The R Package **smicd**: **S**tatistical **M**ethods for **I**nterval **C**ensored **D**ata

Paul Walter*

*Institute of Statistics and Econometrics, Freie Universität Berlin, Germany

Abstract

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 to account for weights in the estimation of statistical indicators. The standard errors of the statistical indicators are estimated by 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. Standard errors are estimated by 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

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 (Moore and Welniak, 2000; Hagenaars and Vos, 1988). Among others, popular surveys and censuses that collect interval censored data are the German Microcensus (Statistisches Bundesamt, 2017), the Columbian census (Departamento Administrativo Nacional De Estadistica, 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 continuous 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 broadly be 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 modelling via splines (Berger and Escobar, 2016) and fitting a parametric distribution function to the censored data (Bandourian et al., 2002; Dagum, 1977; McDonald, 1984). 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 the use of 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 modelling 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 by the function `kdeAlgo()`.

Similarly to the direct estimation of statistical indicators from interval censored data, a variety of ad-hoc approaches and explicitly formulated mathematical methods for the estimation of linear regression models with an interval censored dependent variable exists. 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 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 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 (Gurka et al., 2006).

In order to analyse longitudinal or clustered data (e.g. students within schools) linear mixed regression models are applicable. These kind 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 as linear regression models, can be estimated on the interval midpoints of the censored dependent variable. Furthermore, conceptualizing the model as ordered logit or probit regression model is feasible (Agresti, 2010). These approaches inherit the same advantages and disadvantages as discussed before. Linear mixed regression on the midpoints can be applied by the **lme4** (Pinheiro et al., 2017) or **nlme** (Bates et al., 2015) package and ordered logit regression is implemented in the **ordinal** package (Christensen, 2015). Up 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 fulfil the model assumptions (Gurka et al., 2006).

The paper is structured into two main sections. Section 2 deals with the direct estimation of statistical indicators from interval censored data whereas Section 3 introduces linear and linear mixed regression models with an interval censored dependent variable. Both sections are split up 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 datasets are provided. In Section 4 the main results are summarized and an outlook is given.

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) dataset (European Commission, 2013).

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 Zambom and Dias (2012); Jones

et al. (1996); Loader (1999). When working with interval censored data, standard KDE cannot be applied since x is not observed on a continuous scale. Nevertheless, its unobserved true distribution is of continuous form. As 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, is leading 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 equidistant interval width. Walter and Weimer (2018) show that the algorithm can be adjusted to one-dimensional data with 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 $k = \{k_1, k_2, \ldots, k_n\}$ 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

For fitting 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, which is obtained by the KDE $\hat{f}_h(x)$. 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 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 equally 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 by model-based simulations that a value of 3 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 estimation procedure when the variance cannot be stated as closed form solution (Shao and Tu, 1995). While bootstrapping avoids the problem of 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 was first introduced in Walter and Weimer (2018).

2.2 Core functionality: direct estimation of statistical indicators

The presented KDE-algorithm is implemented in the function `kdeAlgo()` (see Table 1). The arguments and default settings of `kdeAlgo()` are shortly summarized in Table 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 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 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

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) dataset is used (European Commission, 2013). The PUF is a fully synthetic dataset which cannot be used for inferential statistics. Nevertheless, the distribution of the data mimics the distribution of the original dataset (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 12703 observations left in the EU-SILC survey that are used in the analysis. Since the total disposable household income is measured 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 830000 persons in 370000 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`.

Table 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 ∞ e.g. $(15000, \infty)$, then ∞ is replaced by $15000 * upper$	3
`weights`	Survey weights	NULL
`oecd`	Household weights of equivalence scale	NULL

```
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, 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 size 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 2916 Euro and its standard error is 23.124.

```
R> print(Indicators)

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

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

In Walter and Weimer (2018) the performance of the KDE-algorithm is evaluated by 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 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 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 4000 grid points G. Estimating a quantile based on only 4000 unique outcomes leads to equal quantile estimates for almost all iteration steps of the KDE-algorithm.

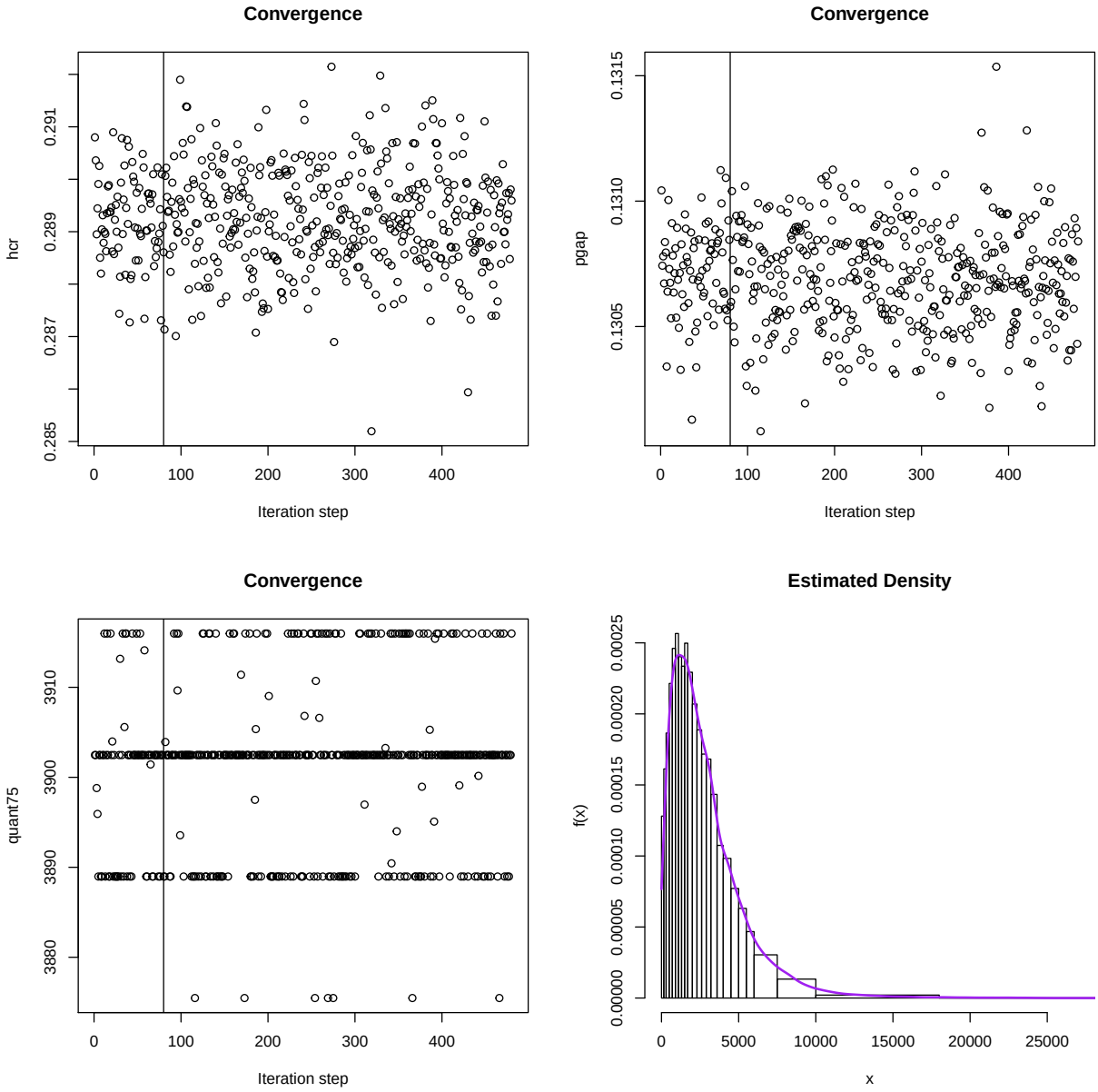

Figure 1: Convergence plots of the statistical indicators and a plot of the estimated final density with a histogram of the observed distribution of the data in the background

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 being exemplary modelled.

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

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v} + \mathbf{e}, \tag{1}$$

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 $\mathbf{e} \sim N(0, \mathbf{R})$ 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 Snijders and Bosker (2011); Searle et al. (1992); McCulloch et al. (2008). In the case of an interval censored dependent variable the parameters of Model 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 $\mathbf{y}$ by using the known intervals $\mathbf{k}$ and the linear relationship stated in Model 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 $\mathbf{y}$. Therefore,

$$f(\mathbf{k}|\mathbf{y}) = \begin{cases} 1 & \text{if } A_{\mathbf{k}-1} \leq \mathbf{y} \leq A_{\mathbf{k}}, \\ 0 & \text{else,} \end{cases}$$

and

$$f(\mathbf{y}|\mathbf{X}, \mathbf{Z}, \mathbf{v}, \boldsymbol{\theta}) \sim \ N(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v}, \mathbf{R}). \tag{2}$$

The relationship in Equation 2 follows from the linear mixed model assumptions (Model 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

For fitting Model 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

$$\hat{f}(\mathbf{y}|\mathbf{X},\mathbf{Z},\hat{\mathbf{v}},\mathbf{k},\hat{\boldsymbol{\theta}}) = \frac{\phi\left(\frac{\mathbf{y}-\hat{\boldsymbol{\mu}}}{\widehat{\mathbf{R}}}\right)}{\widehat{\mathbf{R}}\left(\Phi\left(\frac{A_{\mathbf{k}}-\hat{\boldsymbol{\mu}}}{\widehat{\mathbf{R}}}\right)-\Phi\left(\frac{A_{\mathbf{k}-1}-\hat{\boldsymbol{\mu}}}{\widehat{\mathbf{R}}}\right)\right)}, \tag{3}$$

with $\hat{\boldsymbol{\mu}} = \mathbf{X}\widehat{\boldsymbol{\beta}} + \mathbf{Z}\hat{\mathbf{v}}$. $\phi(\bullet)$ denotes the probability density function of the standard normal distribution and $\Phi(\bullet)$ denotes its cumulative distribution function. From its definition it follows that $\Phi\left(\frac{A_{\mathbf{k}}-\hat{\mu}}{\widehat{\mathbf{R}}}\right) = 1$ if $A_{\mathbf{k}} = \infty$ and $\Phi\left(\frac{A_{\mathbf{k}-1}-\hat{\mu}}{\widehat{\mathbf{R}}}\right) = 0$ 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 1 using the midpoints of the intervals as substitute 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 3) obtaining $(\tilde{\mathbf{y}}, \mathbf{X}, \mathbf{Z})$. The drawn pseudo $\tilde{\mathbf{y}}$ are used as replacement for the unobserved $\mathbf{y}$.
3. **Maximization Step:** Re-estimate the parameter vector $\hat{\boldsymbol{\theta}}$ from Model 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 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:

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

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 $f(\mathbf{y}|\mathbf{X},\boldsymbol{\beta},\sigma_e) \sim N(\mathbf{X}\boldsymbol{\beta},\sigma_e)$ according to the model assumptions of a linear regression model. In the SEM-algorithm for linear models it is then drawn from $N(\mathbf{X}\boldsymbol{\beta},\sigma_e)$ within the given interval.

The standard errors of the regression parameters are estimated by the use of bootstrap methods. For the linear regression model a non-parametric bootstrap (Efron and Stein, 1981; Efron, 1982; B. Efron and Tibshirani, 1986; Efron and Tibshirani, 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()`.

3.2 Core functionality: regression analysis

The introduced SEM-algorithm is implemented in the functions described in Table 3. The arguments and default settings of the estimation functions `semLm()` and `semLme()` are summarized in Table 4. Both functions return a S3 object of class `"sem" "lm"` or `"sem" "lme"`. A detailed explanation of all 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.

Table 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: 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

3.3 Example: regression analysis

To demonstrate the functions `semLm()` and `semLme()` the famous London school dataset that is analysed in Goldstein et al. (1993) is used. The dataset contains examination results of 4059 students from 65 schools in six Inner London Education Authorities. The dataset 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 Examination 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 dataset 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 eximination 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 dataset `Exam` and the vector of interval bounds `intervals` is assigned to the `classes` argument. The arguments `burnin` and `samples` are left as default. 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` has 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)  5.0696954 0.01769550       5.0291108       5.1062928
standLRT     0.5908558 0.01250971       0.5650455       0.6136739
sexM        -0.1713774 0.02697037      -0.2370421      -0.1144653

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 9 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.

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 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.

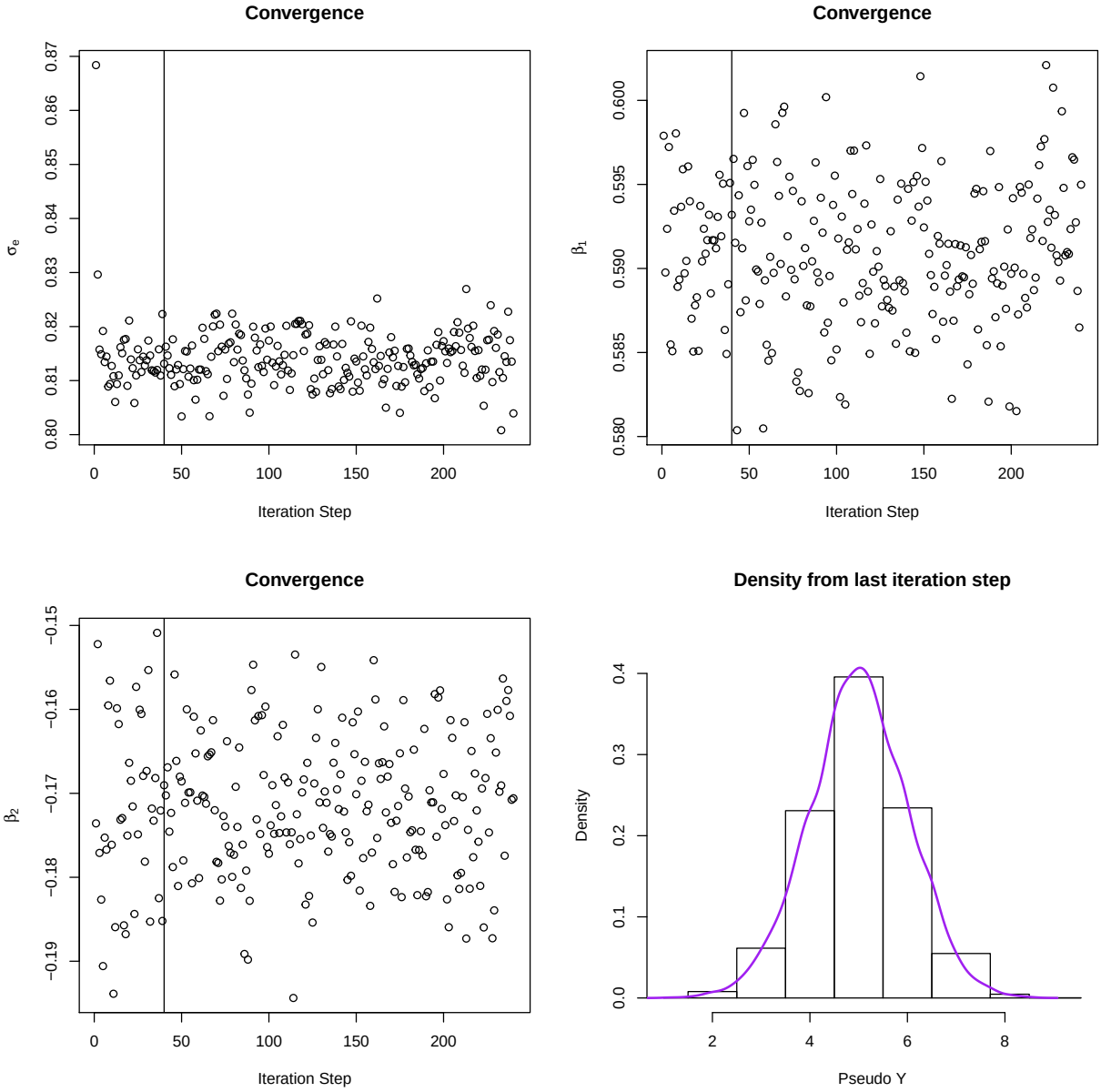

Figure 2: Convergence plots of estimated model parameters and the estimated final density with a histogram of the observed distribution of the data in the background

As already mentioned the **smicd** package also enables the estimation of linear mixed regression models by the function `semLme()`. In the London school dataset 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 random intercept. Furthermore, a random slope parameter on the standardized London reading test score `standLRT` is included into the model to allow for different slopes. Again the variable `sex` is included as 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 are made as before.

```
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. Additionally 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)  5.0657320   0.04352554       4.9735476       5.1595542
standLRT     0.5537966   0.02153048       0.5049930       0.5957868
sexM        -0.1749747   0.03314769      -0.2506864      -0.1053517

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 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 analyse this kind of data. The paper shortly 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 to estimate statistical indicators from interval censored EU-SILC income data and to analyse 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 bootstraped standard errors in parallel computing environments. Additionally, it is planned to allow for the use of survey weights in the linear (mixed) regression models.

References

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

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

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

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

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

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

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.

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

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

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

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

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

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

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

Dutang, C., Goulet, V., and Pigeon, M. (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*. Society for Industrial and Applied Mathematics, Philadelphia.

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

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

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., Kuenstler, R., Pigeot, I., and Tutz, G. (2011). *Statistik - Der Weg zur Datenanalyse*. Springer, Berlin.

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

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

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. Tipogr. di P. Cuppini, Bologna.

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

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

Groß, M., Rendtel, U., Schmid, T., Schmon, S., and Tzavidis, N. (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., Edwards, L. J., Muller, K. E., and Kupper, L. (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 Vos, K. D. (1988). The definition and measurement of poverty. *Journal of Human Resources*, 23(2):211–221.

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.

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

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

Lenau, S. and Münnich, R. (2016). Estimating income poverty and inequality from income classes. In Münnich, R., editor, *InGRID Integrating Expertise in Inclusive Growth: Case Studies*, pages 60–90.

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

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

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

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

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

Nakagawa, S. and Schielzeth, H. (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.

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

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

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

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

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

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

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.

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

Thai, H., Mentre, F., Holford, N., Veyrat-Follet, C., and Comets, E. (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 Nelson, K. (2003). Linear regression with type I interval- and left-censored response data. *Environmental and Ecological Statistics*, 10(2):221–230.

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.

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

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

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

Wang, J., Carpenter, J. R., and Kepler, M. A. (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.

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