation parameter λ over simulation runs.

	Normal scenario 1		Normal scenario 2		Log-scale scenario	
	Median	Mean	Median	Mean	Median	Mean
LME: $\hat{\lambda}$	0.91	0.91	0.91	0.91	0.00	0.00
SEM: $\hat{\lambda}^{(F)}$	0.91	0.91	0.85	0.86	0.00	0.00

Table 3.4: Performance of the bootstrapped root MSE estimator over areas.

	Indicator:	Mean		HCR		PGAP	
		Median	Mean	Median	Mean	Median	Mean
Normal scenario 1 (14 intervals)							
rel.Bias[%]	SEM	7.371	7.054	5.907	5.069	2.307	3.066
	SEM Box-Cox	7.557	7.331	5.611	5.471	-6.856	-5.580
rel.RMSE[%]	SEM	9.497	10.502	10.588	11.379	12.046	13.344
	SEM Box-Cox	9.916	10.854	10.778	11.418	13.033	14.007
Normal scenario 2 (7 intervals)							
rel.Bias[%]	SEM	5.296	5.839	4.710	3.649	-0.178	0.297
	SEM Box-Cox	5.462	6.095	4.591	3.909	-15.770	-14.823
rel.RMSE[%]	SEM	8.586	9.911	10.224	10.978	12.224	13.294
	SEM Box-Cox	8.815	10.297	10.296	11.067	19.269	19.155
Log-scale scenario (7 intervals)							
rel.Bias[%]	SEM Log	7.219	6.559	6.734	7.582	6.737	7.130
	SEM Box-Cox	13.169	25.997	6.780	7.649	7.096	7.573
rel.RMSE[%]	SEM Log	33.486	34.777	13.188	14.248	21.017	21.625
	SEM Box-Cox	42.186	60.854	13.227	14.361	21.332	21.930

method, the LME Box-Cox method, and the LME Log method. Estimation without the use of transformations is not considered in this case because we know that the model assumptions do not hold. Nevertheless, the methods that use a Box-Cox transformation are adaptive and the estimated transformation parameter informs us about the need to use a transformation. The results in Table 3.2 show that the performance of the estimates using the SEM Box-Cox and the SEM Log methods is close to the performance of the estimates using the LME Box-Cox and the LME Log methods that assume that the continuous outcome variable is available. As expected, some accuracy in estimation is compromised when working with the interval-censored outcome. However, the SEM-based estimates remain competitive when compared to the estimates obtained by assuming that full information for the response variable is available. This is also confirmed by looking at how the SEM-based methods recover the true population density in Figure 3.5.

The use of the Box-Cox transformation appears to be working well. Under this scenario the transformation parameter λ should be estimated to be close to zero. This is confirmed by examining the estimation results of $\hat{\lambda}$ in Table 3.3. Finally, as shown in Figure 3.4 and Table 3.4 the proposed bootstrap MSE estimator tracks the empirical (Monte Carlo) well and has reasonably low relative bias. As expected, the MSE under the Box-Cox version of the SEM is somewhat more unstable than the corresponding MSE for the Log SEM. This is due to the fact that in the case of the Box-Cox method the transformation parameter is estimated with each bootstrap sample whereas for the Log method the transformation is held fixed.

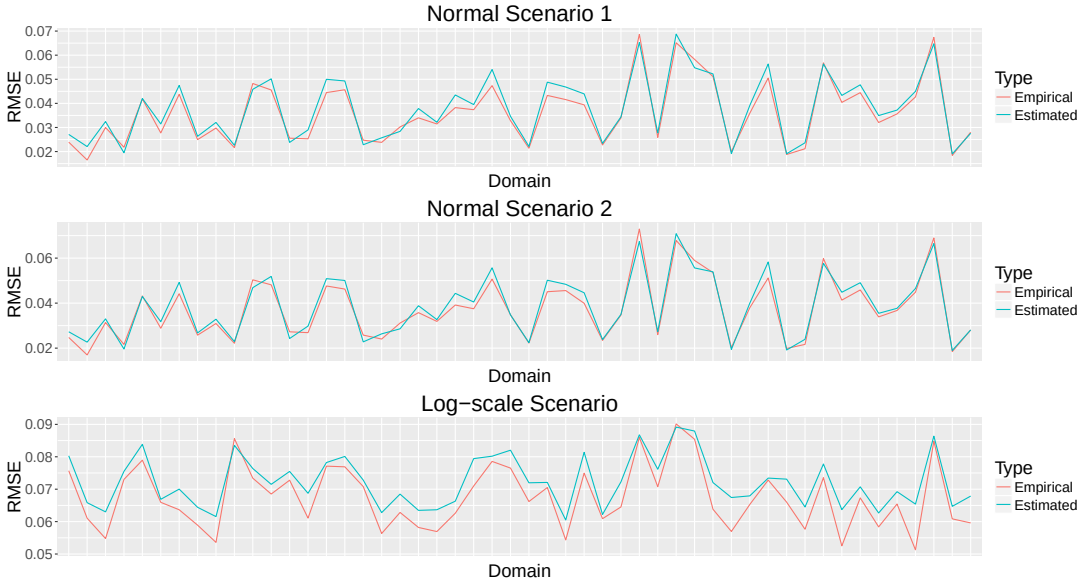


Figure 3.4: Estimated and empirical area-specific RMSEs of the HCR using the SEM Box-Cox estimator.

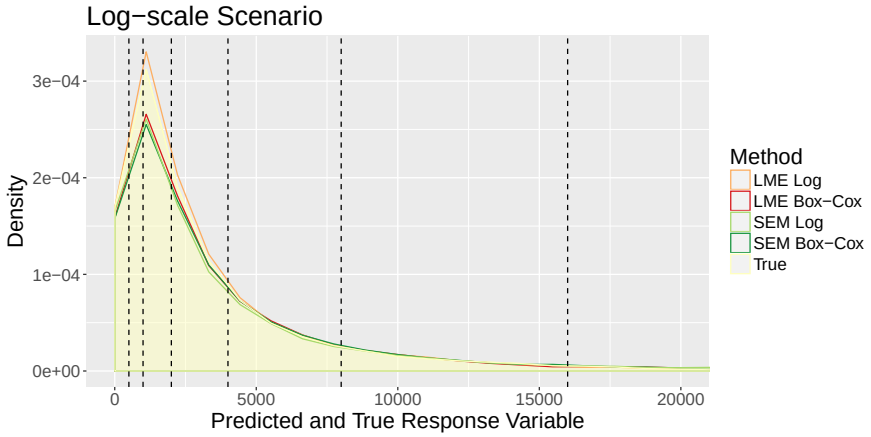


Figure 3.5: Log-scale scenario (7 intervals): Estimate of the true population density and estimate of the predicted population density $\hat{y}_{ij}^{*(l)}$ from a randomly chosen run l.

3.5 Estimating small area deprivation indicators for municipalities in the Mexican state of Chiapas

In this section the proposed methods are applied to real census and survey data from Mexico. Despite Mexico being the 15th largest economy in the world (International Monetary Fund, 2017), the fight against poverty and inequality is of great importance for the country since high poverty rates are omnipresent. During the Mexican peso crisis extreme poverty increased from 21% in 1994 to 37% in 1996 (Pereznieto, 2010). Today, poverty rates remain at considerably high levels. According to the World Bank (2010), 33% of the population in the country experienced moderate poverty and 9% extreme poverty in 2013. This demonstrates the relevance of estimating and mapping poverty at local levels such that appropriate interventions can be designed.

In this paper the target parameters are the average municipal household income, the municipal head count ratio and the municipal poverty gap. Estimation uses the 2010 equivalized household income and expenditure survey called ENIGH (Encuesta Nacional de Ingreso y Gasto de los Hogares) and a large sample of the 2010 National Population and Housing census. Both data sets are collected by the National Institute of Statistics and Geography (INEGI Instituto Nacional de Estadística y Geografía) and they are provided to us by the National Council for the Evaluation of Social Development Policy (CONEVAL Consejo Nacional de Evaluación de la Política de Desarrollo Social). Both the census and survey data set include socioeconomic and regional information at household level. While the data cover all 31 states of Mexico, the application is focusing on the state of Chiapas. Chiapas is one of the poorest states in Mexico with an average income of about 40% of the national median income (Levy et al., 2016). The state is located in the south of Mexico at the boarder to Guatemala. The survey covers 42 out of the 118 municipalities in Chiapas so there are 42 in-sample and 76 out-of-sample municipalities for which no sample data is available. In order to derive precise and reliable estimates at the level of municipality for all 118 municipalities we rely on the use of model-based methods and auxiliary information from the census and survey micro-data.

The total sample size is $n = 2486$ households and the census sample size is $N = 96350$ households. The regional distribution of the sample size is given in Table 3.5. The sample size of the in-sample municipalities varies between 13 and 651 households with a median sample size of 33 households. Since sample sizes are small in many municipalities, SAE methods can potentially improve the accuracy of the small area estimates.

Table 3.5: Distribution of the sample and census household sizes across areas.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Sample	13.00	17.00	33.00	59.19	51.00	651.00
Census	82.00	399.50	617.50	816.50	839.00	7172.00

The estimation methods we consider in this paper rely on the use of a nested error regression model estimated with the survey data. The response variable equivalized household income is measured on a continuous scale. In order to present the performance of the newly

proposed methodology we group equivalized household income to 14 and eight intervals. The distribution of the grouped equivalized household income is given in Table 3.11 and Table 3.12 in Appendix 3.7. The variables in Table 3.6 were identified as possible covariates that predict equivalized household labor income well. The variables in the working model are selected with regards to content and by a – for mixed models suitable – coefficient of determination proposed by Nakagawa and Schielzeth (2013). The conditional R^2_c, interpreted as the variance explained by the whole model, is $R^2_{c,lme} = 0.61$ when estimating the model with the observed continuous response variable on the transformed scale (Box-Cox transformation). When estimating the model with an interval-censored response variable on the transformed scale (Box-Cox transformation) using the SEM algorithm the $R^2_{c,sem(14)} = 0.61$ and $R^2_{c,sem(8)} = 0.62$ for the 14 and eight interval scenario, respectively.

Table 3.6: Variables used in the nested error regression working model.

Variable type	Description
Response variable:	Interval-censored equivalized household labour income
Auxiliary variables:	Value of all household goods
	Value of household communication equipment
	Share of employees in the household
	Educational level of head of household
	Social class of head of household
	Municipalities of Chiapas

The Box-Cox transformation is used as the preferred transformation method because it is data-driven. This is particularly crucial when working with interval-censored data as response variable, because the normality assumption of the residuals cannot be checked. The estimated transformation parameters are $\lambda_{lme} = 0.16$ for the continuous response, $\lambda_{sem(14)} = 0.18$ and $\lambda_{sem(8)} = 0.17$ for the 14 and eight interval-censoring scenarios, respectively. The results show that the algorithm is able to identify the λ we would have estimated should we have modelled the continuous outcome. The results also indicate that the use of a logarithmic transformation or the use of the untransformed response variable may lead to erroneous results. Rojas-Perilla et al. (2017) and Tzavidis et al. (2018) show that even if $\hat{\lambda}$ is close to 0 the the EBP estimates using the Box-Cox transformation may outperform the EBP estimates using the logarithmic transformation.

Estimates of the mean equivalized household labor income, HCR and PGAP for each of the 118 municipalities are obtained by using the SEM Box-Cox method based on 14 and eight intervals, and by using LME Box-Cox based on the observed continuous response variable. The mean and median averaged over all municipalities are given in Table 3.7 and plotted in Figures 3.6 and 3.8. The results show that the point estimates from all three estimation methods are very close. Interval censoring does not appear to impact significantly on the estimation results. Additionally, the relative efficiencies of the estimators (EFF) defined as $EFF(\hat{I}_i^{EBP}) = RMSE_{sem}(\hat{I}_i^{EBP})/RMSE_{lme}(\hat{I}_i^{EBP})$ is reported in the Table 3.7. It is notable that the efficiency loss is small even when the response variable is grouped to only eight intervals. In the 14 interval scenario the point estimates of the mean are even more efficient, but this result is only due to the Monte Carlo variability. The spatial distributions of the HCR

in municipalities in Chiapas are shown in Figure 3.6 for all three estimation methods. The figure supports the priorly-stated results that estimation results are very closed to each other, independently from the estimation method.

Table 3.7: Point estimates and corresponding EFF of the point estimates over municipalities using the SEM Box-Cox algorithm.

	Mean		HCR		PGAP	
	Median	Mean	Median	Mean	Median	Mean
Point estimate LME Box-Cox	814.4	872.6	0.426	0.432	0.220	0.233
Point estimate SEM Box-Cox 14 intervals	812.1	870.8	0.426	0.427	0.224	0.233
EFF	0.983	0.995	1.018	1.024	1.022	1.035
Point estimate SEM Box-Cox 8 intervals	810.5	863.9	0.421	0.428	0.219	0.232
EFF	1.028	1.013	1.029	1.038	1.043	1.046

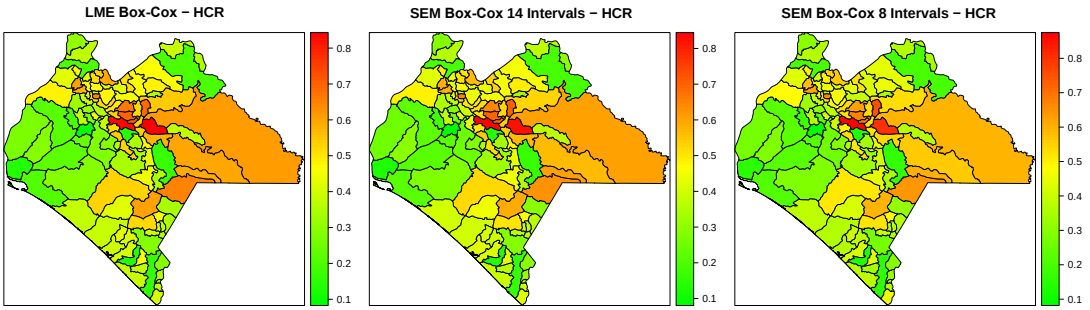


Figure 3.6: Estimated HCR for municipalities in the state of Chiapas based on different estimation methods.

A possible way to further validate the estimation results is by comparing the direct estimates to the model-based estimates. In Figure 3.7 the direct estimates of the mean (based on the observed continuous data) are compared to the model-based estimates (SEM Box-Cox) of the mean using an interval-censored response variable (14 intervals). As expected the left panel shows a positive linear correlation between the estimates. The right panel plots the value of the estimates for both estimation method for each in-sample domain. The pattern shows that as the sample size increases the direct estimates and the SEM Box-Cox estimates are almost identical. This is reasonable because the direct estimates gain precision with increasing sample size.

Figure 3.8 presents municipal estimates of mean income and PGAP for the SEM Box-Cox algorithm based on 14 intervals. The plots for the other estimation methods are omitted, because the results are comparable. We observe that municipalities in the middle and in the east of Chiapas exhibit high rates of HCRs and PGAPs and low levels of mean equivalized household labor income and are thus more strongly affected by poverty. These regions are characterized by high mountains, the Chiapas Highlands and a large concentration of indigenous population. There are, however, two regions in the center of the state with relatively high mean income and low rates of poverty. These are the regions where the capital Tuxtla Gutiérrez and the larger city San Cristóbal de las Casas are located. Also the coastal region – especially in the south –

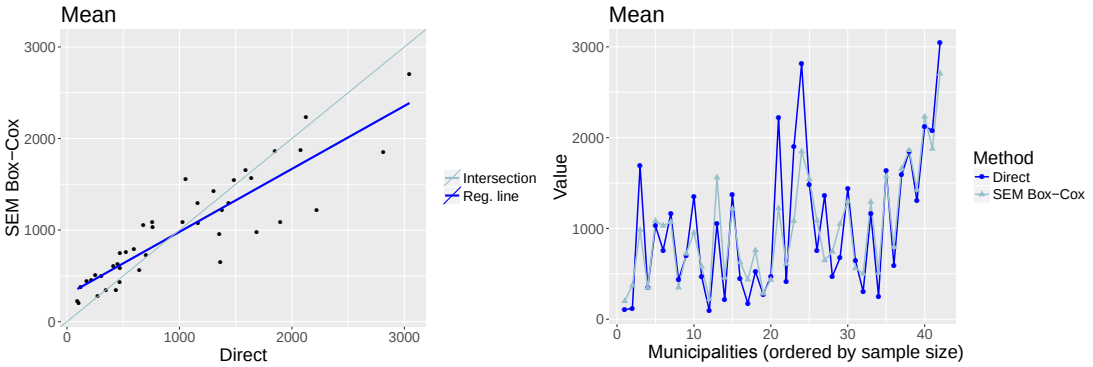


Figure 3.7: The left panel shows a scatter and the right panel a line plot of the direct and the model-based estimate for each in-sample domain (municipality).

where the economically most important city Tapachula is located, is affected less by poverty. The analysis shows that even though Chiapas is one of the poorest states in Mexico, there are spatial variations between the municipalities. These differences can be revealed by using SAE methods designed for grouped data. The proposed SEM Box-Cox method is – to the best of our knowledge – the first approach that allows the use of the popular EBP method in conjunction with a grouped response variable. This enables the estimation of spatially disaggregated target indicators with small sample sizes when confidentiality restrictions or decisions about the survey design require the use of relatively limited information for the response variable.

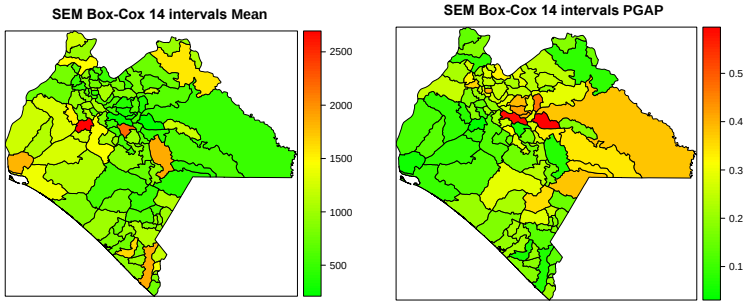


Figure 3.8: Estimated mean and PGAP for municipalities in the state of Chiapas.

3.6 Concluding remarks

The paper proposes small area estimation methodology when working with a response variable that is interval censored. The novel aspects of the paper include the estimation of a nested error regression model when the response is interval censored, the estimation both of linear and non-linear indicators for small areas, the use of data-driven transformation with the nested error regression model and the estimation of the MSE of the small area target parameters that accounts for the fact that we are working with limited information compared to standard small area models.

The proposed methods are evaluated using model-based simulations under different scenarios for the model error terms. The results show the proposed methods work well and in most scenarios the loss of accuracy in the estimates is small when compared to the use of EBPs that are estimated assuming the availability of full information for the response variable. As expected the loss of accuracy also depends on the number of intervals used for censoring the data and the proposed methodology appears to be working well even when the number of groups used is fairly small. The results also show that the use of an adaptive transformation works properly and the transformation parameter is estimated well in the presence of limited information for the response variable. Finally, the proposed MSE estimator appears to be capturing the different sources of variability and appropriately tracks the empirical MSE.

The new methodology is used to estimate disaggregated poverty and inequality indicators for municipalities in Chiapas, a southern state of Mexico, using interval-censored income grouped in eight and 14 intervals. In order to evaluate the proposed methodology estimates of the target parameters are also obtained when income is fully available, i.e., not interval censored. The Box-Cox transformation is applied to ensure that the model assumptions are met. The estimates from the continuous and grouped responses are very close, indicating the validity of the proposed methodology. The plotted poverty maps enable policy makers to get a spatial overview of the distribution of poverty in Chiapas and to target poorer regions more precisely.

Current research focuses on extending the SEM method for fitting nested error regression models for more complex structures, for example models with random slopes. Estimation of the standard errors of the fixed and random parameters and inference is of particular interest. In future research we plan to focus on the situation where interval censoring is also affecting some of the auxiliary variables. This is a more challenging problem but perhaps more realistic if interest is in protecting data confidentiality.

Acknowledgements

The research has received funding from the European Unions Horizon 2020 research and innovation programme under grant agreement No 730998, InGRID-2, Integrating Research Infrastructure for European expertise on Inclusive Growth from data to policy. Furthermore, Schmid and Tzavidis gratefully acknowledge support by grant ES/N011619/1 - Innovations in Small Area Estimation Methodologies from the UK Economic and Social Research Council. The authors are grateful to CONEVAL for providing the data used in empirical work. The views set out in this paper are those of the authors and do not reflect the official opinion of CONEVAL. The numerical results are not official estimates and are only produced for illustrating the methods.

3.7 Appendix

Table 3.8: Normal scenario 1 (14 intervals): Distribution of one arbitrarily chosen Monte Carlo population.

Interval	Frequencies
$[1, 1000)$	314
$[1000, 2000)$	656
$[2000, 2500)$	582
$[2500, 3000)$	785
$[3000, 3500)$	972
$[3500, 4000)$	1091
$[4000, 4500)$	1153
$[4500, 5000)$	1113
$[5000, 5500)$	940
$[5500, 6000)$	827
$[6000, 6500)$	608
$[6500, 7000)$	405
$[7000, 8000)$	400
$[8000, +\infty)$	154

Table 3.9: Normal scenario 2 (7 intervals): Distribution of one arbitrarily chosen Monte Carlo population.

Interval	Frequencies
$[1, 2000)$	970
$[2000, 3000)$	1367
$[3000, 4000)$	2063
$[4000, 5000)$	2266
$[5000, 6000)$	1767
$[6000, 7500)$	1265
$[7500, +\infty)$	302

Table 3.10: Log-scale scenario (7 intervals): Distribution of one arbitrarily chosen Monte Carlo population.

Interval	Frequencies
$[1, 500)$	1473
$[500, 1000)$	1703
$[1000, 2000)$	2113
$[2000, 4000)$	2093
$[4000, 8000)$	1453
$[8000, 16000)$	770
$[16000, +\infty)$	395

Table 3.11: Distribution of grouped equivalized household labor income (14 intervals).

Interval	Frequencies
$[1, 50)$	108
$[50, 100)$	104
$[100, 200)$	156
$[200, 400)$	234
$[400, 600)$	273
$[600, 1000)$	411
$[1000, 1500)$	325
$[1500, 2000)$	226
$[2000, 3000)$	247
$[3000, 4000)$	157
$[4000, 5500)$	92
$[5500, 8000)$	81
$[8000, 12000)$	48
$[12000, +\infty)$	24

Table 3.12: Distribution of grouped equivalized household labor income (8 intervals).

Interval	Frequencies
$[1, 100)$	212
$[100, 400)$	390
$[400, 1000)$	684
$[1000, 2000)$	551
$[2000, 4000)$	404
$[4000, 8000)$	173
$[8000, 1200)$	48
$[12000, +\infty)$	24

Part III

Implementation in the Programming Language R

Chapter 4

The R Package smicd: Statistical Methods for Interval-Censored Data

4.1 Introduction

Interval-censored or grouped data occurs when only the lower A_{k-1} and upper A_k interval bounds (A_{k-1}, A_k) of a variable are observed and its true value remains unknown. Instead of measuring the variable of interest on a continuous scale, for instance income data, the scale is divided into n_k intervals. The variable k $(1 \leq k \leq n_k)$ indicates in which of the n_k intervals an observation falls into. This leads to a loss of information since the shape of the distribution within the intervals remains unknown. In the field of survey statistics, asking for interval-censored data is often done in order to avoid item non-response and thus increase data quality. Item non-response is avoided because interval-censored data offers a higher level of data privacy protection (Hagenaars and Vos, 1988; Moore and Welniak, 2000). Among others, popular surveys and censuses that collect interval-censored data are the German Microcensus (Statistisches Bundesamt, 2017), the Colombian census (Departamento Administrativo Nacional De Estadística, 2005) and the Australian census (Australian Bureau of Statistics, 2011). While item non-response is reduced or avoided, the statistical analysis of the data requires more elaborate mathematical methods. Even statistical indicators that are easily calculated for metric data, e.g., the mean, cannot be estimated using standard formulas (Fahrmeir et al., 2011). Also, estimating linear and linear mixed regression models which are applied in many fields of statistics requires advanced statistical methods when the dependent variable is interval censored. Therefore, the presented R package (R Core Team, 2018) implements three major functions: `kdeAlgo()` to estimate statistical indicators (e.g., the mean) from interval-censored data, `semLm()` and `semLme()` to estimate linear and linear mixed regression models with an interval-censored dependent variable.

For the estimation of statistical indicators from interval-censored data different approaches are described in the literature. These approaches can be broadly categorized into four groups: Estimation on the midpoints (Fahrmeir et al., 2011), linear interpolation of the distribution function (Information und Technik (NRW), 2009), non-parametric modeling via splines (Berger

and Escobar, 2016) and fitting a parametric distribution function to the censored data (Dagum, 1977; McDonald, 1984; Bandourian et al., 2002). Some of these methods are implemented in R packages available on the Comprehensive R Archive Network (CRAN). The method of linear interpolation is implemented for the estimation of quantiles in the R package **actuar** (Dutang et al., 2008). The package also enables the estimation of the mean on the interval midpoints. Fitting a parametric distribution to interval-censored data can be done by using the R package **fitdistrplus** (Delignette-Muller and Dutang, 2015).

In survey statistics, interval-censored data is often collected for income or wealth variables. Thus, the performance of the above-mentioned methods is commonly evaluated by simulation studies that rely on data that follows some kind of income distribution. The German statistical office (DESTATIS) uses the method of linear interpolation for the estimation of statistical indicators from interval-censored income data collected by the German Microcensus (Information und Technik (NRW), 2009). This approach gives the same results as assuming a uniform distribution within the income intervals. Estimation results are reasonably accurate if the estimated indicators do not depend on the whole shape of the distribution, e.g., the median (Lenau and Münnich, 2016). Fitting a parametric distribution to the data enables the estimation of indicators that rely on the whole shape of the distribution. This method works well when the data is censored to only a few equidistant intervals (Lenau and Münnich, 2016). Non-parametric modeling via splines shows especially good results for a high number of intervals in ascending order (Lenau and Münnich, 2016). However, according to Lenau and Münnich (2016) all of the above-mentioned methods show large biases and variances when the estimation is based on a small number of intervals. Therefore, a novel kernel density estimation (KDE) algorithm is implemented in the **smicd** package that overcomes the drawbacks of the previously mentioned methods (Walter and Weimer, 2018). The algorithm bases the estimation of statistical indicators on pseudo samples that are drawn from a fitted non-parametric distribution. The method automatically adapts to the shape of the true unknown distribution and provides reliable estimates for different interval-censoring scenarios. It can be applied via the function `kdeAlgo()`.

Similarly to the direct estimation of statistical indicators from interval-censored data, there exists a variety of ad-hoc approaches and explicitly formulated mathematical methods for the estimation of linear regression models with an interval-censored dependent variable. The following methods and approaches are used for handling interval-censored dependent variables within linear regression models: Ordinary least squares (OLS) regression on the midpoints (Thompson and Nelson, 2003), ordered logit- or probit-regression (McCullagh, 1980), and regression methodology formulated for left-, right-, and interval-censored data (Tobin, 1958; Rosett and Nelson, 1975; Stewart, 1983). All of these methods are implemented in different R packages available on CRAN. OLS regression on the midpoints is applicable by using the `lm()` function from the **stats** Package (R Core Team, 2018), ordered logit regression is implemented in the **MASS** package (Venables and Ripley, 2002), and interval regression is implemented in the **IntReg** (Toomet, 2015) package.

While OLS regression on the midpoints of the intervals is easily applied, it comes with the

disadvantage of giving biased estimation results (Cameron, 1987). This approach disregards the uncertainty stemming from the unknown true distribution of the data within the intervals and therefore leads to biased parameter estimates. Its performance relies on the number of intervals and estimation results are only comparable to more advanced methods when the number of intervals is very large (Fryer and Pethybridge, 1972). Conceptualizing the model as an ordered logit or probit regression is feasible by treating the dependent variable as an ordered factor variable (McCullagh, 1980). However, this approach also neglects the unknown distribution of the data within the intervals. Furthermore, the predicted values are not on a continuous scale but are in terms of probability of belonging to a certain group. To overcome these disadvantages and obtain unbiased estimation results Stewart (1983) introduces regression methodology for models with an interval-censored dependent variable. Walter et al. (2017) further develop his approach and introduce a novel stochastic expectation-maximization (SEM) algorithm for the estimation of linear regression models with an interval-censored dependent variable that is implemented in the **smicd** package. The model parameters are unbiasedly estimated as long as the model assumptions are fulfilled. The function `semLm()` provides the SEM algorithm and enables the use of fixed (logarithmic) and data-driven (Box-Cox) transformations (Box and Cox, 1964). The Box-Cox transformation automatically adapts to the shape of the data and transforms the dependent variable in order to meet the model assumption.

In order to analyze longitudinal or clustered data (e.g., students within schools) linear mixed regression models are applicable. These kinds, of models control for the correlated structure of the data by including random effects in addition to the usual fixed effects. In order to deal with an interval-censored dependent variable in linear mixed regression models there are several approaches described in the literature. Linear mixed regression models, just like linear regression models, can be estimated on the interval midpoints of the censored-dependent variable. Furthermore, conceptualizing the model as an ordered logit or probit regression model is feasible (Agresti, 2010). These approaches inherit the same advantages and disadvantages as previously discussed. Linear mixed regression on the midpoints can be applied by the **lme4** (Pinheiro et al., 2017) or **nlme** (Bates et al., 2015) package and the ordered logit regression is implemented in the **ordinal** package (Christensen, 2015). To my knowledge, there are no R packages for the estimation of linear mixed regression models with an interval-censored dependent variable. Therefore, the package **smicd** contains the SEM algorithm proposed by Walter et al. (2017) for the estimation of linear mixed regression models with an interval-censored dependent variable. If the model assumptions are fulfilled, the method gives unbiased estimation results. The function `semLme()` enables the estimation of the regression parameters and it also allows for the usage of the logarithmic and Box-Cox transformation in order to fulfill the model assumptions (Gurka et al., 2006).

The paper is structured into two main sections. Section 4.2 deals with the direct estimation of statistical indicators from interval-censored data whereas Section 4.3 introduces linear and linear mixed regression models with an interval-censored dependent variable. Both sections have been divided into three subsections: first the statistical methodology is introduced, then the core functions of the **smicd** package are presented, and finally, illustrative examples with

two different data sets are provided. In Section 4.4 the main results are summarized and an outlook is given.

4.2 Direct estimation of statistical indicators

In the following three subsections, the methodology for the direct estimation of statistical indicators from interval-censored data is introduced, the core functionality of the function `kdeAlgo()` is presented and statistical indicators are estimated using the European Union Statistics on Income and Living Conditions (EU-SILC) data set (European Commission, 2013).

4.2.1 Methodology: Direct estimation of statistical indicators

In order to estimate statistical indicators from interval-censored data the proposed algorithm generates metric pseudo samples of an interval-censored variable. These pseudo samples can be used to estimate any statistical indicator. They are drawn from a non-parametrically estimated kernel density. Kernel density estimation was first introduced by Rosenblatt (1956) and Parzen (1962). By its application the density $f(x)$ of a continuous independently and identically distributed random variable is estimated without assuming any distributional shape of the data. The estimator is defined as

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right), \qquad i = 1, \ldots, n,$$

where $K(\cdot)$ is a kernel function, $h > 0$ the bandwidth and $x = \{x_1, x_2, \ldots, x_n\}$ denotes a sample of size n. The performance of the estimator is determined by the optimal choice of h. The selection of an optimal h is widely discussed in the literature, see Jones et al. (1996); Loader (1999); Zambom and Dias (2012). When working with interval-censored data, a standard KDE cannot be applied since x is not observed on a continuous scale. Nevertheless, its unobserved true distribution is of continuous form. As an ad hoc solution the density $\hat{f}_h(x)$ can be estimated based on the interval midpoints. The resulting density estimate will be spiky unless the bandwidth is sufficiently large. A large bandwidth, however, leads to a loss of information (Wang and Wertelecki, 2013). Therefore, Walter and Weimer (2018) propose an iterative KDE algorithm for density estimation from interval-censored data. The approach is based on Groß et al. (2017) who introduce a similar KDE algorithm in a two-dimensional setting with an equidistant interval width. Walter and Weimer (2018) show that the algorithm can be adjusted to one-dimensional data with an arbitrary class width. For the estimation of linear and non-linear statistical indicators the unknown distribution of x has to be reconstructed by using the observed interval $k = \{k_1, k_2, \ldots, k_n\}$ that an observation falls into. From Bayes' theorem (Bayes, 1763) it follows that the conditional distribution of $x|k$ is:

$$\pi(x|k) \propto \pi(k|x)\pi(x)$$

with $\pi(k|x)$ is defined by a product of a Dirac distribution $\pi(k|x) = \prod_{i=1}^{n} \pi(k_i|x_i)$ with

$$\pi(k_i|x_i) = \begin{cases} 1 & \text{if } A_{k-1} \leq x_i \leq A_k, \\ 0 & \text{else,} \end{cases}$$

for $i = 1, \ldots, n$. Since $\pi(x)$ is unknown it is replaced by a kernel density estimate $\hat{f}_h(x)$.

Estimation and computational details

To fit the model, pseudosamples of x_i are drawn from the conditional distribution

$$\pi(x_i|k_i) \propto \mathbf{I}(A_{k-1} \leq x_i \leq A_k) f(x_i),$$

where $\mathbf{I}(\cdot)$ denotes the indicator function. The conditional distribution of $\pi(x_i|k_i)$ is given by the product of a uniform distribution and density $f(x_i)$. As the density is unknown it is replaced by an estimate $\hat{f}_h(x)$, which is obtained by the KDE. In particular, x_i is repeatedly drawn from the given interval (A_{k-1}, A_k) by using the current density estimate $\hat{f}_h(x)$ as a sampling weight. The explicit steps of the iterative algorithm as given in Walter and Weimer (2018) are stated below:

1. Use the midpoints of the intervals as pseudo $\tilde{x}_i$ for the unknown x_i. Estimate a pilot estimate of $\hat{f}_h(x)$, by applying KDE. Note: Choose a sufficiently large bandwidth h in order to avoid rounding spikes.

2. Evaluate $\hat{f}_h(x)$ on an equal-spaced grid $G = \{g_1, \ldots, g_j\}$ with grid points $g_1, \ldots, g_j$. The width of the grid is denoted by δ_g. It is given by

   $$\delta_g = \frac{|A_0 - A_{n_k}|}{j - 1},$$

   and the grid is defined as:

   $$G = \{g_1 = A_0, g_2 = A_0 + \delta_g, g_3 = A_0 + 2\delta_g, \ldots, g_{j-1} = A_0 + (j-2)\delta_g, g_j = A_{n_k}\}.$$

3. Sample from $\pi(x|k)$ by drawing randomly from $G_k = \{g_j | g_j \in (A_{k-1}, A_k)\}$ with sampling weights $\hat{f}_h(\tilde{x}_i)$ for $k = 1, \ldots, n_k$. The sample size for each interval is given by the number of observations within each interval. Obtain $\tilde{x}_i$ for $i = 1, \ldots, n$.

4. Estimate any statistical indicator of interest $\hat{I}$ using $\tilde{x}_i$.

5. Recompute the density $\hat{f}_h(x)$, using the pseudo samples $\tilde{x}_i$ obtained in iteration Step 3.

6. Repeat Steps 2-5, with $B^{(KDE)}$ burn-in and $M^{(KDE)}$ additional iterations.

7. Discard the $B^{(KDE)}$ burn-in iterations and estimate the final $\hat{I}$ by averaging the obtained $M^{(KDE)}$ estimates.

For open-ended intervals, e.g., $(15000, +\infty)$ the upper bound has to be replaced by a finite number. Walter and Weimer (2018) show through model-based simulations that a value of three times the value of the lower bound $(15000, 45000)$ gives appropriate estimation results when working with income data.

The variance of the statistical indicators is estimated by bootstrapping. Bootstrap methods were first introduced by Efron (1979). These methods serve as an estimation procedure when the variance cannot be stated as closed-form solution (Shao and Tu, 1995). While bootstrapping avoids the problem of the non-availability of a closed-form solution, it comes with the disadvantage of long computational times. In the package, a non-parametric bootstrap that accounts for the additional uncertainty coming from the interval-censored data is implemented. This non-parametric bootstrap is introduced in Walter and Weimer (2018).

4.2.2 Core functionality: Direct estimation of statistical indicators

The presented KDE algorithm is implemented in the function `kdeAlgo()` (see Table 4.1). The arguments and default settings of `kdeAlgo()` are briefly summarized in Table 4.2. The function gives back an S3 object of class `"kdeAlgo."` A detailed explanation of all components of an `"kdeAlgo"` object can be found in the package documentation. The generic functions `plot()` and `print()` can be applied to `"kdeAlgo"` objects to output the main estimation results (see Table 4.1). In the next section the function `kdeAlgo()` is used to estimate a variety of statistical indicators from interval-censored EU-SILC data and its arguments are explained in more detail.

Table 4.1: Implemented functions for the direct estimation of statistical indicators.

Function Name	Description
`kdeAlgo()`	Estimates statistical indicators and its standard errors from interval-censored data
`plot()`	Plots convergence of the estimated statistical indicators and estimated density of the pseudo $\tilde{x}_i$
`print()`	Prints estimated statistical indicators and its standard errors

4.2.3 Example: Direct estimation of statistical indicators

To demonstrate the function `kdeAlgo()`, the total disposable household income and the corresponding household weight from the public use file (PUF) of the European Union Statistics on Income and Living Condition (EU-SILC) data set is used (European Commission, 2013). The PUF is a fully synthetic data set which cannot be used for inferential statistics. Nevertheless, the distribution of the data mimics the distribution of the original data set (Eurostat, 2018). The PUF has the advantage (over the scientific use file) of being easily available on the Eurostat website (Eurostat, 2018). The analysis is carried out using the German PUF from 2013. After the deletion of missing values there are 12,703 observations left in the EU-SILC survey that are used in the analysis. Since the total disposable household income is measured

Table 4.2: Arguments of function `kdeAlgo()`.

Argument	Description	Default
`xclass`	Interval-censored variable	
`classes`	Numeric vector of interval bounds	
`threshold`	Threshold used for poverty indicators (60% of the median of the target variable)	0.6
`burnin`	Number of burn-in iterations $B^{(KDE)}$	80
`samples`	Number of additional iterations $M^{(KDE)}$	400
`bootstrap.se`	If TRUE, standard errors of the statistical indicators are estimated	FALSE
`b`	Number of bootstraps for the estimation of the standard errors	100
`bw`	Smoothing bandwidth used	"nrd0"
`evalpoints`	Number of evaluation grid points	4000
`adjust`	Bandwidth multiplier $bw = adjust * bw$	1
`custom_indicator`	A list of user-defined statistical indicators	NULL
`upper`	If upper bound of the upper interval is $+\infty$ e.g., $(15000, +\infty)$, then $+\infty$ is replaced by $15000 * upper$	3
`weights`	Survey weights	NULL
`oecd`	Household weights of equivalence scale	NULL

on a continuous scale, it is censored to 24 intervals for demonstration purposes. For a realistic censoring scheme the interval bounds are chosen such that they match the interval bounds used in the German Microcensus from 2013 (Statistisches Bundesamt, 2014). The German Microcensus is a representative household survey that covers 830,000 persons in 370,000 households (1% of the German population) in which income is only collected as interval-censored variable (Statistisches Bundesamt, 2016).

In a first step the variable total disposable household income called `hhincome_net` is interval censored according to the 24 intervals in the German Microcensus using the function `cut()`. The vector of interval bounds is called `intervals` and the newly obtained interval-censored income variable is called `c.hhincome`.

```
R> intervals <- c(0,150,300,500,700,900,1100,1300,1500,1700,
+    2000,2300,2600,2900,3200,3600,4000,4500,5000,5500,6000,
+    7500,10000,18000,Inf)
R> c.hhincome <- cut(hhincome_net, breaks = intervals)
```

In order to get a descriptive overview of the distribution of the censored income data the function `table()` is applied.

```
R> table(c.hhincome)

c.hhincome
          (0,150]         (150,300]         (300,500]
```

```
             229              283              442
       (500,700]        (700,900]    (900,1.1e+03]
             532              576              609
(1.1e+03,1.3e+03] (1.3e+03,1.5e+03] (1.5e+03,1.7e+03]
             570              555              586
  (1.7e+03,2e+03]   (2e+03,2.3e+03] (2.3e+03,2.6e+03]
             819              744              673
(2.6e+03,2.9e+03] (2.9e+03,3.2e+03] (3.2e+03,3.6e+03]
             612              604              685
  (3.6e+03,4e+03]   (4e+03,4.5e+03]   (4.5e+03,5e+03]
             510              587              461
  (5e+03,5.5e+03]   (5.5e+03,6e+03]   (6e+03,7.5e+03]
             375              279              536
  (7.5e+03,1e+04]   (1e+04,1.8e+04]    (1.8e+04,Inf]
             392              198               23
```

Most incomes are in interval $(1700, 2000]$ and only 23 incomes are in the upper interval. For the estimation of the statistical indicators the function `kdeAlgo()` of the `smicd` package is called with the following arguments.

```
R> Indicators <- kdeAlgo(xclass = c.hhincome, classes =
+    intervals, bootstrap.se = TRUE, custom_indicator =
+    list(quant05 = function(y, treshold, weights)
+    {wtd.quantile(y, probs = 0.05, weights)}, quant95 =
+    function(y, treshold, weights){wtd.quantile(y, probs =
+    0.95, weights)}), weights = hhweight)
```

The variable `c.hhincome` is assigned to the argument `xclass` and the vector of interval bounds `intervals` is assigned to the argument `classes`. The default settings of the arguments `burnin`, `samples`, `bw`, `evalpoints`, `adjust` and `upper` are retained. Simulation results from Walter and Weimer (2018) and Groß et al. (2017) show that these settings give good results when working with income data. Changing these arguments has an impact on the performance of the KDE algorithm. As default, the statistical indicators: mean, Gini coefficient, headcount ratio (HCR), the quantiles (10%, 25%, 50%, 75%, 90%), the poverty gap (PGAP) and the quintile share ratio (QSR) are estimated (Gini, 1912; Foster et al., 1984). The HCR and PGAP rely on a poverty threshold. The default choice of the `threshold` argument is 60% of the median of the target variable as suggested by Eurostat (2014). Besides the mentioned indicators, any other statistical indicator can be estimated via the argument `custom_indicator`. In the example the argument is assigned a list that holds functions to estimate the 5% and 95% quantile. The custom indicators must depend on the target variable, the threshold (even if it is not needed for the specified indicator) and optionally on the weights argument, if the estimation of a weighted indicator is required. To estimate

the standard errors of all indicators `bootstrap.se = TRUE` and the number of bootstrap samples is 100 (the default value as suggested in Walter and Weimer (2018)). Lastly, the household weight (`hhweight`) is assigned to the argument `weights` in order to estimate weighted statistical indicators. It can also be controlled for households of different sizes, by assigning `oecd` a variable with household equivalence weights. By applying the `print()` function to the `"kdeAlgo"` object the estimated statistical indicators (default and custom indicators) as well as their standard errors are printed. For instance, in this example the estimated mean is about 2,916 Euro and its standard error is 23.124.

```
R> print(Indicators)
```

```
Value:
    mean    gini     hcr  quant10   quant25   quant50
2916.041   0.425   0.289  591.783  1203.239  2295.574
 quant75   quant90    pgap      qsr  quant05   quant95
3901.166  5935.196   0.131  11.929  343.548  7583.327

Standard error:
   mean    gini     hcr quant10 quant25 quant50
 23.124   0.004   0.003  11.050  15.289  25.819
quant75 quant90    pgap     qsr quant05 quant95
 38.855  57.051   0.002  0.251  11.451  82.597
```

In Walter and Weimer (2018) the performance of the KDE algorithm is evaluated via detailed simulation studies. By applying the function `plot()` `"kdeAlgo"` objects can be plotted. Thereby, convergence plots for all estimated statistical indicators and a plot of the estimated final density are obtained.

```
R> plot(Indicators)
```

Figure 4.1 shows convergence plots for three of the estimated indicators (panel 1-3). Additionally, a plot of the estimated final density with a histogram of the observed data in the background (panel 4) is obtained. In panel 1-3 the estimated statistical indicator (HCR, PGAP, 75% quantile) is plotted for each iteration step of the KDE algorithm. A vertical line marks the end of the burn-in period. All convergence plots in Figure 4.1 demonstrate that the number of iterations is chosen sufficiently large for the estimates to converge. If convergence were not achieved the arguments `burnin` and `samples` should be increased. It is notable that the estimated 75% quantile has the same value for almost all iterations steps. This is the case because the quantile, as any other statistical indicator, is estimated using the pseudo samples that are drawn on 4,000 grid points G. Estimating a quantile based on 12,703 observations on only 4,000 unique outcomes (pseudo values) leads to equal quantile estimates for almost all iteration steps of the KDE algorithm.

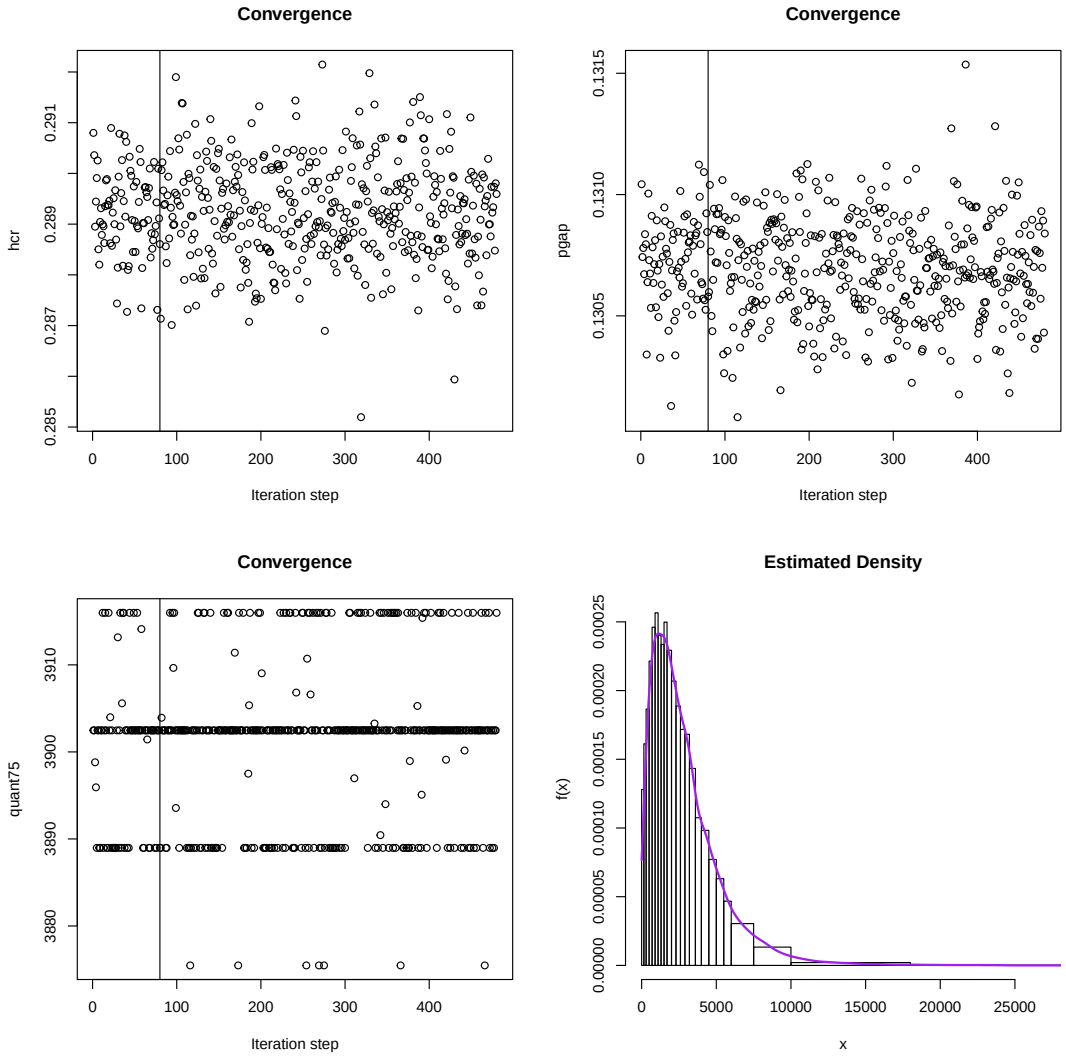


Figure 4.1: Convergence plots of the statistical indicators and a plot of the estimated final density with a histogram of the observed distribution of the data in the background.

4.3 Regression analysis

In the following three subsections the statistical methodology for linear and linear mixed regression models with an interval-censored dependent variable is introduced, the core functionality of the functions `semLM()` and `semLME()` is presented and examination scores of students from schools in London are exemplary modeled.

4.3.1 Methodology: Regression analysis

The theoretical introduction of the new regression method, proposed by Walter et al. (2017), is presented for linear mixed regression models. The theory for linear regression models can be obtained by simplifying the introduced method. In its standard form the linear mixed regression model serves to analyze the linear relationship between a continuous dependent variable and

some independent variables (Goldstein, 2003). Random parameters (random slopes and random intercepts) are included in the model to account for correlated data e.g., students within schools. The model in matrix notation (Laird and Ware, 1983) is given by

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v} + \mathbf{e}, \tag{4.1}$$

where $\mathbf{y}$ is a $n \times 1$ column vector of the dependent variable, n is the sample size, $\mathbf{X}$ is a $n \times p$ matrix where p is equal to the number of predictors, $\boldsymbol{\beta}$ is a column vector of the fixed effects regression parameters of size $p \times 1$, $\mathbf{Z}$ is the $n \times q$ design matrix with q random effects, $\mathbf{v}$ is a $q \times 1$ vector of random effects, and $\mathbf{e}$ is the residual vector of size $n \times 1$. The distribution of the random effects is given by

$$\mathbf{v} \sim N(\mathbf{0}, \mathbf{G}), \qquad \text{where } \mathbf{G} = \begin{bmatrix} \sigma_0^2 & \sigma_{01} & \dots & \sigma_{0q} \\ \sigma_{10} & \sigma_1^2 & \dots & \sigma_{1q} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{q0} & \sigma_{q1} & \dots & \sigma_q^2 \end{bmatrix},$$

and the distribution of the residuals is given by $\mathbf{e} \sim N(\mathbf{0}, \mathbf{R})$ with $\mathbf{R} = \mathbf{I}_n\sigma_e^2$ where $\mathbf{I}_n$ is the identity matrix and σ_e^2 is the residual variance. The random effects $\mathbf{v}$ and the residuals $\mathbf{e}$ are assumed to be independent. For a more detailed introduction of mixed models see Searle et al. (1992); McCulloch et al. (2008); Snijders and Bosker (2011). In the case of an interval-censored dependent variable the parameters of Model (4.1) have to be estimated without observing $\mathbf{y}$ on a continuous scale. Instead, only the interval identifier $\mathbf{k}$, now defined as $n \times 1$ column vector, is observed. Open-ended interval bounds $A_0 = -\infty$ and $A_{n_k} = +\infty$ and unequal interval widths are allowed. Since the true distribution of $\mathbf{y}$ is unknown the aim is to reconstruct the distribution of $\mathbf{y}$ using the known intervals $\mathbf{k}$ and the linear relationship stated in Model (4.1). As presented in Walter et al. (2017) in order to reconstruct the unknown distribution of $f(\mathbf{y}|\mathbf{X}, \mathbf{Z}, \mathbf{v}, \mathbf{k}, \boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\boldsymbol{\beta}, \mathbf{R}, \mathbf{G})$, the Bayes theorem (Bayes, 1763) is applied. Hence,

$$f(\mathbf{y}|\mathbf{X}, \mathbf{Z}, \mathbf{v}, \mathbf{k}, \boldsymbol{\theta}) \propto f(\mathbf{k}|\mathbf{y}, \mathbf{X}, \mathbf{Z}, \mathbf{v}, \boldsymbol{\theta}) f(\mathbf{y}|\mathbf{X}, \mathbf{Z}, \mathbf{v}, \boldsymbol{\theta}),$$

with $f(\mathbf{k}|\mathbf{y}, \mathbf{X}, \mathbf{Z}, \mathbf{v}, \boldsymbol{\theta}) = f(\mathbf{k}|\mathbf{y})$ because the conditional distribution of the interval identifier $\mathbf{k}$ only depends on $\mathbf{y}$. It is given by $f(\mathbf{k}|\mathbf{y}) = \mathbf{r}$ with $\mathbf{r}$ being a $n \times 1$ column vector $\mathbf{r} = (r_1, r_2, \dots, r_n)^T$ with

$$r_i = \begin{cases} 1 & \text{if } A_{k-1} \leq y_i \leq A_k, \\ 0 & \text{else,} \end{cases}$$

for $(i = 1, \dots, n)$ and

$$f(\mathbf{y}|\mathbf{X}, \mathbf{Z}, \mathbf{v}, \boldsymbol{\theta}) \sim \ N(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v}, \mathbf{R}). \tag{4.2}$$

The relationship in Equation (4.2) follows from the linear mixed model assumptions (Model (4.1)). The unknown parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}, \mathbf{R}, \mathbf{G})$ are estimated based on pseudo samples $\tilde{\mathbf{y}}$

(since $\mathbf{y}$ is unknown) that are iteratively drawn from $f(\mathbf{y}|\mathbf{X}, \mathbf{Z}, \mathbf{v}, \mathbf{k}, \boldsymbol{\theta})$. The next subsection states the computational details of the SEM algorithm.

Estimation and computational details

To fit Model (4.1), the parameter vector $\hat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\beta}}, \widehat{\mathbf{R}}, \widehat{\mathbf{G}})$ is estimated and pseudo samples of the unknown $\mathbf{y}$ are iteratively generated by the following SEM algorithm. The pseudo samples $\tilde{\mathbf{y}}$ are drawn from the conditional distribution

$$f(\mathbf{y}|\mathbf{X}, \mathbf{Z}, \mathbf{v}, \mathbf{k}, \boldsymbol{\theta}) \propto \mathbf{I}(A_{\mathbf{k}-1} \leq \mathbf{y} \leq A_{\mathbf{k}}) \times N(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v}, \mathbf{R}),$$

where $\mathbf{I}(\cdot)$ denotes the indicator function. Hence, for $\mathbf{y}$ with explanatory variables $\mathbf{X}$ the corresponding $\tilde{\mathbf{y}}$ is drawn from $N(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v}, \mathbf{R})$ conditional on the given interval $(A_{\mathbf{k}-1} \leq \mathbf{y} \leq A_{\mathbf{k}})$. If $\hat{\boldsymbol{\theta}}$ is estimated the conditional distribution $f(\mathbf{y}|\mathbf{X}, \mathbf{Z}, \mathbf{v}, \mathbf{k}, \boldsymbol{\theta})$ follows a two-sided truncated normal distribution. Its probability density function equals

$$\hat{f}(\mathbf{y}|\mathbf{X}, \mathbf{Z}, \hat{\mathbf{v}}, \mathbf{k}, \hat{\boldsymbol{\theta}}) = \frac{\phi\left(\frac{\mathbf{y}-\hat{\boldsymbol{\mu}}}{\hat{\sigma}_e}\right)}{\hat{\sigma}_e\left(\Phi\left(\frac{A_{\mathbf{k}}-\hat{\boldsymbol{\mu}}}{\hat{\sigma}_e}\right) - \Phi\left(\frac{A_{\mathbf{k}-1}-\hat{\boldsymbol{\mu}}}{\hat{\sigma}_e}\right)\right)}, \tag{4.3}$$

with $\hat{\boldsymbol{\mu}} = \mathbf{X}\widehat{\boldsymbol{\beta}} + \mathbf{Z}\hat{\mathbf{v}}$. $\phi(\cdot)$ denotes the probability density function of the standard normal distribution and $\Phi(\cdot)$ denotes its cumulative distribution function. From its definition it follows that $\Phi\left(\frac{A_{\mathbf{k}}-\hat{\boldsymbol{\mu}}}{\hat{\sigma}_e}\right) = 1$ if $A_{\mathbf{k}} = +\infty$ and $\Phi\left(\frac{A_{\mathbf{k}-1}-\hat{\boldsymbol{\mu}}}{\hat{\sigma}_e}\right) = 0$ if $A_{\mathbf{k}-1} = -\infty$. The steps of the SEM algorithm as described in Walter et al. (2017) are:

1. Estimate $\hat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\beta}}, \widehat{\mathbf{R}}, \widehat{\mathbf{G}})$ from Model (4.1) using the midpoints of the intervals as substitutes for the unknown $\mathbf{y}$. The parameters are estimated by restricted maximum likelihood theory (REML) (Thompson, 1962).

2. **Stochastic step:** For $i = 1, \ldots, n$, draw randomly from $N(\mathbf{X}\widehat{\boldsymbol{\beta}} + \mathbf{Z}\hat{\mathbf{v}}, \widehat{\mathbf{R}})$ within the given interval $(A_{\mathbf{k}-1} \leq \mathbf{y} \leq A_{\mathbf{k}})$ (the two-sided truncated normal distribution given in Equation (4.3)) obtaining $(\tilde{\mathbf{y}}, \mathbf{X}, \mathbf{Z})$. The drawn pseudo $\tilde{\mathbf{y}}$ are used as replacements for the unobserved $\mathbf{y}$.

3. **Maximization step:** Re-estimate the parameter vector $\hat{\boldsymbol{\theta}}$ from Model (4.1) by using the pseudo samples $(\tilde{\mathbf{y}}, \mathbf{X}, \mathbf{Z})$ from Step 2. Again, parameter estimation is carried out by REML.

4. Iterate Steps 2-3 $B^{(SEM)}+M^{(SEM)}$ times, with $B^{(SEM)}$ burn-in iterations and $M^{(SEM)}$ additional iterations.

5. Discard the burn-in iterations $B^{(SEM)}$ and estimate $\hat{\boldsymbol{\theta}}$ by averaging the obtained $M^{(SEM)}$ estimates.

If open-ended intervals $A_0 = -\infty$ and $A_{n_k} = +\infty$ are present, the midpoints M_1 and M_{n_k} of these intervals in iteration Step 1 are computed as follows:

$$M_1 = (A_1 - \overline{D})/2,$$
$$M_{n_k} = (A_{n_k-1} + \overline{D})/2,$$

where

$$\overline{D} = \frac{1}{(n_k - 2)} \sum_{k=2}^{n_k-1} |A_{k-1} - A_k|.$$

These midpoints serve as proxies for the unknown interval midpoints in Step 1 of the algorithm. The SEM algorithm for the linear regression model is obtained by simplifying the conditional distribution $f(\mathbf{y}|\mathbf{X}, \mathbf{Z}, \mathbf{v}, \boldsymbol{\theta}) \sim \ N(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v}, \mathbf{R})$ to $f(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma_e^2) \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma_e^2)$ according to the model assumptions of a linear regression model. In the SEM algorithm for linear models it is then drawn from $N(\mathbf{X}\boldsymbol{\beta}, \sigma_e^2)$ within the given interval.

The standard errors of the regression parameters are estimated using bootstrap methods. For the linear regression model a non-parametric bootstrap (Efron and Stein, 1981; Efron, 1982; Efron and Tibshirani, 1986, 1993) and for the linear mixed regression model a parametric bootstrap (Wang et al., 2006; Thai et al., 2013) is used to estimate the standard errors. The non-parametric as well as the parametric bootstrap are further developed to account for the additional uncertainty that is due to the interval-censored dependent variable. Both newly proposed bootstraps are available in the **smicd** package.

To assure that the model assumptions are fulfilled the logarithmic and the Box-Cox transformations are incorporated into the function `semLm()` and `semLme()`.

4.3.2 Core functionality: Regression analysis

The introduced SEM algorithm is implemented in the functions described in Table 4.3. The arguments and default settings of the estimation functions `semLm()` and `semLme()` are summarized in Table 4.4. Both functions return a an S3 object of class `"sem" "lm"` or `"sem" "lme"`. A detailed explanation of all the components of these objects can be found in the **smicd** package documentation. The generic functions `plot()`, `print()` and `summary()` can be applied to objects of class `"sem" "lm"` and `"sem" "lme"` in order to summarize the main estimation results. In the next section the functionality of `semLm()` and `semLme()` is demonstrated based on an illustrative example.

4.3.3 Example: Regression analysis

To demonstrate the functions `semLm()` and `semLme()` the famous London school data set that is analyzed in Goldstein et al. (1993) is used. The data set contains the examination results of 4,059 students from 65 schools in six Inner London Education Authorities. The data set is available in the R package **mlmRev** (Bates et al., 2014) and also included in the package **smicd**. The variables used in the following example are: general certificate of secondary ex-

Table 4.3: Implemented functions for the estimation of linear and linear mixed regression models.

Function Name	Description
`semLm()`	Estimates linear regression models with an interval-censored dependent variable
`semLme()`	Estimates linear mixed regression models with an interval-censored dependent variable
`plot()`	Plots convergence of the estimated parameters and estimated density of the pseudo $\tilde{\mathbf{y}}$ from the last iteration step
`print()`	Prints basic information of the estimated linear and linear mixed regression models
`summary()`	Summary of the estimated linear and linear mixed regression models

Table 4.4: Arguments of functions `semLm()` and `semLme()`.

Argument	Description	Default
`formula`	A two-sided linear formula object	
`data`	A data frame containing the variables of the model	
`classes`	Numeric vector of interval bounds	
`burnin`	Burn-in iterations	40
`samples`	Additional iterations	200
`trafo`	Transformation of the dependent variable: None, logarithmic or Box-Cox transformation	"None"
`adjust`	Extends the number of iterations for the estimation of the Box-Cox transformation parameter: $(burnin + samples) * adjust$	2
`bootstrap.se`	If TRUE standard errors and confidence intervals of the regression parameters are estimated	FALSE
`b`	Number of bootstraps for the estimation of the standard errors	100

amination scores (`examsc`), the standardized London reading test scores at the age of 11 years (`standLRT`), the sex of the student (`sex`), and the school identifier (`school`). In the original data set the variable `examsc` is measured on a continuous scale. In order to demonstrate the functionality of the functions `semLm()` and `semLme()` the variable is arbitrarily censored to nine intervals. As before, the censoring is carried out by the function `cut()` and the vector of interval bounds is called `intervals`.

```
R> intervals <- c(1,1.5,2.5,3.5,4.5,5.5,6.5,7.7,8.5,Inf)
R> Exam$examsc.class<- cut(Exam$examsc, intervals)
```

The newly created interval-censored variable is called `examsc.class`. The distribution is visualized by applying the function `table()`.

```
R> table(Exam$examsc.class)

  (1,1.5] (1.5,2.5] (2.5,3.5] (3.5,4.5] (4.5,5.5] (5.5,6.5]
        1        32       249       937      1606       951
```

```
(6.5,7.7] (7.7,8.5] (8.5,Inf]
      267        15         1
```

It can be seen that most examination scores are concentrated in the center intervals. To fit the linear regression model the function `semLM()` is called.

```
R> LM <- semLm(formula = examsc.class ~ standLRT + sex,
+    data = Exam, classes = intervals,  bootstrap.se = TRUE)
```

The formula argument is assigned the model equation, where `examsc.class` is regressed on `standLRT` and `sex`. The argument `data` is assigned the name of the data set `Exam` and the vector of interval bounds `intervals` is assigned to the `classes` argument. The arguments `burnin` and `samples` are left as defaults. The specified number of default iterations is sufficiently large for most regression models, however, convergence of the parameters has to be checked by plotting the estimation results with the function `plot()` after the estimation. No transformation is specified for the interval-censored dependent variable therefore `trafo` is assigned its default value. The argument `adjust` is only relevant if the Box-Cox transformation `trafo="bc"` is chosen. In this case the number of iterations for the estimation of the Box-Cox transformation parameter λ can be specified by this argument. The convergence of the transformation parameter λ has to be checked using the function `plot()`. More information on the Box-Cox transformation and on the estimation of the transformation parameter is given in Walter et al. (2017). For the estimation of the standard errors of the regression parameters the argument `bootstrap.se` is set to `TRUE`. The number of bootstrap samples `b` is 100, its default value, which again is reasonable for most settings. A summary of the estimation results is obtained by the application of the function `summary()`.

```
R> summary(LM)

Call:
semLm(formula = examsc.class ~ standLRT + sex, data = Exam,
     classes = intervals, bootstrap.se = TRUE)

Fixed effects:
            Estimate Std.Error Lower 95%-level Upper 95%-level
(Intercept) 5.069695  0.0176955       5.029111        5.106293
standLRT    0.590856  0.0125097       0.565046        0.613674
sexM       -0.171377  0.0269704      -0.237042       -0.114465

Multiple R-squared: 0.3501  Adjusted R-squared: 0.3498
Variable examsc.class is divided into 9 intervals.
```

The output shows the function call, the estimated regression coefficients, the bootstrapped standard errors, and the confidence intervals as well as the R-squared and the adjusted R-squared. Furthermore, the output reminds the user that the dependent variable is censored to

nine intervals. All estimates are interpreted as in a linear regression model with a continuous dependent variable, hence, if `standLRT` increases by one unit and all other parameters are kept constant, `examsc.class` increases by 0.59 on average. The bootstraped confidence intervals indicate that all regressors have a significant effect on the dependent variable.

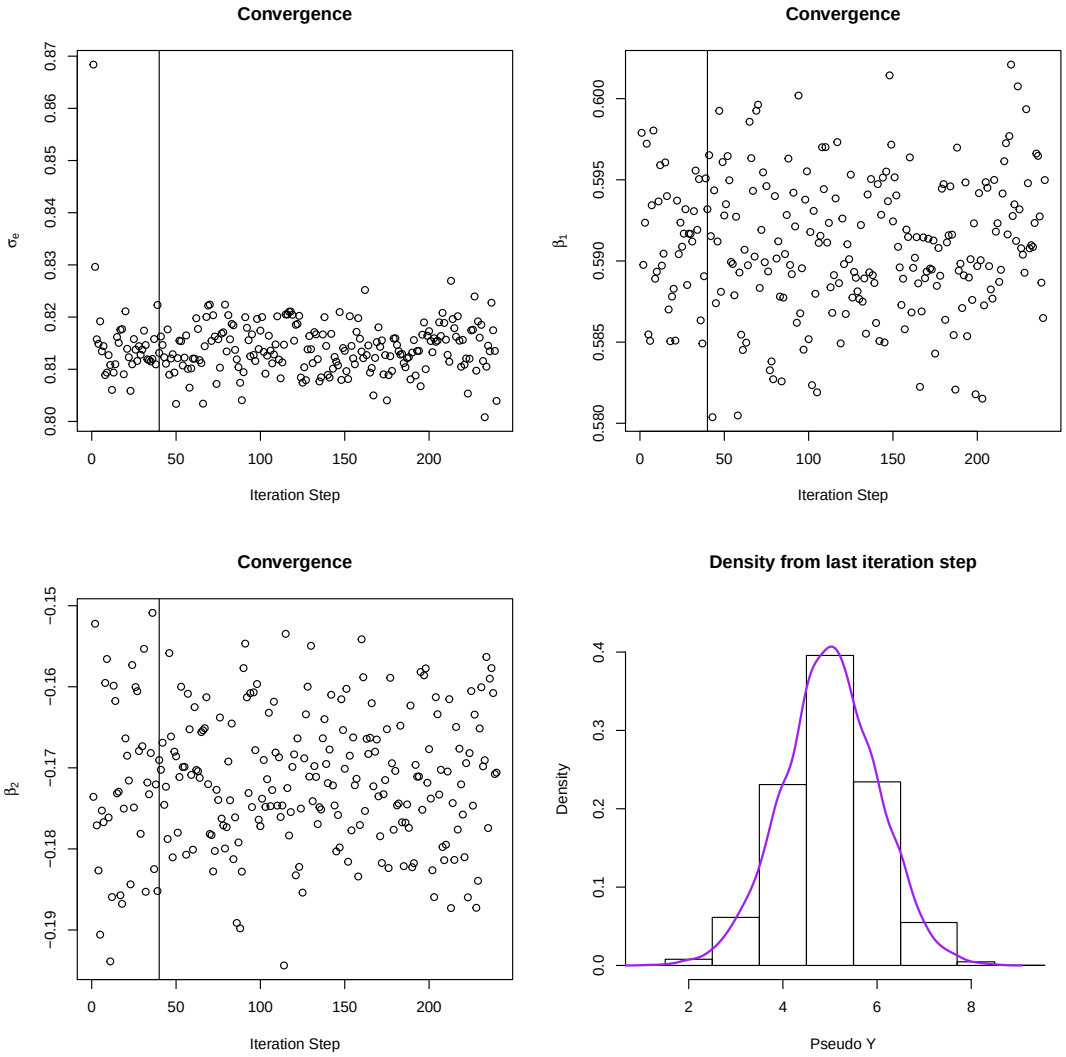


Figure 4.2: Convergence plots of estimated model parameters and the estimated final density with a histogram of the observed distribution of the data in the background.

By using the generic function `plot()` on an object of class `"sem" "lm"` convergence plots of each estimated regression parameter and of the estimated residual variance are obtained. Furthermore, the density of the generated pseudo $\tilde{\mathbf{y}}$ variable from the last iteration step is plotted with a histogram of the observed distribution of the interval-censored variable `examsc.class` in the background.

```
R> plot(LM)
```

In Figure 4.2 a selection of convergence plots is given in panel 1-3 and the density of the pseudo $\tilde{\mathbf{y}}$ from the last iteration step of the SEM algorithm is given in panel 4. The estimated

parameter is plotted for each iteration step of the SEM algorithm. A vertical line indicates the end of the burn-in period (40 iterations). The final parameter estimate is obtained by averaging the $M^{(SEM)}$ additional iterations (200). The selected 240 iterations are enough to obtain reliable estimates in this example, because the estimates have converged.

As already mentioned, the **smicd** package also enables the estimation of linear mixed regression models by the function `semLme()`. In the London school data set students are nested within schools, therefore it is necessary to control for the correlation within-schools. In order to do that the variable `school` is specified as a random intercept. Furthermore, a random slope parameter on the standardized London reading test score `standLRT` is included in the model to allow for different slopes. Again, the variable `sex` is included as an additional regressor. Hence, the `formula` argument is assigned the following model equation `examsc.class` $\sim$ `standLRT + sex + (standLRT|school)`. So far, the function `semLme()` enables the estimation of linear mixed models with a maximum of one random slope and one random intercept parameter. Regarding all other arguments, the same specifications as before are made.

```
R> LME <- semLme(formula = examsc.class ~ standLRT + sex +
+    (standLRT|school), data = data, classes = intervals,
+    bootstrap.se = TRUE)
```

By using the generic function `summary()` the estimation results are printed. In addition to the fixed effects, the estimated random effects are obtained as in the **lme4** and **nlme** packages. Since the R-squared and the adjusted R-squared are not defined for mixed models the `summary()` function prints the Marginal R-squared and Conditional R-squared (Nakagawa and Schielzeth, 2013; Johnson, 2014).

```
> summary(LME)

Call:
semLme(formula = examsc.class ~ standLRT + sex + (standLRT |
    school), data = data, classes = intervals,
    bootstrap.se = TRUE)

Random effects:
    Groups        Name   Variance  Std.Dev.
    school (Intercept) 0.08524761 0.2919719
  standLRT             0.01515524 0.1231066
  Residual             0.57213169 0.7563939

Fixed effects:
            Estimate Std.Error Lower 95%-level Upper 95%-level
(Intercept) 5.065732 0.0435255        4.973548        5.159554
standLRT    0.553797 0.0215305        0.504993        0.595787
sexM       -0.174975 0.0331477       -0.250686       -0.105352
```

```
Marginal R-squared: 0.319 Conditional R-squared: 0.4205
Variable examsc.class is divided into 9 intervals.
```

Again, interpretation is the same as in linear mixed models with a continuous dependent variable. By applying the generic function `plot()` to an `"sem" "lme"` object the same plots as for the linear regression model are plotted.

4.4 Discussion and outlook

Asking for interval-censored data can lead to lower item non-response rates and increased data quality. While item non-response is potentially avoided, applying traditional statistical methods becomes infeasible because the true distribution of the data within each interval is unknown. The functions of the **smicd** package enable researchers to easily analyze this kind of data. The paper briefly introduces the new statistical methodology and presents, in detail, the core functions of the package:

- `kdeAlgo()` for the direct estimation of any statistical indicator,
- `semLm()` to estimate linear models with an interval-censored dependent variable,
- `semLme()` to estimate linear mixed models with an interval-censored dependent variable.

The functions are applied in order to estimate statistical indicators from interval-censored EU-SILC income data and to analyze interval-censored examination scores of students from London with linear and linear mixed regression models.

Further developments of the **smicd** package will include the possibility to estimate the bootstrapped standard errors in parallel computing environments. Additionally, it is planned to allow for the use of survey weights in the linear (mixed) regression models.

Bibliography

Abraham, K. and S. Houseman (1995). Earnings inequality in Germany. In R. B. Freeman and L. F. Katz (Eds.), Differences and Changes in Wage Structures, pp. 371–404. Chicago: Nber Comparative Labor Markets.

Agresti, A. (2010). Analysis of Ordinal Categorical Data. New Jersey: Wiley.

Aldashev, A., J. Gernandt, and S. L. Thomsen (2008). The immigrant wage gap in Germany. Technical report, Centre for European Economic Research.

Alfons, A. and M. Templ (2013). Estimation of social exclusion indicators from complex surveys: the R package laeken. Journal of Statistical Software 54(15), 1–25.

Australian Bureau of Statistics (2011). Census household form. `https://unstats.un.org/unsd/demographic/sources/census/quest/AUS2011en.pdf`. Accessed: 2018-04-05.

Bandourian, R., J. McDonald, and R. Turley (2003). Income distributions: an inter-temporal comparison over countries. Estadistica 55(1), 135–152.

Bandourian, R., J. McDonald, and R. S. Turley (2002). A comparison of parametric models of income distribution across countries and over time. Technical report, Luxembourg Income Study.

Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences 160(901), 268–282.

Bates, D., M. Mächler, B. Bolker, and S. Walker (2015). Fitting linear mixed-effects models using lme4. Journal of Statistical Software 67(1), 1–48.

Bates, D., M. Maechler, and B. Bolker (2014). mlmRev: Examples from Multilevel Modelling Software Review. R package version 1.0-6.

Battese, G. E., R. M. Harter, and W. A. Fuller (1988). An error component model for prediction of county crop areas using survey and satellite data. Journal of the American Statistical Association 83(401), 28–36.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S. Philosophical Transactions 53, 370–418.

Bell, B., S. Nickell, and G. Quintini (2002). Wage equations, wage curves and all that. Labour Economics 9(3), 341–360.

Berger, Y. G. and E. L. Escobar (2016). Variance estimation of imputed estimators of change for repeated rotating surveys. International Statistical Review 85(3), 421–438.

Blum, U., H. S. Buscher, H. Gabrisch, J. Günther, G. Heimpold, C. Lang, U. Ludwig, M. T. W. Rosenfeld, and L. Schneider (2010). Ostdeutschlands Transformation seit 1990 im Spiegel wirtschaftlicher und sozialer Indikatoren. Technical report, Institut für Wirtschaftsforschung Halle.

Boehle, M. (2015). Armutsmessung mit dem Mikrozensus: Methodische Aspekte und Umsetzung für Querschnitts- und Trendanalysen. Technical report, Gesis Leibniz-Institut für Sozialwissenschaften.

Bönke, T., G. Corneo, and H. Lüthen (2014). Lifetime earnings inequality in Germany. Journal of Labor Economics 33(1), 171–208.

Bordley, R. F., J. B. McDonald, and A. Mantrala (1997). Something new, something old: parametric models for the size of distribution of income. Journal of Income Distribution 6(1), 91–103.

Borgoni, R., P. D. Bianco, N. Salvati, T. Schmid, and N. Tzavidis (2018). Modelling the distribution of health-related quality of life of advanced melanoma patients in a longitudinal multi-centre clinical trial using M-quantile random effects regression. Statistical Methods in Medical Research 27(2), 549–563.

Box, G. E. P. and D. R. Cox (1964). An analysis of transformations. Journal of the Royal Statistical Society: Series B 26(2), 211–252.

Bruch, C., R. Münnich, and S. Zins (2011). Variance estimation for complex surveys. Technical report, European Commission.

Burridge, J. (1981). A note on maximum likelihood estimation for regression models using grouped data. Journal of the Royal Statistical Society: Series B 43(1), 41–45.

Cameron, T. A. (1987). The impact of grouping coarseness in alternative grouped-data regression models. Journal of Econometrics 35(1), 37–57.

Carlin, B. and T. Louis (2000). Bayes and Empirical Bayes Methods for Data Analysis. London: Chapman & Hall.

Celeux, G., D. Chauveau, and J. Diebolt (1996). Stochastic versions of the EM algorithm: an experimental study in the mixture case. Journal of Statistical Computation and Simulation 55(4), 287–314.

Celeux, G. and J. Dieboldt (1985). The SEM algorithm: a probalistic teacher algorithm derived from the EM algorithm for the mixture problem. Computational Statistics Quarterly 2, 73–82.

Charlotte, L. and V. Steiner (1999). Returns to human capital in Germany: review of the empirical literature. In R. Asplund and P. T. Pereira (Eds.), Returns to Human Capital in Europe, pp. 125–146. Helsinki: ETLA.

Chen, Y. T. (2017). A unified approach to estimating and testing income distributions with grouped data. Journal of Business & Economic Statistics 36(3), 1–18.

Chotikapanich, D., W. E. Griffiths, and D. S. P. Rao (2007). Estimating and combining national income distributions using limited data. Journal of Business & Economic Statistics 25(1), 97–109.

Christensen, R. H. B. (2015). ordinal: Regression Models for Ordinal Data. R package version 2015.6-28.

Collins, D. and A. White (1996). In search of an income question for the 2001 census. Survey Methodology Bulletin 39(7), 2–10.

Corrado, A. (2007). Returns to education and wage equations: a dynamic approach. Applied Economics Letters 14(8), 577–579.

Dagum, C. (1977). A new model of personal income distribution: specification and estimation. Economie Appliquee 30, 413–437.

Dastrup, S. R., R. Hartshorn, and J. B. McDonald (2007). The impact of taxes and transfer payments on the distribution of income: a parametric comparison. Journal of Economic Inequality 5(3), 353–369.

Davison, A. C. and D. Hinkley (1997). Bootstrap Methods and Their Application. New York: Cambridge University Press.

Delaigle, A. (2007). Nonparametric density estimation from data with a mixture of Berkson and classical errors. Canadian Journal of Statistics 35(1), 89–104.

Delignette-Muller, M. L. and C. Dutang (2015). fitdistrplus: an R package for fitting distributions. Journal of Statistical Software 64(4), 1–34.

Demidenko, E. (2004). Mixed Models: Theory and Applications. New Jersey: Wiley.

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

Departamento Administrativo Nacional De Estadística (2005). Censo general 2005. `https://www.dane.gov.co/files/censos/libroCenso2005nacional.pdf?&`. Accessed: 2018-04-05.

Deville, J. (1999). Variance estimation for complex statistics and estimators: linearization and residual techniques. Survey Methodology 25(2), 193–203.

Diebolt, J. and E. H. S. Ip (1996). Stochastic EM: method and application. In W. Gilks, S. Richardson, and D. Speigelhalter (Eds.), Markov Chain Monte Carlo in Practice, pp. 259–268. London: Chapman & Hall.

Draper, N. R. and D. R. Cox (1969). On distributions and their transformation to normality. Journal of the Royal Statistical Society: Series B 31(3), 472–476.

Dutang, C., V. Goulet, and M. Pigeon (2008). actuar: an R package for actuarial science. Journal of Statistical Software 25(7), 38.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. The Annals of Statistics 7(1), 1–26.

Efron, B. (1982). The Jackknife, the Bootstrap and Other Resampling Plans. Philadelphia: Society for Industrial and Applied Mathematics.

Efron, B. and C. Stein (1981). The jackknife estimate of variance. The Annals of Statistics 9(3), 586–596.

Efron, B. and R. Tibshirani (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. Statistical Science 1(1), 54–75.

Efron, B. and R. Tibshirani (1993). An Introduction to the Bootstrap. New York: Chapman & Hall.

Elbers, C., J. O. Lanjouw, and P. Lanjouw (2003). Microlevel estimation of poverty and inequality. Econometrica 71(1), 355–364.

Ette, E. (1997). Stability and performance of a population pharmacokinetic model. Journal of Clinical Pharmacology 37(6), 486–95.

European Commission (2013). Description of target variables: cross-sectional and longitudinal. https://circabc.europa.eu/sd/a/d7e88330-3502-44fa-96ea-eab5579b4d1e/SILC065%20operation%202013%20VERSION%20MAY%202013.pdf. Accessed: 2018-04-09.

Eurostat (2014). Statistics explained: at-risk-of-poverty rate. http://ec.europa.eu/eurostat/statistics-explained/index.php/Glossary:At-risk-of-poverty_rate. Accessed: 2018-05-30.

Eurostat (2018). Statistics on income and living conditions (silc). http://ec.europa.eu/eurostat/de/web/microdata/statistics-on-income-and-living-conditions. Accessed: 2018-04-09.

Fahrmeir, L., R. Künstler, I. Pigeot, and G. Tutz (2011). Statistik - Der Weg zur Datenanalyse. Berlin: Springer.

Foster, J., J. Greer, and E. Thorbecke (1984). A class of decomposable poverty measures. Econometrica 52(3), 761–766.

Fryer, J. G. and R. J. Pethybridge (1972). Maximum likelihood estimation of a linear regression function with grouped data. Journal of the Royal Statistical Society: Series C 21(2), 142–154.

Fuchs-Schündeln, N., D. Krueger, and M. Sommer (2010). Inequality trends for Germany in the last two decades: a tale of two countries. Review of Economic Dynamics 13(1), 103–132.

Geraci, M. and M. Bottai (2014). Linear quantile mixed models. Statistics and Computing 24(3), 461–479.

Gini, C. (1912). Variabilità e mutabilità: contributo allo studio delle distribuzioni e delle relazioni statistiche. Studi economico-giuridici pubblicati per cura della facoltà di Giurisprudenza della R. Università di Cagliari. Bologna: Tipogr. di P. Cuppini.

Goldstein, H. (2003). Multilevel Statistical Models. New York: Wiley.

Goldstein, H., J. Rasbash, M. Yang, G. Woodhouse, H. Pan, D. Nuttall, and S. Thomas (1993). A multilevel analysis of school examination results. Oxford Review of Education 19(4), 425–433.

González-Manteiga, W., M. J. Lombardia, I. Molina, D. Morales, and L. Santamaria (2008). Analytic and bootstrap approximations of prediction errors under a multivariate Fay-Herriot model. Computational Statistics & Data Analysis 52(12), 5242–5252.

Graf, M. and D. Nedyalkova (2014). Modeling of income and indicators of poverty and social exclusion using the generalized beta distribution of the second kind. Review of Income and Wealth 60(4), 821–842.

Groß, M. and U. Rendtel (2016). Kernel density estimation for heaped data. Journal of Survey Statistics and Methodology 4(3), 339–361.

Groß, M., U. Rendtel, T. Schmid, S. Schmon, and N. Tzavidis (2017). Estimating the density of ethnic minorities and aged people in Berlin: multivariate kernel density estimation applied to sensitive georeferenced administrative data protected via measurement error. Journal of the Royal Statistical Society: Series A 180(1), 161–183.

Gurka, M. J., L. J. Edwards, K. E. Muller, and L. Kupper (2006). Extending the Box-Cox transformation to the linear mixed model. Journal of the Royal Statistical Society: Series A 169(2), 273–288.

Hagenaars, A. and K. D. Vos (1988). The definition and measurement of poverty. Journal of Human Resources 23(2), 211–221.

Hall, P. (1982). The influence of rounding errors on some nonparametric estimators of a density and its derivatives. SIAM Journal on Applied Mathematics 42(2), 390–399.

Hall, P. and M. P. Wand (1996). On the accuracy of binned kernel density estimators. Journal of Multivariate Analysis 56(2), 165–184.

Heckman, J. J., L. J. Lochner, and P. E. Todd (2003). Fifty years of Mincer earnings regressions. Technical report, National Bureau of Economic Research.

Henderson, D. J. and C. F. Parmeter (2015). Applied Nonparametric Econometrics. New York: Cambridge University Press.

Hsiao, C. (1983). Regression analysis with a categorized explanatory variable. In S. Karlin, T. Amemiya, and L. Goodman (Eds.), Studies in Econometrics, Time Series and Multivariate Statistics, pp. 93–129. Cambridge: Academic Press.

Hyndman, R. J. and A. B. Koehler (2006). Another look at measures of forecast accuracy. International Journal of Forecasting 22(4), 679–688.

Information und Technik (NRW) (2009). Berechnung von Armutsgefährdungsquoten auf Basis des Mikrozensus. `http://www.amtliche-sozialberichterstattung.de/pdf/Berechnung%20von%20Armutsgefaehrdungsquoten_090518.pdf`. Accessed: 2018-04-09.

International Monetary Fund (2017). World economic outlook database. `http://www.imf.org/external/pubs/ft/weo/2017/01/weodata/index.aspx`. Accessed: 2017-10-14.

Jenkins, S. P. (2009). Distributionally sensitive inequality indices and the GB2 income distribution. Review of Income and Wealth 55(2), 392–398.

Johnson, P. (2014). Extension of Nakagawa & Schielzeth's R^2_{GLMM} to random slopes models. Methods in Ecology and Evolution 5(9), 944–946.

Jones, M. C., J. S. Marron, and S. J. Sheather (1996). A brief survey of bandwidth selection for density estimation. Journal of the American Statistical Association 91(433), 401–407.

Kakwani, N. C. and N. Podder (2008). Efficient estimation of the Lorenz curve and associated inequality measures from grouped observations Lorenz curve and associated inequality measures from grouped observations. In D. Chotikapanich (Ed.), Modeling Income Distributions and Lorenz Curves, pp. 57–70. New York: Springer.

Kleiber, C. (2008). A guide to the Dagum distributions Lorenz curve and associated inequality measures from grouped observations. In D. Chotikapanich (Ed.), Modelig Income Distributions and Lorenz Curves, pp. 97–117. New York: Springer.

Kleiber, C. and S. Kotz (2003). Statistical Size Distributions in Economics and Actuarial Sciences. New York: Wiley.

Laird, M. N. and J. H. Ware (1983). Random-effects models for longitudinal data. Biometrics 38(4), 963–74.

Lemieux, T. (2006). The Mincer Equation thirty years after schooling, experience, and earnings. In S. Grossbard (Ed.), Jacob Mincer A Pioneer of Modern Labor Economics, pp. 127–145. New York: Springer.

Lenau, S. and R. Münnich (2016). Estimating income poverty and inequality from income classes. Technical report, InGRID Integrating Expertise in Inclusive Growth: Case Studies.

Levy, D., R. Hausmann, M. A. Santos, L. Espinoza, and M. Flores (2016). Why is Chiapas poor? Technical report, Center for International Development at Harvard University.

Lindstrom, M. J. and D. M. Bates (1990). Nonlinear mixed effects models for repeated measures data. Biometrics 46(3), 673–687.

Loader, C. R. (1999). Bandwidth selection: classical or plug-in? Annals of Statistics 27(2), 415–438.

Lok-Dessallien, R. (1999). Review of poverty concepts and indicators. Technical report, United Nations Development Programme.

Marhuenda, Y., I. Molina, D. Morales, and J. N. K. Rao (2017). Poverty mapping in small areas under a twofold nested error regression model. Journal of the Royal Statistical Society: Series A 180(4), 1111–1136.

McCullagh, P. (1980). Regression models for ordinal data. Journal of the Royal Statistical Society: Series B 42(2), 109–142.

McCulloch, C. E., S. R. Searle, and J. M. Neuhaus (2008). Generalized, Linear, and Mixed Models. New Jersey: Wiley.

McDonald, J. B. (1984). Some generalized functions for the size distribution of income. Econometrica 52(3), 647–663.

McDonald, J. B. and M. Ransom (2008). The generalized beta distribution as a model for the distribution of income: estimation of related measures of inequality. In D. Chotikapanich (Ed.), Modeling Income Distributions and Lorenz Curves, pp. 147–166. New York: Springer.

McDonald, J. B. and M. R. Ransom (1979). Functional forms, estimation techniques and the distribution of income. Econometrica 47(6), 1513–1525.

McDonald, J. B. and Y. J. Xu (1995). A generalization of the beta distribution with applications. Journal of Econometrics 66(1), 133–152.

McLachlan, G. and T. Krishnan (2008). The EM Algorithm and Extensions. New York: Wiley.

Meng, X.-L. and D. B. Rubin (1991). Using EM to obtain asymptotic variance-covariance matrices: the SEM algorithm. Journal of the American Statistical Association 86(416), 899–909.

Micklewright, J. and S. Schnepf (2010). How reliable are income data collected with a single question? Journal of the Royal Statistical Society: Series A 173(2), 409–429.

Milanovic, B. (2003). The Ricardian vice: why Sala-i-Martins calculations of world income inequality are wrong. Technical report, Development Research Group World Bank.

Mincer, J. (1958). Investment in human capital and personal income distribution. Journal of Political Economy 66(4), 281–302.

Mincer, J. (1974). Schooling, Experience, and Earnings. Columbia: University Press.

Minoiu, C. and S. Reddy (2008). Kernel density estimation based on grouped data: the case of poverty assessment. Technical report, International Monetary Fund.

Molina, I. and J. N. K. Rao (2010). Small area estimation of poverty indicators. The Canadian Journal of Statistics 38(3), 369–385.

Moore, J. C. and E. J. Welniak (2000). Income measurement error in surveys: a review. Journal of Official Statistics 16(4), 331.

Münnich, R. (2008). Varianzschätzung in komplexen Erhebungen. Austrian Journal of Statistics 37(3 & 4), 319–334.

Nakagawa, S. and H. Schielzeth (2013). A general and simple method for obtaining R^2 from generalized linear mixed-effects models. Methods in Ecology and Evolution 4(2), 133–142.

Nielsen, S. F. (2000). The stochastic EM algorithm: estimation and asymptotic results. Bernoulli 6(3), 457–489.

Osier, G. (2009). Variance estimation for complex indicators of poverty and inequality using linearization techniques. Survey Research Methods 3(3), 167–195.

Parzen, E. (1962). On estimation of a probability density function and mode. The Annals of Mathematical Statistics 33(3), 1065–1076.

Pereznieto, P. (2010). The case of Mexico's 1995 peso crisis and Argentina's 2002 convertibility crisis: including children in policy responses to previous economic crises. Technical report, UNICEF: Social and economic policy.

Pinheiro, J. and D. Bates (2000). Mixed-Effects Models in S and S-Plus. New York: Springer.

Pinheiro, J., D. Bates, S. Debroy, D. Sarkar, and R Core Team (2017). nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1-131.

Pretson, J. (2008). Rescaled bootstrap for stratified multistage sampling. Survey Methodology 35(2), 227–234.

R Core Team (2018). R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing.

Rao, J. and I. Molina (2015). Small Area Estimation. New York: Wiley.

Rao, J., C. Wu, and K. Yue (1992). Some recent work on resampling methods for complex surveys. Survey Methodology 18(2), 209–217.

Rao, J. N. K. and C. F. J. Wu (1988). Resampling inference with complex survey data. Journal of the American Statistical Association 83(401), 231–241.

Raudenbush, S. W. and A. S. Bryk (2002). Hierarchical Linear Models: Applications and Data Analysis Methods. Thousand Oaks: Sage.

Reed, W. J. and F. Wu (2008). New four- and five-parameter models for income distributions. In D. Chotikapanich (Ed.), Modeling Income Distributions and Lorenz Curves, pp. 211–224. New York: Springer.

Rojas-Perilla, N., S. Pannier, T. Schmid, and N. Tzavidis (2017). Data-driven transformations in small area estimation. Technical report, Freie Universität Berlin, School of Business & Economics.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. The Annals of Mathematical Statistics 27(3), 832–837.

Rosett, R. N. and F. D. Nelson (1975). Estimation of the two-limit probit regression model. Econometrica 43(1), 141–146.

Ruud, P. A. (1991). Extensions of estimation methods using the EM algorithm. Journal of Econometrics 49(3), 305–341.

Schwarz, N. (2001). The German Microcensus. Schmollers Jahrbuch 132(1), 1–26.

Scott, D. W. and S. J. Sheather (1985). Kernel density estimation with binned data. Communications in Statistics - Theory and Methods 14(6), 1353–1359.

Searle, S. R., G. Casella, and C. E. McCulloch (1992). Variance Components. New York: Wiley.

Shao, J. and D. Tu (1995). The Jackknife and Bootstrap. New York: Springer.

Singh, S. and G. Maddala (1976). A function for the size distribution of incomes. Econometrica 44(5), 963–970.

Snijders, T. and R. Bosker (2011). Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling. London: Sage.

Statistical Offices of the Federation and the Federal States (2016). Data supply: Microcensus. http://www.forschungsdatenzentrum.de/en/database/microcensus/index.asp. Accessed: 2018-06-11.

Statistics New Zealand (2013). New Zealand census of population and dwellings. https://unstats.un.org/unsd/demographic/sources/census/quest/NZL2013enIn.pdf. Accessed: 2018-05-13.

Statistisches Bundesamt (2014). Codebook microsensus 2014. http://www.forschungsdatenzentrum.de/en/database/microcensus/codebook_microcensus_2014.pdf. Accessed: 2018-04-09.

Statistisches Bundesamt (2016). Data supply: microcensus. http://www.forschungsdatenzentrum.de/en/database/microcensus/index.asp. Accessed: 2018-04-09.

Statistisches Bundesamt (2017). Datenhandbuch zum Mikrozensus Scientific-Use-File 2012. http://www.forschungsdatenzentrum.de/bestand/mikrozensus/suf/2012/fdz_mz_suf_2012_schluesselverzeichnis.pdf. Accessed: 2017-07-22.

Statistisches Bundesamt (2018a). Der Mikrozensus stellt sich vor. https://www.destatis.de/DE/ZahlenFakten/GesellschaftStaat/Bevoelkerung/Mikrozensus.html. Accessed: 2018-09-25.

Statistisches Bundesamt (2018b). Microcensus. https://www.destatis.de/EN/FactsFigures/SocietyState/Population/HouseholdsFamilies/Methods/Microcensus.html. Accessed: 2018-06-11.

Statistisches Bundesamt (2018c). Wirtschaftsrechnungen: Einkommens- und Verbrauchsstichsprobe Einkommensverteilung in Deutschland. https://www.destatis.de/DE/Publikationen/Thematisch/EinkommenKonsumLebensbedingungen/EinkommenVerbrauch/Einkommensverteilung2152606139004.pdf?__blob=publicationFile. Accessed: 2018-05-22.

Stauder, J. and W. Hüning (2004). Die Messung von Äquivalenzeinkommen und Armutsquoten auf der Basis des Mikrozensus. Technical report, Statistische Analysen und Studien NRW.

Stewart, M. (1983). On least square estimation when the dependent varaible is grouped. The Review of Economic Studies 50(4), 737–753.

Tepping, B. (1968). Variance estimation in complex surveys. Proceedings of the American Statistical Association Social Statistics Section, 11–18.

Thai, H., F. Mentre, N. Holford, C. Veyrat-Follet, and E. Comets (2013). A comparison of bootstrap approaches for estimating uncertainty of parameters in linear mixed-effects models. Pharmaceutical Statistics 12(3), 129–140.

Thompson, J. W. A. (1962). The problem of negative estimates of variance components. Annals of Mathematical Statistics 33(1), 273–289.

Thompson, M. L. and K. Nelson (2003). Linear regression with type I interval- and left-censored response data. Environmental and Ecological Statistics 10(2), 221–230.

Tille, Y. (2001). Theorie des sondages: Echantillonnage et estimation en populations finies. Paris: Dunod.

Tobin, J. (1958). Estimation of relationships for limited dependent variables. Econometrica 26(1), 24–36.

Toomet, O. (2015). intReg: Interval Regression. R package version 0.2-8.

Tzavidis, N., N. Salvati, T. Schmid, E. Flouri, and E. Midouhas (2016). Longitudinal analysis of the strengths and difficulties questionnaire scores of the Millennium Cohort Study children in England using M-quantile random-effects regression. Journal of the Royal Statistical Society: Series A 179(2), 427–452.

Tzavidis, N., L.-C. Zhang, A. Luna, T. Schmid, and N. Rojas-Perilla (2018). From start to finish: a framework for the production of small area official statistics. Journal of the Royal Statistical Society: Series A 181(4), 927–979.

Venables, W. N. and B. D. Ripley (2002). Modern Applied Statistics with S. New York: Springer.

Verbeke, G. and G. Molenberghs (2000). Linear Mixed Models for Longitudinal Data. New York: Springer.

Vijverberg, W. P. M. (1986). Consistent estimates of the wage equation when individuals choose among income-earning activities. Southern Economic Journal 52(4), 1028–1042.

Walter, P. (2018). smicd: Statistical Methods for Interval Censored Data. R package version 1.0.2.

Walter, P., M. Groß, T. Schmid, and N. Tzavidis (2017). Estimation of linear and non-linear indicators using interval censored income data. Technical report, Freie Universität Berlin, School of Business & Economics.

Walter, P. and K. Weimer (2018). Estimating poverty and inequality indicators using interval censored income data from the German microcensus. Technical report, Freie Universität Berlin, School of Business & Economics.

Wand, M. (2015). KernSmooth: Functions for Kernel Smoothing. R package version 2.23-15.

Wand, M. and M. Jones (1995). Kernel smoothing. London: Chapman & Hall.

Wang, B. and M. Wertelecki (2013). Density estimation for data with rounding errors. Computational Statistics & Data Analysis 65, 4–12.

Wang, J., J. R. Carpenter, and M. A. Kepler (2006). Using SAS to conduct nonparametric residual bootstrap multilevel modeling with a small number of groups. Computer Methods and Programs in Biomedicine 82(2), 130–143.

Wehrens, R., H. Putter, and L. Buydens (2000, 12). The bootstrap: a tutorial. Chemometrics and Intelligent Laboratory Systems 54(1), 35–52.

Wolter, K. (1985). Introduction to Variance Estimation. New York: Springer.

Woodruff, R. S. (1971). A simple method for approximating the variance of a complicated estimate. Journal of the American Statistical Association 66(334), 411–414.

World Bank (2010). Poverty & equity data portal. `http://povertydata.worldbank.org/poverty/country/MEX/`. Accessed: 2017-10-14.

World Economic Forum (2017). Global risks 2017. `http://reports.weforum.org/global-risks-2017/part-1-global-risks-2017/`. Accessed: 2017-09-28.

Wu, C. F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. The Annals of Statistics 14(4), 1261–1295.

Yang, Y., H. K. T. Ng, and N. Balakrishnan (2016). A stochastic expectation-maximization algorithm for the analysis of system lifetime data with known signature. Computational Statistics 31(2), 609–641.

Zambom, A. Z. and R. Dias (2012). A review of kernel density estimation with applications to econometrics. International Econometric Review 5(1), 20–42.

Summaries

Summaries in English

Abstract: Estimating Linear Mixed Regression Models with an Interval-Censored Dependent Variable using a Stochastic Expectation-Maximization Algorithm applied to German Microcensus Data

Linear mixed regression analysis is a well-established statistical method used in various research fields. In its standard form the dependent variable is measured on a continuous scale. Parameter estimates are commonly obtained by maximum likelihood or residual maximum likelihood theory. However, when working with income data, the dependent variable might be censored to specific intervals in order to increase data privacy protection and to lower item non-response. This is the case in many surveys and censuses, such as the German Microcensus. To enable parameter estimation in these situations a stochastic expectation-maximization algorithm (SEM) is proposed. In order to estimate the standard errors of the fixed effects, a parametric bootstrap that accounts for the additional uncertainty coming from the interval-censored dependent variable is introduced. Model-based simulation results show the validity of the new methodology. The SEM algorithm is applied to data from the German Microcensus. In the application, the relationship between an interval-censored income variable and relevant explanatory variables is modeled with a linear mixed regression model. A random intercept on the variable nationality controls for the within-cluster correlation.

Keywords: grouped data, banded data, income, parametric bootstrap, linear regression, multilevel regression, hierarchical linear regression

Abstract: Estimating Poverty and Inequality Indicators using Interval-Censored Income Data from the German Microcensus

Rising poverty and inequality increases the risk of social instability in countries all around the world. To measure poverty and inequality there exists a variety of statistical indicators. Estimating these indicators is trivial as long as the income variable is measured on a metric scale. However, estimation is not possible, using standard formulas, when the income variable is interval-censored (or grouped), as in the German Microcensus. This is the case for numerous censuses due to confidentiality constraints or in order to decrease item non-response. To

enable the estimation of statistical indicators in these scenarios, we propose an iterative kernel density algorithm that generates metric pseudo samples from the interval-censored income variable. Based on these pseudo samples, poverty and inequality indicators are estimated. The standard errors of the indicators are estimated with a non-parametric bootstrap. Simulation results demonstrate that poverty and inequality indicators from interval-censored data can be unbiasedly estimated by the proposed kernel density algorithm. Also, the standard errors are correctly estimated with the non-parametric bootstrap. The kernel density algorithm is applied in this work to estimate regional poverty and inequality indicators from German Microcensus data. The results show the regional distribution of poverty and inequality in Germany.

Keywords: direct estimation, grouped data, kernel density estimation, non-parametric bootstrap, income

Abstract: Small Area Estimation with Interval-Censored Income Data

Among a variety of small area estimation methods one popular approach for estimating linear and non-linear poverty and inequality indicators is the use of the empirical best predictor under the unit-level nested error (random effects) regression model. The empirical best predictor relies on fitting a nested error regression model with a continuous dependent variable. Fitting the nested error regression model and parameter estimation is more challenging when the response variable is interval censored. Interval censoring of sensitive variables for example, income is sometimes the preferred approach for collecting data due to data confidentiality concerns or concerns about response burden. The work in this paper proposes methodology that enables fitting a nested error regression model when the dependent variable is interval censored. Model parameters are then used for small area prediction of finite population parameters of interest. Model fitting in the case of an interval-censored response variable is based on the use of a stochastic expectation-maximization algorithm. Since the stochastic expectation-maximization algorithm relies on the Gaussian assumptions of the model error terms, adaptive transformations are incorporated for handling departures from normality. The estimation of the mean squared error of the small area parameters is facilitated by a parametric bootstrap that captures the additional uncertainty due to the interval-censoring mechanism and the possible use of adaptive transformations. The empirical properties of the proposed methodology are assessed by using model-based simulations. The relevance of the proposed methodology in policy work is illustrated by estimating deprivation indicators for municipalities in the Mexican state of Chiapas.

Keywords: empirical best predictor, nested error regression model, grouped data, stochastic expectation-maximization

Abstract: The R Package smicd: Statistical Methods for Interval-Censored Data

The package **smicd** supports two new statistical methods for the analysis of interval-censored data: 1) direct estimation/prediction of statistical indicators, and 2) linear (mixed) regression analysis. Direct estimation of statistical indicators, for instance poverty and inequality indicators, is facilitated by a non-parametric kernel density algorithm. The algorithm allows us to account for weights in the estimation of statistical indicators. The standard errors of the statistical indicators are estimated with a non-parametric bootstrap. Furthermore, the package offers statistical methods for the estimation of linear and linear mixed regression models with an interval-censored dependent variable, particularly random slope and random intercept models. Parameter estimates are obtained through a stochastic expectation-maximization algorithm. Standard errors are estimated using a non-parametric bootstrap in the linear regression model and by a parametric bootstrap in the linear mixed regression model. To handle departures from the model assumptions, fixed (logarithmic) and data-driven (Box-Cox) transformations are incorporated into the algorithm. The functionality of the package is illustrated with example data sets to estimate poverty indicators from interval-censored data in Germany and to linear model interval-censored examination scores of students from London schools.

Keywords: grouped data, kernel density estimation, regression models, income data, stochastic expectation-maximization algorithm, direct estimation

Kurzzusammenfassungen in Deutsch

Zusammenfassung: Parameterschätzung in linear gemischten Modellen mit gruppierter abhängiger Variable durch einen Stochastic Expectation-Maximization Algorithmus angewandt auf den deutschen Mikrozensus

Die linear gemischte Regressionsanalyse ist ein etabliertes statistisches Verfahren, welches in verschiedenen Forschungsbereichen verwendet wird. Standardmäßig ist die abhängige Variable metrisch skaliert. In diesem Fall werden die Regressionsparameter mit der Maximum-Likelihood-Methode oder der restringierten Maximum-Likelihood-Methode geschätzt. Bei der Analyse von Einkommensdaten kann es jedoch vorkommen, dass die Daten nicht stetig, sondern gruppiert erhoben wurden. Daten werden in dieser Form erhoben um z.B. höheren Datenschutz zu gewährleisten oder die Rate an fehlenden Werten zu verringern. Gruppierte Daten werden im Rahmen von vielen Umfragen und Zensus erhoben, z.B. vom deutschen Mikrozensus. Um die Schätzung der Regressionsparameter mit gruppierter abhängiger Variable zu ermöglichen, wird ein Stochastic Expectation-Maximization (SEM) Algorithmus vorgeschlagen. Die Schätzung der Standardfehler der fixen Regressionsparameter erfolgt mit einem parametrischen Bootstrapverfahren. Dieses berücksichtigt die zusätzliche Unsicherheit in der Parameterschätzung, die durch die gruppierte abhängige Variable entsteht. Modellbasierte Simulationsergebnisse bestätigen die Validität des vorgeschlagenen Verfahrens. Im Anschluss wird der SEM Algorithmus auf Daten des deutschen Mikrozensuses angewandt. In der Anwendung

wird der Zusammenhang zwischen gruppiertem Einkommen und relevanten erklärenden Variablen durch ein linear gemischtes Modell geschätzt. Ein zufälliger Achsenabschnitt kontrolliert für die Korrelation der Beobachtungen von Personen mit gleicher Nationalität.

Stichworte: Intervalklassierte Daten, Einkommen, parametrischer Bootstrap, lineare Regression, Mehrebenenanalyse, hierarchische lineare Regression

Zusammenfassung: Berechnung von Armuts- und Ungleichheitsindikatoren aus gruppierten Einkommensdaten des deutschen Mikrozensus

Ansteigende Armut und Ungleichheit erhöht das Risiko von sozialer Instabilität in Ländern überall auf der Erde. Um Armut und Ungleichheit zu messen, existieren eine Vielzahl von statistischen Indikatoren. Diese Indikatoren können einfach berechnet werden solange die erhobene Einkommensvariable metrisch skaliert ist. Jedoch ist ihre Berechnung mit Hilfe von Standardformeln nicht möglich, wenn Einkommen, wie im deutschen Mikrozensus, nur gruppiert erhoben wird. Auch andere Zensus erheben Einkommen nur gruppiert, um Vertraulichkeit der Antworten zu gewährleisten und Antwortausfälle möglichst zu vermeiden. Um in diesen Fällen die Berechnung von statistischen Indikatoren zu ermöglichen, wird ein iterativer Kerndichteschätzalgorithmus vorgestellt, der aus gruppierten Daten metrische generiert. Mittels der so gewonnenen metrischen Daten können die interessierenden Armuts- und Ungleichheitsindikatoren berechnet werden. Die Standardfehler der statistischen Indikatoren werden unter Verwendung eines nicht-parametrischen Bootstrap-Verfahrens berechnet. Simulationsergebnisse zeigen, dass der vorgeschlagene Algorithmus die unverzerrte Berechnung von Armuts- und Ungleichheitsindikatoren ermöglicht. Auch die mit dem Boostrap-Verfahren berechneten Standardfehler sind valide. Der Kerndichteschätzalgorithmus wird anschließend verwendet, um regionale Armuts- und Ungleichheitsindikatoren aus deutschen Mikrozensusdaten zu berechnen. Die Analyseergebnisse veranschaulichen die räumliche Verteilung von Armut und Ungleichheit in Deutschland.

Stichworte: Direkte Schätzung, intervalklassierte Daten, Kerndichteschätzer, nicht-parametrischer Bootstrap, Einkommen

Zusammenfassung: Small Area Estimation mit gruppierten Einkommensdaten

Ein populärer Ansatz zur Schätzung linearer und nichtlinearer Armuts- und Ungleichheitsindikatoren für kleinräumige Gebiete ist der Empirical Best Predictor. Dieser basiert auf einem linear gemischten Modell (zufälliger Achsenabschnitt) mit einer stetigen abhängigen Variable. Wenn die abhängige Variable jedoch nicht stetig, sondern gruppiert erhoben wird, ist die Schätzung der Modellparameter erschwert. Sensible Daten, z.B. Einkommensdaten, werden mitunter gruppiert abgefragt, um ihre Vertraulichkeit zu gewährleisten und Antwortausfälle zu minimieren. In dieser Arbeit wird ein Stochastic Expectation-Maximization Algorithmus vorgeschlagen, der es ermöglicht, die Parameter eines gemischten Modells mit zufälligem Ach-

senabschnitt zu schätzen, wenn die abhängige Variable gruppiert ist. Die Parameter werden anschließend verwendet, um lineare und nichtlineare Indikatoren für kleine Gebiete zu berechnen. Der Stochastic Expectation-Maximization Algorithmus setzt die Normalverteilung der Fehlerterme des Modells voraus. Um Abweichungen von der Normalverteilungsannahme entgegenzuwirken, werden adaptive Transformationen in den Algorithmus integriert. Die Schätzung des mittleren quadratischen Fehlers der berechneten Indikatoren erfolgt durch einen parametrischen Bootstrap. Dieser berücksichtigt die zusätzliche Unsicherheit, die durch die gruppierte abhängige Variable und durch die adaptive Transformation entsteht. Die empirischen Eigenschaften der vorgeschlagenen Methode werden mit modellbasierten Simulationen evaluiert. Ihre Relevanz für politische Entscheidungen wird durch die Schätzung von Armutsindikatoren für Gemeinden im mexikanischen Bundesstaat Chiapas illustriert.

Stichworte: Empirical Best Predictor, linear gemischtes Modell mit zufälligem Achsenabschnitt, intervallklassierte Daten, Stochastic Expectation-Maximization Algorithmus

Zusammenfassung: Das R Paket smicd: Statistische Methoden für gruppierte Daten

Das Paket smicd ermöglicht die Analyse von gruppierten Daten mit Hilfe von zwei neuen statistischen Methoden: 1) Direkte Schätzung/Vorhersage von statistischen Indikatoren und 2) linear (gemischte) Regressionsanalyse. Die direkte Schätzung von Indikatoren, z.B. Armuts- und Ungleichheitsindikatoren, erfolgt mittels eines Kerndichteschätzalgorithmus. Gewichtete Indikatoren können berechnet werden, indem Stichproben und/oder Haushaltsgewichte mit in den Algorithmus aufgenommen werden. Ein nicht-parametrischer Bootstrap erlaubt die Berechnung der Standardfehler. Das Paket ermöglicht außerdem die Parameterschätzung in linear und linear gemischten Regressionsmodellen mit gruppierter abhängiger Variable (Modelle mit zufälligem Achsenabschnitt und Steigungsparameter). Die Parameter werden von einem Stochastic Expectation-Maximization Algorithmus geschätzt. Die Schätzung der Standardfehler erfolgt mit einem nicht-parametrischen Bootstraps im linearen Modell und mit einem parametrischen Bootstrap im linear gemischten Modell. Um Verletzungen von Modellannahmen zu beheben, beinhaltet der Algorithmus fixe und datengetriebene Transformationen (logarithmische und Box-Cox Transformation). Die Funktionalität des Pakets wird anhand von zwei Datensätzen präsentiert. Es werden beispielhaft Armutsindikatoren in Deutschland berechnet und Prüfungsnoten von Londoner Schülern linear (gemischt) modelliert.

Stichworte: Intervalklassierte Daten, Kerndichteschätzer, Regressionsmodell, Einkommensdaten, Stochastic Expectation-Maximization Algorithmus, direkte Schätzung

Declaration of Authorship

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text.

Berlin, October, 2018

Paul Walter