Freie Universität Berlin

Estimation of Linear and Non-Linear Indicators using Interval Censored Income Data

Paul Walter
Marcus Groß
Timo Schmid
Nikos Tzavidis

School of Business & Economics
Discussion Paper

Economics

2017/22

Estimation of Linear and Non-Linear Indicators using Interval Censored Income Data

Paul Walter[*], Marcus Groß[*], Timo Schmid[*], and Nikos Tzavidis[**]

[*]Institute of Statistics and Econometrics, Freie Universität Berlin, Berlin, Germany
[**]Department of Social Statistics & Demography, University of Southampton

Abstract

Among a variety of small area estimation methods, one popular approach for the estimation of linear and non-linear indicators is the empirical best predictor. However, parameter estimation using standard maximum likelihood methods is not possible, when the dependent variable of the underlying nested error regression model, is censored to specific intervals. This is often the case for income variables. Therefore, this work proposes an estimation method, which enables the estimation of the regression parameters of the nested error regression model using interval censored data. The introduced method is based on the stochastic expectation maximization algorithm. Since the stochastic expectation maximization method relies on the Gaussian assumptions of the error terms, transformations are incorporated into the algorithm to handle departures from normality. The estimation of the mean squared error of the empirical best predictors is facilitated by a parametric bootstrap which captures the additional uncertainty coming from the interval censored dependent variable. The validity of the proposed method is validated by extensive model-based simulations.

Keywords: Small area estimation, empirical best predictor, nested error regression model, grouped data.

1 Introduction

Extreme poverty rates have been cut by more than half since 1990. While this is a remarkable achievement, it is still one of the 17 sustainable development goals defined by the United Nations to eradicate extreme poverty by 2030 (United Nations, 2017). To fight poverty, it is essential to have knowledge about its spatial distribution. In order to estimate disaggregated poverty indicators small area estimation (SAE) methods can be used. These methods enable the estimation of poverty (e.g. linear and non-linear) indicators at a geographical level where direct estimation is either not possible or very imprecise, due to a lack of sample size (Rao and I.Molina, 2015). Most SAE methods use linking models that borrow strength across areas and incorporate auxiliary informations from censuses or administratives (Pfeffermann, 2002).

Among a variety of SAE methods e.g. the World Bank method, developed by Elbers et al. (2003) or the M-Quantile approach introduced by Chambers and Tzavidis (2006), one popular approach for the estimation is the empirical best predictor (EBP) proposed by Molina and Rao (2010). The EBP method is based on a nested error linear regression model (Battese et al., 1988) and it enables the estimation of linear and non-linear indicators. As long as the dependent variable of the nested error linear regression model is measured on a metric scale parameter estimates are obtained using standard maximum likelihood methods.

However, standard estimation procedure can not be applied, when the dependent variable of the underlying nested error regression model, such as income or consumption, is censored to specific intervals, also known as grouping. While this is rarely the case in underdeveloped countries, in developed countries income is often only observed, due to confidentially constraints or other reasons, as grouped variable e.g. in the German micro census (Statistisches Bundesamt, 2017). Even though, absolute poverty, as defined by United Nations (1995) is not an issue in most developed countries, its politicians are still interested in the spatial distribution of income to target less wealthy areas more accurately.

Therefore, this work introduces an estimation method, which enables the estimation of the regression parameters of the nested error regression model when the dependent variable is interval censored. The proposed estimation method is based on the stochastic expectation maximization (SEM) algorithm (Caleux and Dieboldt, 1985). Since the proposed method relies on the Gaussian assumption of the error terms, transformations are incorporated into the algorithms to handle departures from normality. To the best of our knowledge there is no comparable approach proposed in the literature yet. Following Gonzalez-Manteiga et al. (2008) the estimation of the mean squared error (MSE) of the EBPs is facilitated by a parametric bootstrap. Their proposed method is progressed to capture the additional uncertainty coming from the interval censored dependent variable.

The paper is organized as follows. Since the EBP method relies on the parameter estimates of the nested error regression model, Section 2 starts with introducing the SEM-algorithm, which enables the estimation of the model parameters when the dependent variable is grouped. Then, Section 3 introduces the EBP method with grouped dependent data. The proposed MSE estimation procedure is introduced in Section 4. And in order to evaluate the performance of the SEM-algorithm and the parametric bootstrap for the MSE estimation, model-based simulation results are presented in Section 5. Finally, the main results are discussed and summarized in Section 6.

2 Nested error regression models with interval censored dependent variable

The later introduced EBP method relies on the parameter estimates of the nested error linear regression model, defined by Battese et al. (1988). It is given by:

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

where y_{ij} is the unobserved dependent variable, x_{ij} is a $p \times 1$ vector of explanatory variables, with p is the number of explanatory variables, β is a $p \times 1$ vector of regressors, $j = 1, \dots, n_i$ refers to the j-th individual and $i = 1, \dots, D$ to the i-th area. The error terms u_i and e_{ij} are assumed to be independent.

Standard estimation procedures, such as maximum likelihood (ML) or residual maximum likelihood (REML) are utilized for parameter estimation when y_{ij} is observed on a metric scale. However, when the dependent variable is grouped, standard estimation methods can not be applied.

For linear regression models there are different approaches to handle interval censored dependent variables. A very naive approach is ordinary least square regression on the midpoints of the intervals. This approach has two major drawbacks. First, the uncertainty regarding the nature of the exact value of each observation within each interval is not being reflected in the model and secondly, dealing adequately with open ended intervals is not possible. A different approach is, conceptualizing the model as ordered probit or logit regression (McCullagh, 1980). But since, the predicted values are then in terms of probability of membership in each interval, these models cannot be used for predicting the true unknown value of the dependent variable on a continuous scale.

To overcome these drawbacks interval regression as proposed by Stewart (1983) can be applied to interval censored data in the standard linear regression context. Stewart (1983) describes the possibility of using an iterative expectation-maximization (EM) algorithm as discussed by Dempster et al. (1977) to estimate the model parameters of a linear regression model. Following and extending this idea, a stochastic expectation-maximization algorithm (SEM) (Caleux and Dieboldt, 1985) is proposed that can easily be adapted to a variety of model classes (e.g. linear regression models or nested error linear regression models). Also the EM-algorithm was considered for parameter estimation, but since its performance in terms of accuracy in the prediction was worse under transformations, only the SEM-algorithm is introduced.

Consider model 1, where the only observed information concerning the dependent variable is, that it falls into a certain interval on a continuous scale. 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} $(1 \leq k_i \leq K)$ 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 $A_K = +\infty$ is possible. Of course, situations in which either or none of the outer intervals are open ended can also be handle by the SEM-algorithm. Furthermore, the interval length is allowed to be arbitrary and can vary between intervals.

Since the true distribution of y_{ij} is unknown, the aim is to estimate the distribution of y_{ij}, by using the known interval k_{ij} and the linear relationship given by model 1. To reconstruct the unknown distribution $f(y_{ij}|x_{ij}, k_{ij}, u_i)$ the Bayes theorem (Bayes, 1763) is applied:

$$f(y_{ij}|x_{ij}, k_{ij}, u_i) \propto f(k_{ij}|y_{ij}, x_{ij}, u_i) f(y_{ij}|x_{ij}, u_i),$$

with the conditional distribution of k_{ij},

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

and the conditional distribution of y_{ij},

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

The unknown model parameters $\theta = (\beta, u_i, \sigma_e^2, \sigma_u^2)$ and the conditional distribution $f(y_{ij}|x_{ij}, k_{ij}, u_i)$ are iteratively estimated using the SEM-algorithm described in the following section.

2.1 Parameter estimation and computational details

For fitting model1 pseudo samples of y_{ij} are generated iteratively from the following conditional distribution:

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

where $I(\cdot)$ denotes the indicator function. The steps of the SEM-algorithm are given by:

1. Estimate $\hat{\theta} = (\hat{\beta}, \hat{u}_i, \hat{\sigma}_e^2, \hat{\sigma}_u^2)$ from Equation 1 using the midpoints of the intervals as a substitute for the unknown y_{ij}.

2. Sample from the conditional distribution $f(y_{ij}|x_{ij}, k_{ij}, u_i)$ 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 y_{ij} are denoted by $\tilde{y}_{ij}$.

3. Re-estimate the vector $\hat{\theta}$ from equation 1 by using the pseudo sample $(\tilde{y}_{ij}, x_{ij})$ obtained in step 2.

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 obtained M estimates.

In the presence of open ended intervals $A_0 = -\infty$ and/or $A_K = +\infty$, the midpoints M_1 and M_K from the open ended intervals in iteration step 1 are computed as follows:

$$\begin{aligned} M_1 &= (A_1 - \overline{D})/2, \\ M_K &= (A_K + \overline{D})/2, \end{aligned} \tag{3}$$

where

$$\overline{D} = \frac{1}{K} \sum_{k=1}^{K} |A_{k-1} - A_k|. \tag{4}$$

During the algorithm it is repeatedly drawn 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 strongly on the Gaussian assumption of the error terms. To assure this assumption holds, the SEM-algorithm is upgraded by the incorporation of transformations.

2.2 The SEM-algorithm under transformations

Since the normality assumption is crucial for the performance of the SEM-algorithm, it is extended by the use of transformations. Transformations are applied to the target variable to meet distributional assumptions for multiple statistical methods. It is broadly distinguished between non-adaptive and adaptive transformation (Draper and Cox, 1969).

When the target variable is income, the logarithmic transformation, which is a non-adaptive transformation, is probably the most applied one in the field of econometrics. It is given by:

$$y^*_{ij} = log(y_{ij}), \tag{5}$$

where y^*_{ij} denotes the transformed variable. While the logarithmic transformation is easy to use, it has the downside of not adapting to the specific distributional shape of the data. This is particularly crucial, whenever the fulfilment of the distributional assumption cannot accurately be tested after the application of the transformation. This is the case, when the dependent variable of the nested error regression model is interval censored, since the true distribution of the residuals cannot be identified. Therefore, the SEM-algorithm is further extended by adaptive transformations.

Adaptive, or data-driven transformations automatically adapt to the specific shape of the data. Box and Cox (1964) introduced a family of adaptive power transformations. These transformations analytically adapt to the shape of the data. There are numerous alternative data driven transformation in the literature, see Manly (1976), In-Known and Johnson (2000) or Yang (2000) that can easily be incorporated into the SEM-algorithm. But, since the Box-Cox transformation shows fruitful results in the EBP context (Rojas-Perilla et al., 2017) it is focused on its implementation. The Box-Cox transformation (Box and Cox, 1964) is given by:

$$y^*_{ij}(\lambda) = \begin{cases} \frac{y^{\lambda}_{ij}-1}{\lambda} & \text{if } \lambda \neq 0, \\ ln\; y_{ij} & \text{if } \lambda = 0. \end{cases} \tag{6}$$

The formula holds, whenever $y_{ij} > 0$. In the case of negative y_{ij} values there is an extension available, including a parameter, that shifts the data into the positive region. The presented Box-Cox transformation depends on the transformation parameter λ. The aim is to optimize λ using residual maximum likelihood (REML) such that normally distributed residuals are obtained. Even if normality is not reached, at least symmetry is often obtained.

The SEM-algorithm under the logarithmic transformation

The implementation of a fixed transformation in the SEM-algorithm context is quite easy, since no parameter has to optimized. Therefore, embedding the log-transformation in the SEM-algorithm is easily done by transforming the intervals before iteration step 1 of the algorithm. Thus, the K intervals, where the k-th interval is given by (A_{k-1}, A_k) are simply transformed by taking the logarithm $(\log(A_{k-1}), \log(A_k))$. Again, the variable k_{ij} indicates in which of the intervals the dependent variable falls into. Only strictly positive intervals can be transformed by talking the logarithm, so $A_0 > 0$ has to hold or the logarithmic transformation needs to be extended by the inclusion of a shift parameter. Open ended upper intervals however, are still possible $A_k = +\infty$. After transforming the intervals, the SEM-algorithm is applied as described before.

The SEM-algorithm under the Box-Cox transformation

The application of any data-driven transformation in the SEM-algorithm is computational extensive and the algorithm has to be altered in order to optimize λ in each iteration step. The SEM-algorithm under the Box-Cox transformation is structured in two parts and given by:

Part 1

1. Perform the Box-Cox transformation on the interval midpoints, as a substitute for the unknown y_{ij}, estimate $\hat{\lambda}$ and obtain the transformed pseudo $\tilde{y}^*_{ij}$. Using $\hat{\lambda}$ to transform the intervals (A_{k-1}, A_k)

gives the transformed intervals (A^*_{k-1}, A^*_k).

2. Estimate $\hat{\theta} = (\hat{\beta}, \hat{u}_i, \hat{\sigma}^2_e, \hat{\sigma}^2_u)$ from equation 1 using the transformed $\tilde{y}^*_{ij}$.
3. Generate new pseudo samples as a proxy of the unobserved y^*_{ij}. Sample from the conditional distribution $f(y^*_{ij}|x_{ij}, k_{ij}, u_i)$ by drawing randomly from $N(x^T_{ij}\hat{\beta} + \hat{u}_i, \hat{\sigma}^2_e)$ within the given interval $(A^*_{k-1} \leq y^*_{ij} \leq, A^*_k)$, obtain $(\tilde{y}^*_{ij}, x_{ij})$.
4. Transform $\tilde{y}^*_{ij}$ back to $\tilde{y}_{ij}$ with $\hat{\lambda}$ from the previous iteration step. Apply the Box-Cox transformation on $\tilde{y}_{ij}$ to estimate a new $\hat{\lambda}$ and to get new transformed pseudo $\tilde{y}^*_{ij}$. Again, using $\hat{\lambda}$ to transform the original classes (A_{k-1}, A_k), gives (A^*_{k-1}, A^*_k).
5. Re-estimate the vector $\hat{\theta}$ from equation 1 by using the pseudo sample $(\tilde{y}^*_{ij}, x_{ij})$ obtained in step 2.
6. Iterate steps 2-5 $B + M$ 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 obtained M estimates.

Part 2

8. Transform the original classes (A_{k-1}, A_k) using $\hat{\lambda}^{(F)}$ to get (A^*_{k-1}, A^*_k).
9. Restart the original SEM-algorithm proposed in the previous section with the transformed classes (A^*_{k-1}, A^*_k) and iterate B+M times.
10. Discard the B burn in iterations and estimate $\hat{\theta}$ by averaging the obtained M estimates.

Because of the non-linear relationship between the parameter estimates and $\hat{\lambda}$, as seen in Figure 1, obtaining $\hat{\theta}$ and $\hat{\lambda}$ by simply averaging the obtained M estimates would lead to non matching results and to erroneous EBPs.

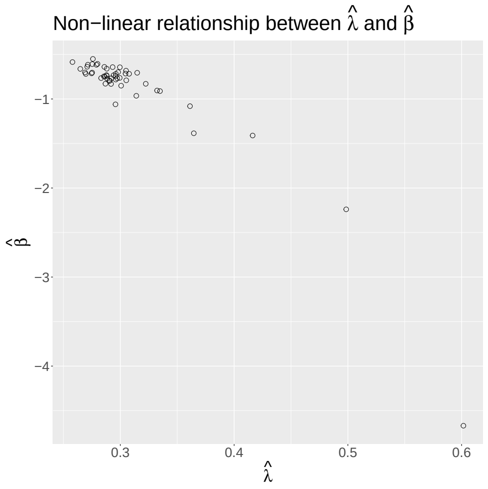

Figure 1: The estimated λ is plotted against the estimated β for each iteration of the SEM-algorithm.

Therefore the estimation is structured into two parts. In Part 1 the correct $\hat{\lambda}^{(F)}$ is identified. And afterwards, $\hat{\lambda}^{(F)}$ is used to transform the original classes and the SEM-algorithm is restarted (Part 2) with the transformed classes. The final estimates $\hat{\theta}$ are obtained by averaging over the M estimates, and since $\hat{\lambda}^{(F)}$ is fixed, $\hat{\beta}$ matches $\hat{\lambda}^{(F)}$ and the EBP method can be applied.

3 The EBP method

This section introduces the EBP method that is proposed by Molina and Rao (2010). The EBP method uses Monte carlo approximations for the estimation and it is based on the previous defined nested error linear regression model. By its application, EBPs for non-linear estimates are obtained at the domain or area level. The obtained estimates are best in terms of minimizing the MSE under the assumed SAE model. Simulations by Molina and Rao (2010) show that these estimates outperform direct estimates in terms of MSE. While this method is applicable for the estimation of non-linear indicators in general, this paper is focusing on poverty and inequality indicators. Since the performance of the EBP method strongly relies on the Gaussian assumption of the error terms, Rojas-Perilla et al. (2017) included data-driven transformations into the EBP method.

As before, consider a finite population U of size N, divided into D regions. The sample size of each of the D-regions $U_1, U_2, \ldots, U_D$ is given by $N_1, N_2, \ldots, N_D$. Further, the target variable with whom the poverty indicator of interest is estimated is denote by y_{ij}, where j indicates the jth individual belonging to the ith region. $j = 1, 2, \ldots, N_i$. The data matrix X is defined as $X = (x_1, \ldots, x_p)^T$, where p denotes the number of explanatory variables. The EBP approach further differentiates between sampled units s and non-samples units r. The sampled-units in area i are defined as s_i, whereby the non-sampled units are denoted by r_i. For each area i the sample size n_i is given by $n \sum_{i=1}^{D} n_i$ and the population vector y_i for area i consists of the sampled and non-sampled units $y_i^T = (Y_{is}^T, Y_{ir}^T)$. As mentioned before a nested error linear regression model serves to model the relationship between the variable of interest and auxiliary information, the unexplained between area variation is captures by u_i. Because y_{ij} is unobserved, since only the group information known, the EBP method (Molina and Rao, 2010) is extended by the use of the SEM-algorithm. The extended EBP method without transformations is given by:

1. Estimate $\hat{\theta} = (\hat{\beta}, \hat{\sigma}_u^2, \hat{\sigma}_e^2)$ by the SEM-algorithm and obtain the weighting factor, $\hat{\gamma}_i = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \frac{\hat{\sigma}_e^2}{n_i}}$ using the sample data.
2. For $l = 1, \ldots, L$:
   (a) Generate a bootstrap population using the nested error regression model $y_{ij}^{(l)} = x_{ij}^T \hat{\beta} + \hat{u}_i + v_i^{(l)} + e_{ij}^{(l)}$, where x_{ij} are auxiliary information from the population and v_i is drawn from $v_i \overset{iid}{\sim} N(0, \hat{\sigma}_u^2(1 - \hat{\gamma}_i))$, and e_{ij} is drawn from $e_{ij} \overset{iid}{\sim} N(0, \hat{\sigma}_e^2)$. The random effect u_i is given by $\hat{u}_i = E(u_i | y_i)$, the conditional expectation of u_i given y_i
   (b) In each area, estimate the poverty measure of interest $\hat{I}_i^{(l)}$ using $y_{ij}^{(l)}$.
3. Finally, estimate the poverty indicator of interest, by averaging over the L Monte Carlo estimates $\hat{I}_i^{(l)}$ in each area i:

$$\hat{I}_i^{EBP} = \frac{1}{L} \sum_{l=1}^{L} \hat{I}_i^{(l)}. \tag{7}$$

Whenever the SEM-algorithm under any transformation is used, the EBP method is slightly modified as described by Rojas-Perilla et al. (2017). In step 1 of the algorithm, $\hat{\theta}$ is estimated using the SEM algorithm under transformation. In step 2 (a), a transformed pseudo populations is obtained $y_{ij}^{*(l)}$, which has to be transformed back to the original scale $y_{ij}^{(l)}$. The poverty indicator of interest $\hat{I}_i^{(l)}$ is then estimated in each area.

In the presence of non-sampled areas, the bootstrap in 2(a) is altered as follows: For $l = 1, \ldots, L$ bootstrap from $y_{ij}^{(l)} = x_{ij}^T \hat{\beta} + u_i^{(l)} + e_{ij}^{(l)}$, where the error terms are drawn from $u_i^{(l)} \overset{iid}{\sim} N(0, \hat{\sigma}_u^2)$ and $e_{ij}^{(l)} \overset{iid}{\sim} N(0, \hat{\sigma}_e^2)$.

4 MSE estimation

To asses the quality of the EBPs the MSE is estimated for each indicator $\hat{I}_i^{EBP}$. The MSE estimation is facilitated by a parametric bootstrap, that was first introduced by Gonzalez-Manteiga et al. (2008). The method is further extended by Rojas-Perilla et al. (2017) to account for the additional variability coming from the estimation of the transformation parameter λ. For capturing the additional uncertainty due to the interval censored dependent variable, the algorithm is advanced further as follows:

1. Use the sample estimates $\hat{\theta} = (\hat{\beta}, \hat{\sigma}_u^2, \hat{\sigma}_e^2)$ obtained by the SEM-algorithm, generate $u_i \overset{iid}{\sim} N(0, \sigma_u^2)$ and $e_{ij} \overset{iid}{\sim} N(0, \sigma_e^2)$ and simulate a bootstrap superpopulation model $y_{ij}^{(b)} = x_{ij}^T \hat{\beta} + u_i^{(b)} + e_{ij}^{(b)}$.
2. Estimate the population indicator $I_{i,b}$ using $y_{ij}^{(b)}$.
3. Extract the bootstrap sample $y_{ij}^{(b)}$, group $y_{ij}^{(b)}$ according to the K intervals (A_{k-1}, A_k) and then apply the EBP method using only the interval informations and treating $y_{ij}^{(b)}$ as unknown.
4. Obtain $\hat{I}_{i,b}^{EBP}$.
5. Iterate steps 2-4, $b = 1, \dots, B$ times. The MSE-estimate for each area i given by:

$$\widehat{MSE}(\hat{I}_i^{EBP}) = B^{-1} \sum_{b=1}^{B} (\hat{I}_{i,b}^{EBP} - I_{i,b})^2. \tag{8}$$

Again, if transformations are incorporated into the algorithm the procedure needs some adjustments. In step 1. the sample estimates and additionally $\hat{\lambda}^{(F)}$ are obtained by the SEM-algorithm under transformation and the bootstrap superpopulation model yields transformed $y_{ij}^{*(b)}$. The $y_{ij}^{*(b)}$ are then transformed back to the original scale in step 2, and the population indicator $I_{i,b}$ is estimated using $y_{ij}^{(b)}$. In the 3.step, the EBP method is applied using the SEM-algorithm under transformation and the data is transformed back again. The estimation of $\hat{\lambda}^{(F)}$ is newly done for each bootstrap sample b.

5 Model-based simulations

This section presents model-based simulation results for the proposed methods to evaluate the performance of the SEM-algorithm in the EBP context, with and without transformations. Also, the results of the newly presented MSE method are evaluated. For the evaluation of the EBPs it is concentrated on the following popular poverty and inequality measures. The Gini coefficient (Gini) (Gini, 1912), the head count ratio (HCR) (Foster et al., 1984) with a threshold equal to 60% of the median of the target variable and the mean.

Even though, there are various quality measures in the literature, it is focused on the root mean squared error (RMSE) of each indicator $\hat{I}^{EBP}$ in each area i. The RMSE is a scale dependent measure that can be used to evaluate the performance of different methods applied to the same data (Hyndman and Koehler, 2006). It is given by:

$$RMSE(\hat{I}_i^{EBP}) = \left[\frac{1}{M} \sum_{m=1}^{M} \left(\hat{I}_i^{EBP(m)} - I_i^{(m)} \right)^2 \right]^{1/2}, \tag{9}$$

where M corresponds to the number of Monte Carlo populations. It measures the difference between the estimated poverty or inequality measure and its corresponding true value. The domain of the RMSE lies between 0 and ∞.

Three different super-population models (Table 1) are used for evaluation. The normal scenario is used to evaluate the performance of the EBP approach when all model assumptions are met. In contrast, the Log-scale and the GB2 scenario try to mimic an equalized income distribution of the dependent variable (Graf et al., 2011). Thus, the Gaussian assumption of the error terms is not fulfilled. In all scenarios a finite population U of size $N = 10000$, which is partitioned into $D = 50$ regions $U_1, U_2, \dots, U_D$ of sizes $N_i = 200$ is generated. Afterwards, a sample using an unbalanced design with sample sizes n_i between $8 \leq n_i \leq 29$ leading to a total sample size of $\sum_{i=1}^{D} n_i = 921$ is drawn from the population. The number of Monte Carlo iterations equals 100. Furthermore, $L = 50$ and $B = 100$.

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	$exp(10 - x_{ij} - 0.5z_{ij}u_i + e_{ij})$	$N(\mu_i, 2)$	$N(0, 1)$	$U(-3, 3)$	$N(0, 0.4^2)$	$N(0, 0.8^2)$
GB2	$8000 - 400x_{ij} + u_i + e_{ij} - \overline{e}$	$N(\mu_i, 5)$	-	$U(-1, 1)$	$N(0, 500^2)$	$GB2(2.5, 1700, 18, 1.46)$

Table 1: Model based simulation scenarios for the evaluation of the EBPs and the MSE

There are different methods applied for the estimation of the parameters of the nested error regression model. This is done in order to compare the performance of the EBP approach using y_{ij} on a metric scale (abbreviated by LME) to the performance of the EBP approach whenever y_{ij} is grouped and the SEM-algorithm is used for parameter estimation (abbreviated by SEM). Furthermore, LME and SEM are also applied under the log transformation (LME Log and SEM Log) and under the Box-Cox transformation (LME Box-Cox and SEM Box-Cox). The SEM-algorithm is applied with 40 burn-in and 200 additional iterations. The rate of convergence depends strongly on the variability of the regressors. While the number of iterations is sufficient in the proposed set-ups larger variability in the regressors will lead to longer convergence time. Therefore, the practitioner should check the convergence of the parameters and find a sufficiently high number of iterations.

5.1 Evaluation under normality

For the evaluation of the EBPs under normality, two different interval censoring scenarios (Table 2) are simulated. This is done in order to study the influence of the number of intervals on the performance of the EBPs.

Normal scenario 1

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

Normal scenario 2

Interval	Frequencies
$[1, 3000)$	2337
$[3000, 5000)$	4329
$[5000, 7500)$	3032
$[7500, Inf)$	302

Table 2: Normal scenarios, distribution for one arbitrary chosen population.

Figure 2 and 3 present the results for the SEM-algorithm, the SEM Box-Cox algorithm, the LME and the LME Box-Cox approach. The results indicate, that the performance of the EBPs using the SEM-algorithm is close to the performance of the EBPs using LME. The exact numbers are given in the appendix in Table 3 and 4. If the number of intervals decreases, the performance of the EBPs gets worse. This is not surprising, since fewer information is used (4 compared to 7 intervals). Since four intervals is a really extreme case, e.g. in comparison the German micro census uses 24 intervals (Statistisches Bundesamt, 2017), the SEM-algorithm can be used without concerns in most practical applications.

The performance of the EBPs using SEM and SEM Box-Cox are very similar. This should be the case whenever the error terms are normally distributed, because no transformation is needed and the adaptivity of the Box-Cox transformation assures that no transformation is applied. Hence, the Box-Cox transformation adapts well to the shape of the data, even though only the interval information is used for the estimation of λ.

The MSE results for the different indicators are given in Figure 4 and in the appendix in Figure 9. The estimated RMSE tracks the empirical RMSE well. Therefore, the proposed bootstrap sufficiently

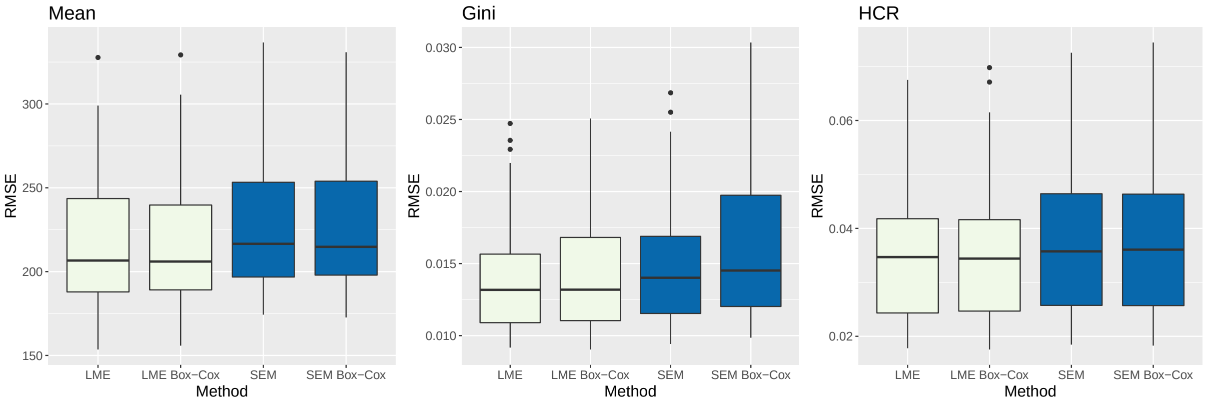

Figure 2: Normal scenario 1, RMSE for the proposed methods.

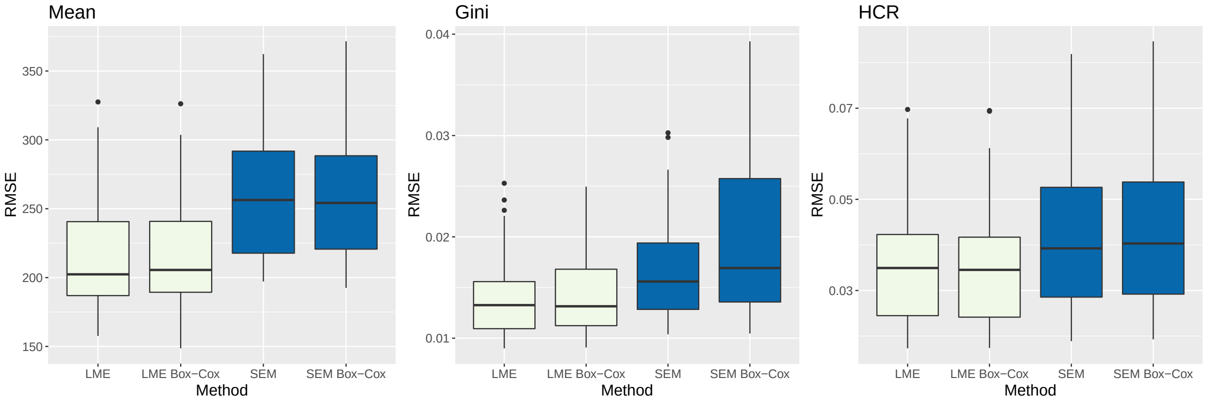

Figure 3: Normal scenario 2, RMSE for the proposed methods.

accounts for the additional variability that is due to the grouping.

In the appendix in Figure 10 and 13 the true density of the population y_{ij} is plotted against the prediction from one arbitrary chosen simulation run, using the different methods for parameter estimation. The density plots emphasize the prior results. The prediction of the SEM-algorithm is close to the prediction using LME and the performance of the SEM-algorithm depends on the number of intervals.

5.2 Evaluation under departures from normality

In order to evaluate the performance of the EBP approach when there are departures from normality the GB2 and the Log-scale scenario given in Table 1 are used.

Log-scale scenario

In the Log-scale scenario the dependent variable is grouped into seven intervals. The scenario with four intervals is not being considered, because the prior shown effect, that the performance of the EBPs decreases is independent from the chosen scenario. The distribution for one arbitrary chosen population is given in Table 5 in the appendix. Again, the results show (see Figure 5 and Table 6), that the performance of the EBPs using the SEM Box-Cox or SEM Log approach is close to the performance using LME Box-Cox or LME Log. Parameter estimation without transformation is not considered in the Log-scale scenario, because the results are clearly worse. Of course, some accuracy is lost, due to the grouping, but considering that only seven intervals are used for estimation the results are really promising. Also the Box-Cox transformation is working well, since its results are closed to the results of the Log transformation, indicating the capability of the Box-Cox transformation to adapt to the specific shape of the data. The Log transformation is the gold standard in the Log-scale scenario, due to the set up of the data and therefore its results can not be met by the Box-Cox transformation. The results are again backed up by the density Plot 15 given in the appendix.

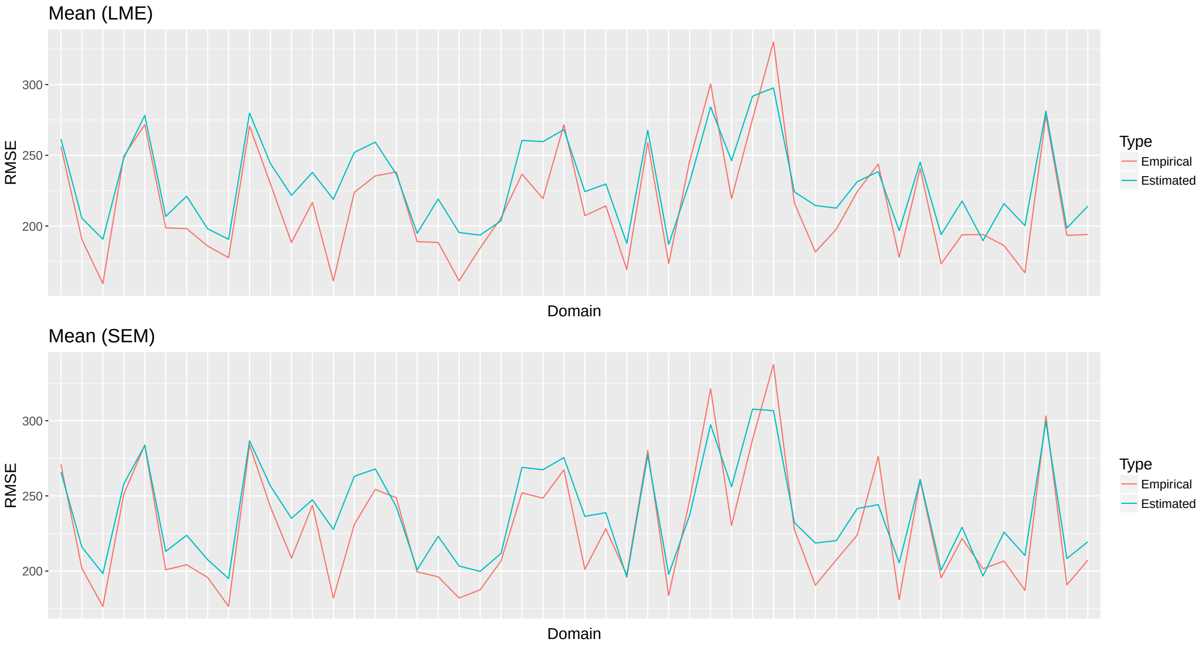

Figure 4: Normal scenario 1, MSE for the mean.

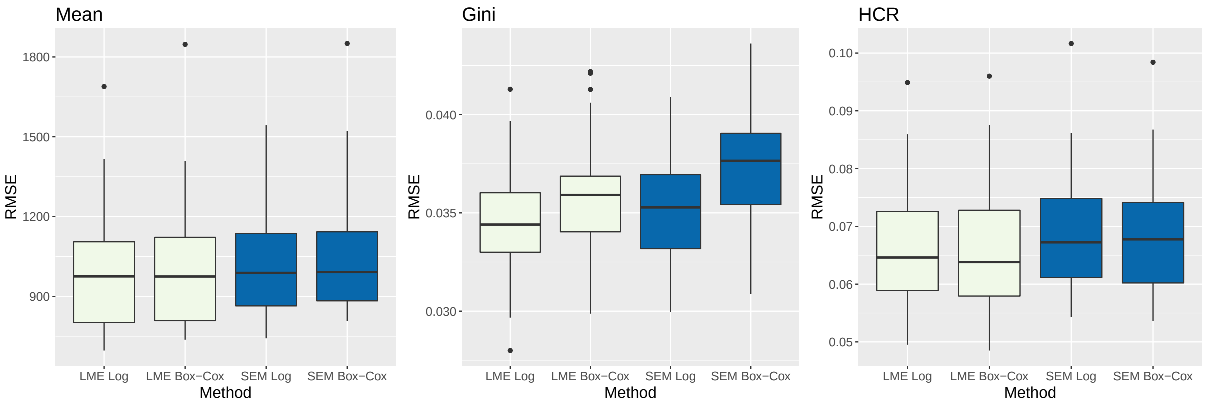

Figure 5: Log scenario, RMSE for the proposed methods.

The proposed MSE is also working under the logarithmic transformation (see Figure 6 and Figure 14 in the appendix) . The estimated MSE tracks the empirical MSE well.

GB2 scenario

The dependent variable in the GB2 scenario is again censored to seven intervals (see Table 7 in the appendix). The results show (see Figure 7 and Table 8) that the RMSE of the Gini and HCR using the SEM algorithm is smaller compared to the RMSE using LME. While this seems counter-intuitive, one reason might be that the SEM-algorithm is robust against outliers and thus performs better in the case of skewed data. Furthermore, the results highlight the functioning of the Box-Cox transformation even when the data is grouped. Again the results are backed up by a density plot given in the appendix (see Figure 17). It is seen that the SEM Box-Cox method reconstructs the true distribution best, even though it is only using the interval informations.

The Figures 8 and 16 in the appendix show, that the estimated RMSE is tracking the empirical RMSE well. Therefore the proposed bootstrap provides useful results even under the Box-Cox transformation.

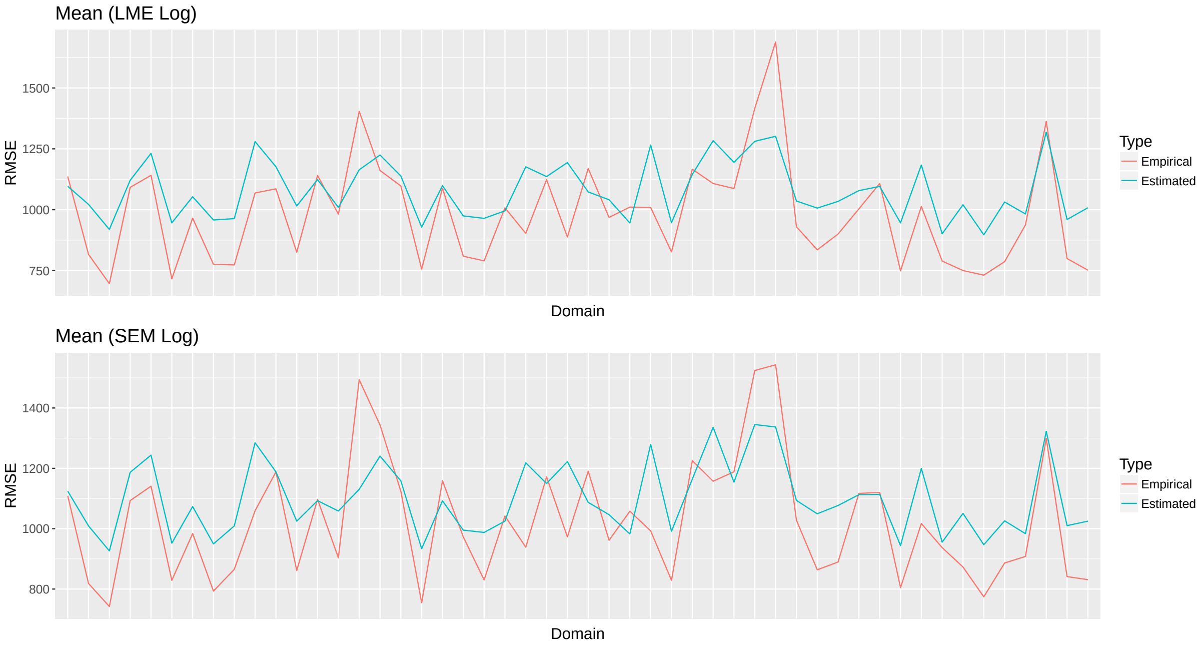

Figure 6: Log scenario, MSE for the mean.

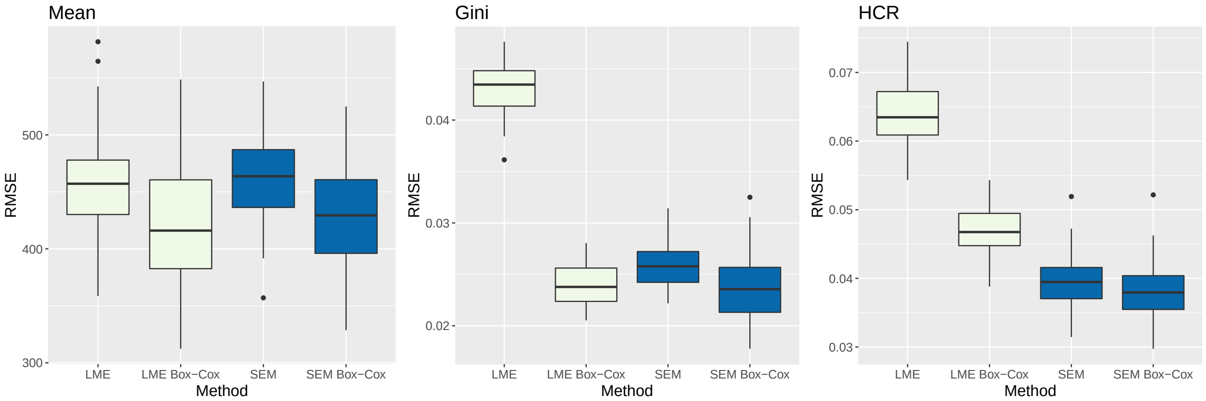

Figure 7: GB2 scenario, RMSE for the proposed methods.

6 Conclusion

The paper introduces a SEM-algorithm that estimates the parameter of a nested error linear regression model when the dependent variable is grouped. This enables the use of the popular EBP method, which is based on these models, when the target variable is grouped. The proposed SEM-algorithm relies on normally distributed error terms. Since this is rarely the case in applied work, transformations are incorporated into the algorithm to handle departures from normality. Furthermore, an MSE estimation procedure that accounts for the additional variability, coming from the interval censored dependent variable, is proposed.

To validate the proposed estimation methods extensive model based simulations were performed using different scenarios. The simulation results validate the functioning of the proposed SEM-algorithm and show that in most scenarios the loss of accuracy in the EBPs, is minimal compared to the use of the uncensored data. That accuracy is lost is obvious, because only the group information, hence less information is used by the SEM-algorithm. The amount of accuracy lost in the EBPs using the SEM-algorithm compared to the use of the uncensored data, strongly depends on the number of intervals the data is censored to. In the GB2 scenario the EBPs using the SEM-algorithm even outperform the EBPs using the true uncensored data. A possible explanation is that the SEM-algorithm is robust against outliers. Finally, empirical evaluation also validate the functioning of the proposed parametric bootstrap, also under transformations.

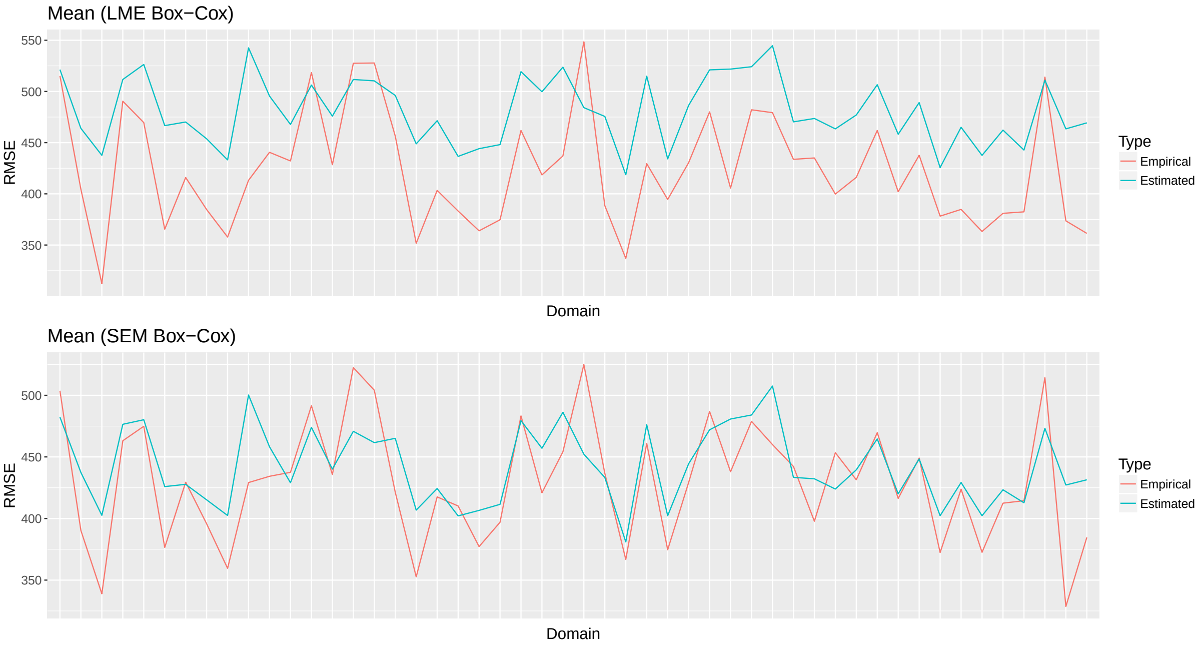

Figure 8: GB2 scenario, MSE for the mean.

Further research will focus on the robustness properties of the proposed methods against outliers and on inferential statistics.

7 Appendix

	LME	LMEBox	SEM	SEMBox
Mean	206.6279	206.0218	216.5873	214.7854
Gini	0.0132	0.0132	0.0140	0.0145
HCR	0.0347	0.0344	0.0357	0.0361

Table 3: Normal scenario 1, median RMSE for the proposed methods.

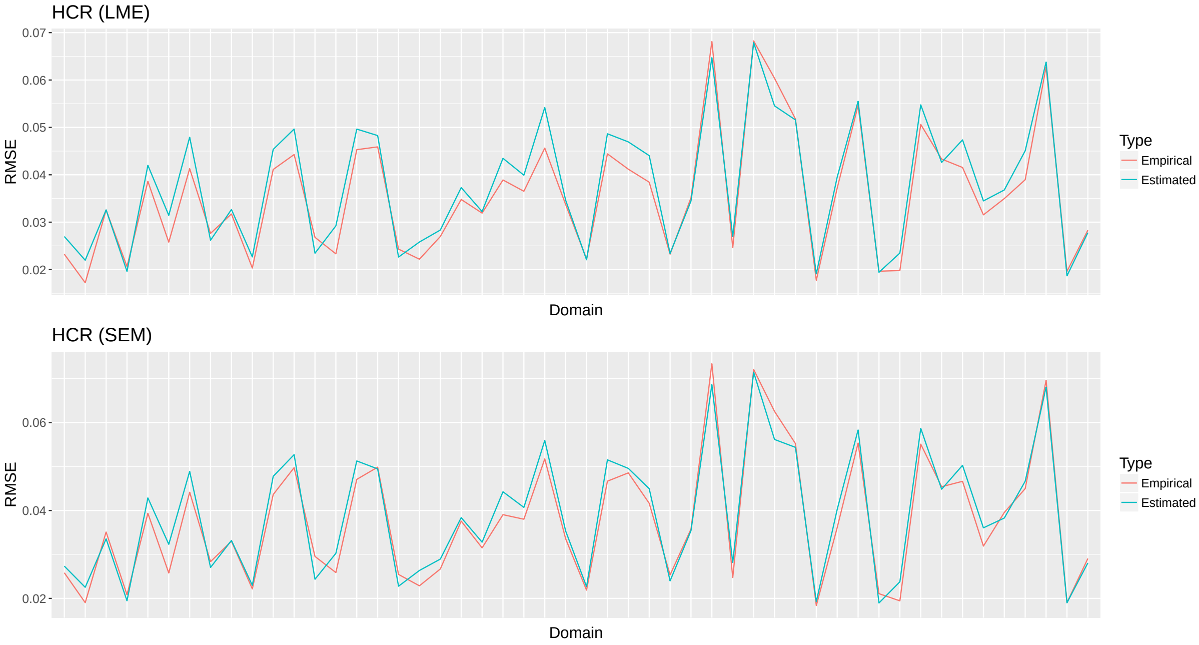

Figure 9: Normal scenario 1, MSE for the HCR.

	LME	LMEBox	SEM	SEMBox
Mean	202.3627	205.5423	256.3309	254.2082
Gini	0.0133	0.0132	0.0156	0.0169
HCR	0.0349	0.0345	0.0392	0.0403

Table 4: Normal scenario 2, median RMSE for the proposed methods.

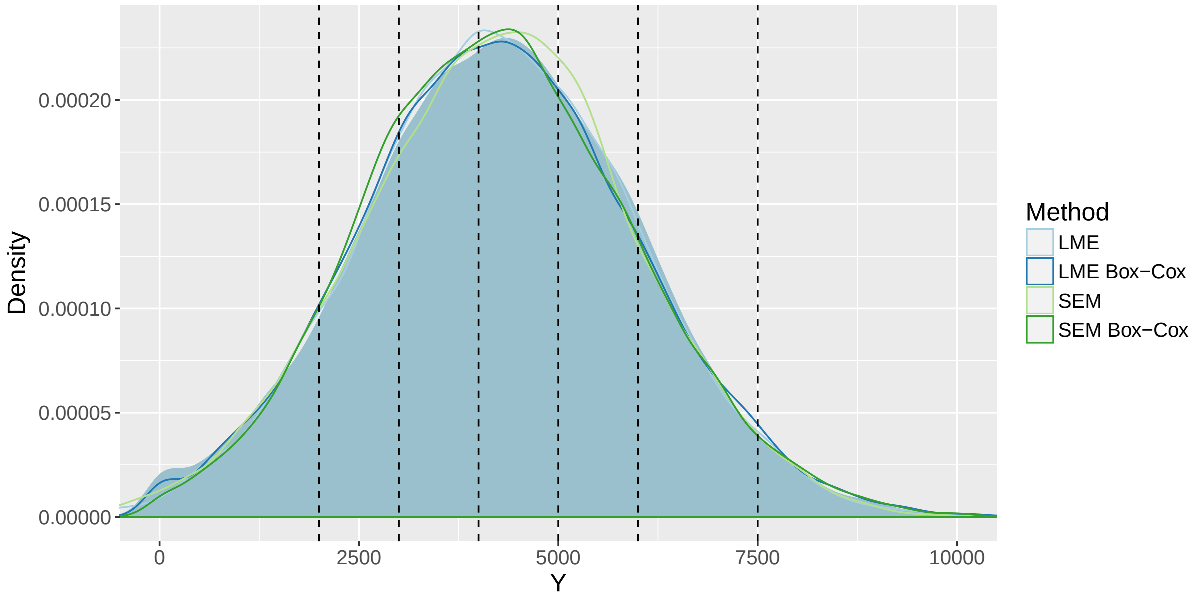

Figure 10: Normal scenario 1, predicted and true density.

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

Table 5: Log-scale scenario, distribution for one arbitrary chosen population.

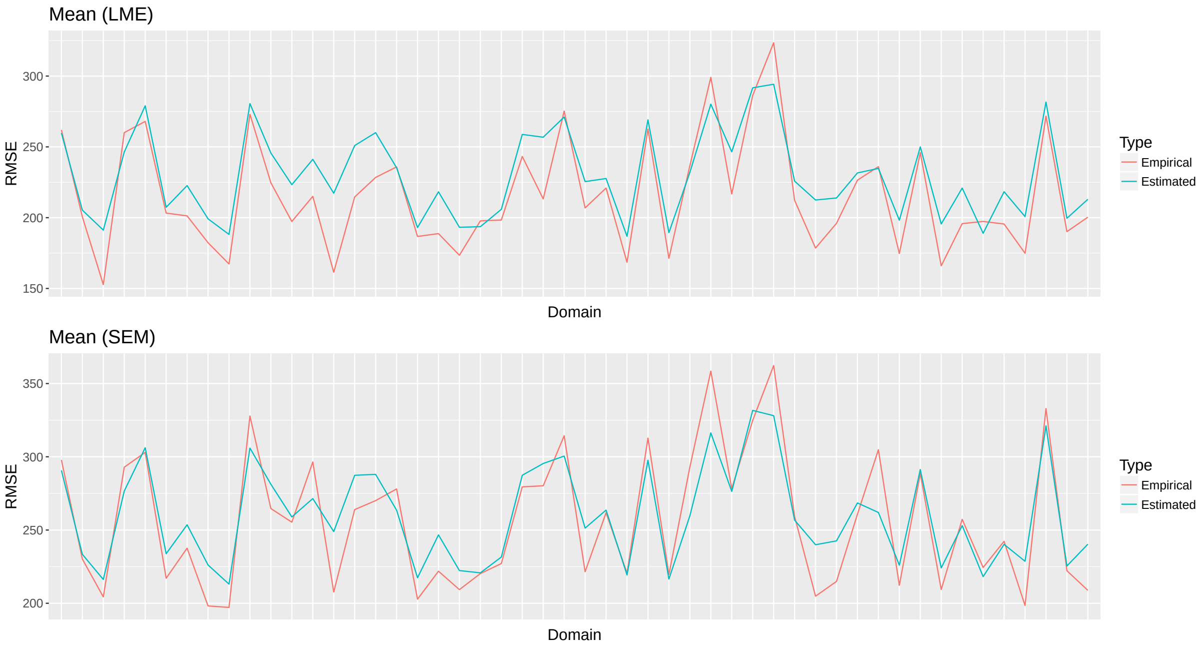

Figure 11: Normal scenario 2, MSE for the mean.

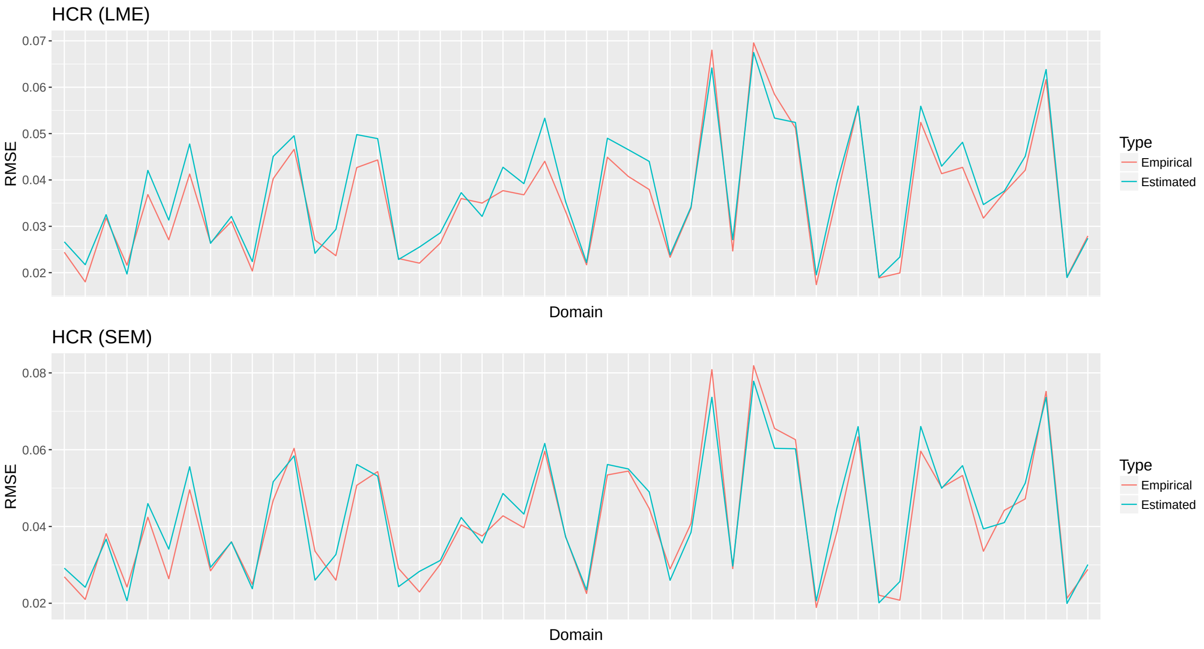

Figure 12: Normal scenario 2, MSE for the HCR

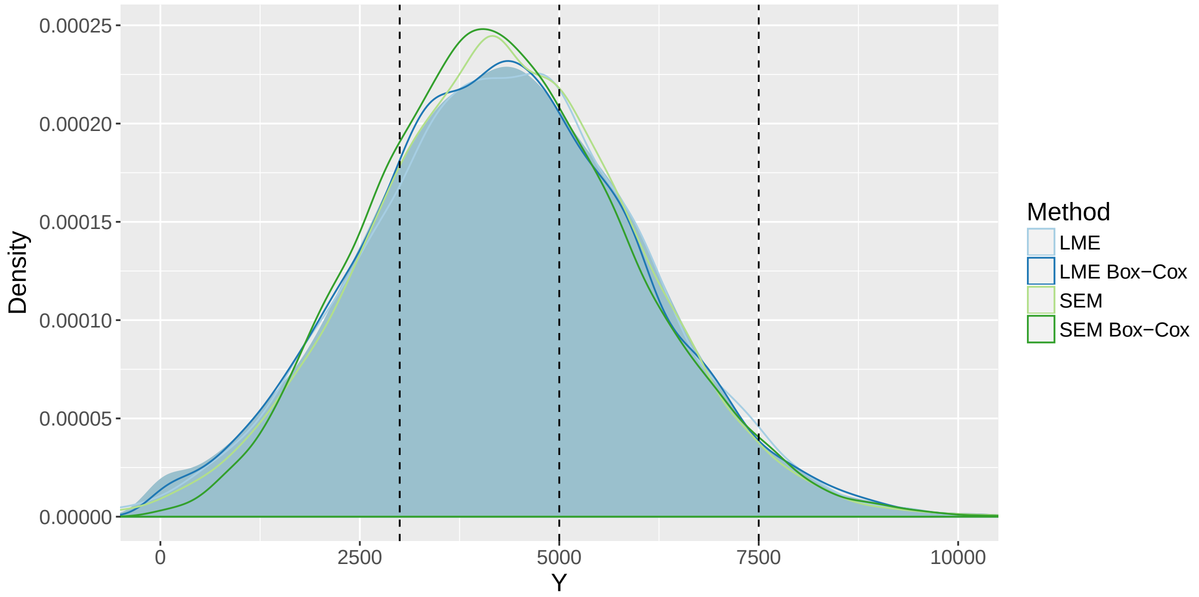

Figure 13: Normal scenario 2, predicted and true density

	LME Log	LME Box-Cox	SEM Log	SEM Box-Cox
Mean	975.0483	974.6505	988.2902	991.2858
Gini	0.0344	0.0359	0.0353	0.0377
HCR	0.0646	0.0638	0.0672	0.0678

Table 6: Log scenario, median RMSE for the proposed methods.

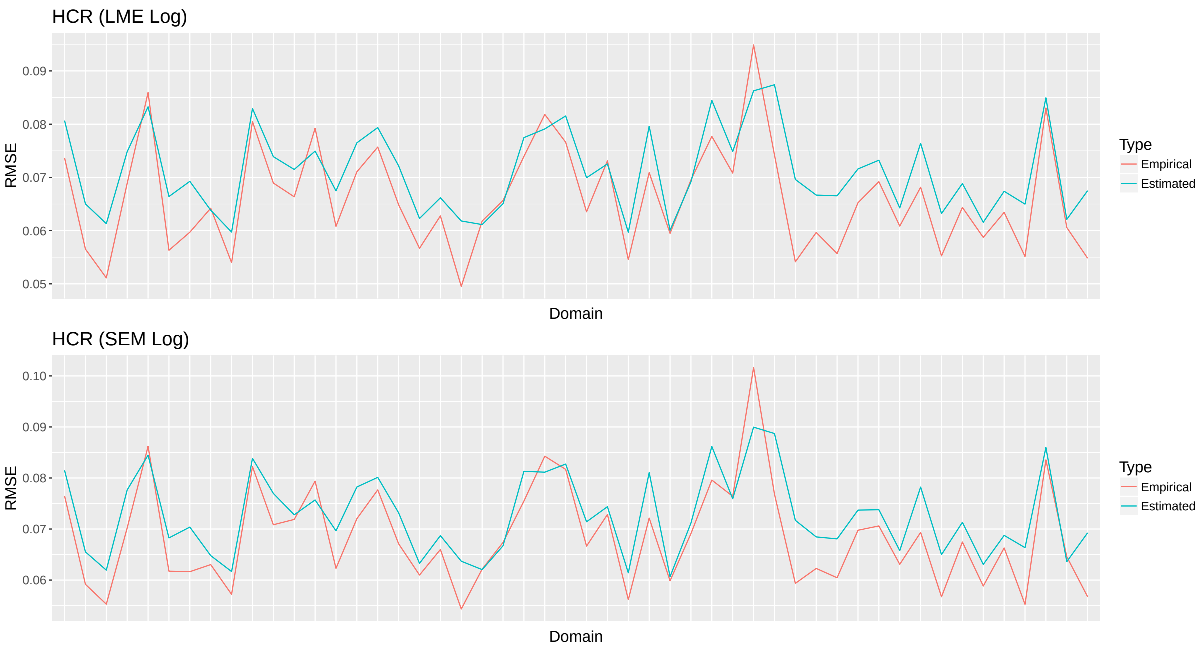

Figure 14: Log scenario, MSE for the HCR.

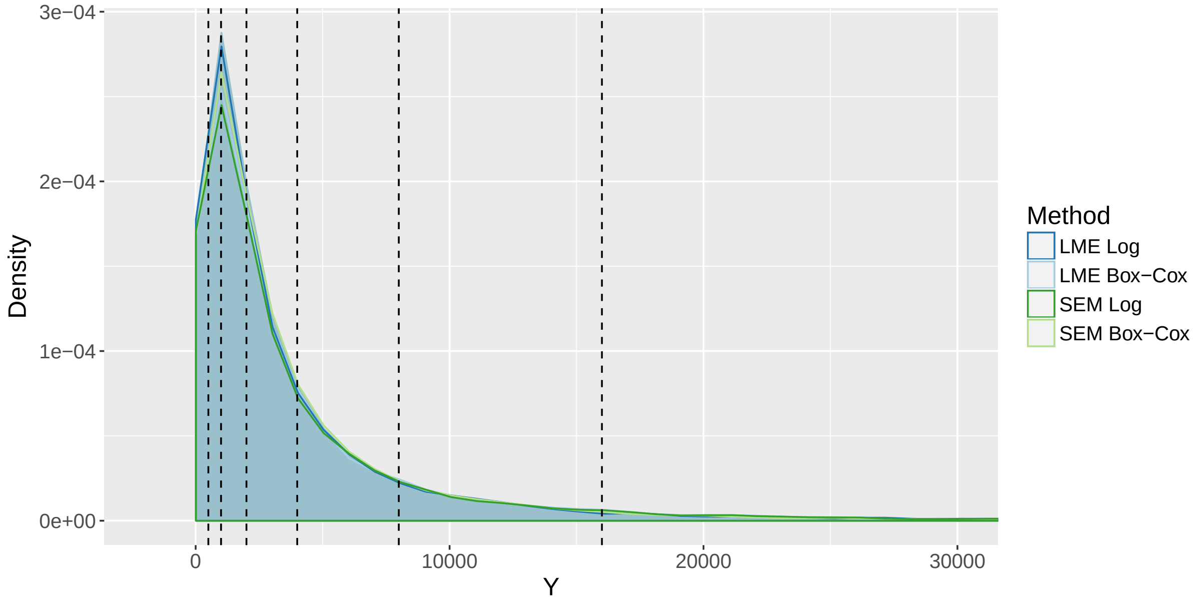

Figure 15: Log scenario, predicted and true density.

Interval	Frequencies
$[1, 3000)$	408
$[3000, 5000)$	1361
$[5000, 7000)$	2580
$[7000, 9000)$	2554
$[9000, 11000)$	1624
$[11000, 13000)$	778
$[13000, Inf)$	695

Table 7: GB2 scenario, distribution for one arbitrary chosen population.

	LME	LME Box-Cox	SEM	SEM Box-Cox
Mean	457.1321	416.0161	463.7191	429.3658
Gini	0.0434	0.0238	0.0258	0.0236
HCR	0.0635	0.0467	0.0395	0.0379

Table 8: GB2 scenario, median RMSE for the proposed methods.

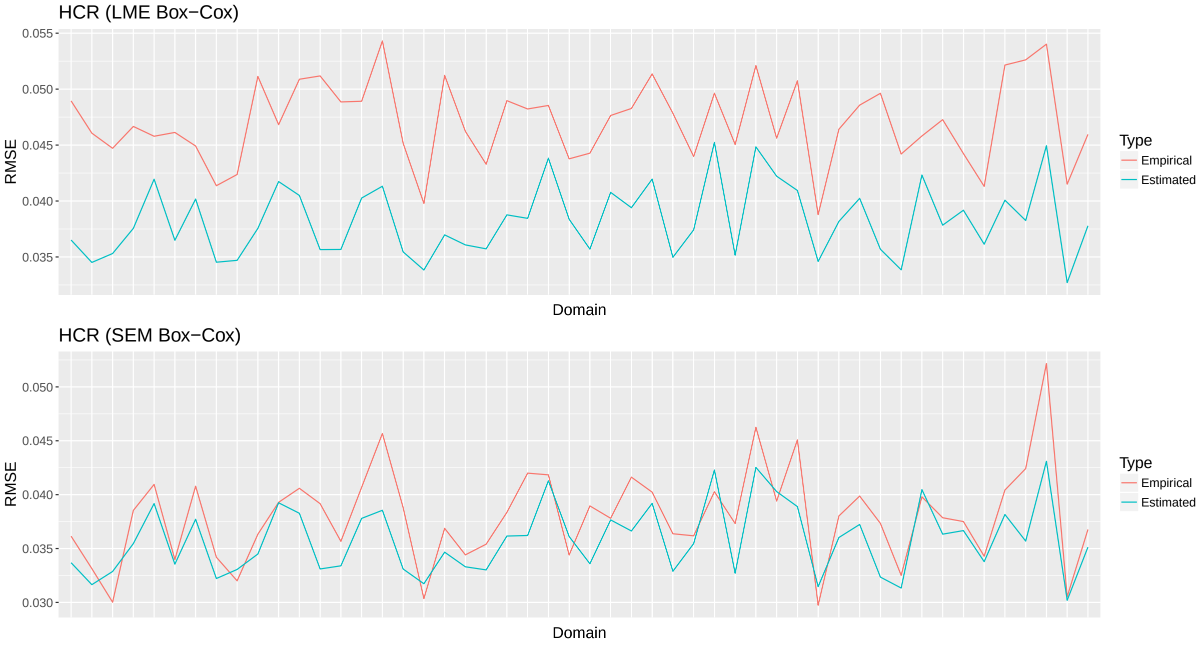

Figure 16: GB2 scenario, MSE for the HCR.

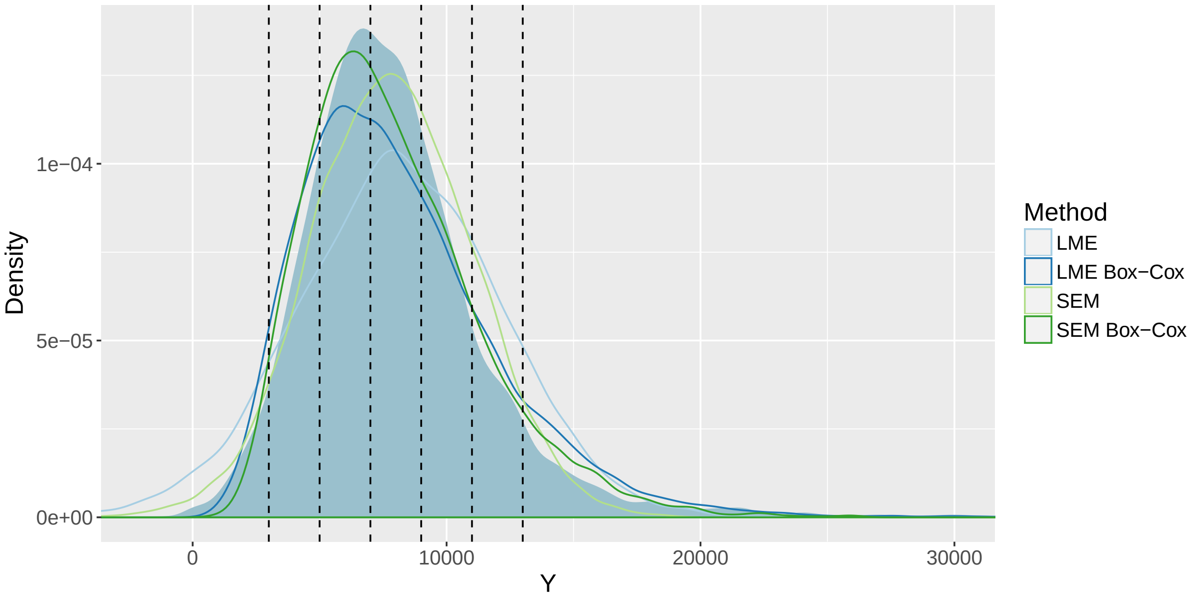

Figure 17: GB2 scenario, predicted and true density.

References

Battese, G. E., Harter, R. M., and Fuller, W. A. (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.

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

Caleux, G. and Dieboldt, J. (1985). The sem algorithm: A probalistic teacher algorithm derived from the em algorithm for the mixture problem. *Computational Statistics Quarterly*, 2:73–82.

Chambers, R. and Tzavidis, N. (2006). M-quantile models for small area estimation. *Biometrika*, 93 (2):255–268.

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

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

Elbers, C., Lanjouw, J. O., and Lanjouw, P. (2003). Micro-level estimation of poverty and inequality. *Econometrica*, 71(1):355–364.

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

Gini, C. (1912). Variabilità e mutabilità. *Reprinted in Memorie di metodologica statistica (Ed. Pizetti E, Salvemini, T). Rome: Libreria Eredi Virgilio Veschi.*

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

Graf, M., Nedyalkova, D., Münich, R., Seger, J., and Zins, S. (2011). Parametric estimation of income distributions and indicators of poverty and social exclusion. *Advanced Methodology for European Laeken Indicators*.

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

In-Known, Y. and Johnson, R. A. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika*, 87(4):954–959.

Manly, B. F. J. (1976). Exponential data transformations. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 25(1):37–42.

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

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

Pfeffermann, D. (2002). Small area estimation - new developments and directions. *International Statistical Review*, 70 (1):125–143.

Rao, J. and I.Molina (2015). *Small Area Estimation*. John Wiley & Sons, Inc.

Rojas-Perilla, N., Pannier, S., Schmid, T., and Tzavidis, N. (2017). Data-driven tansformation in small area estimation.

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.

United Nations (1995). The copenhagen declaration and programme of action. *World Summit for Social Development*.

United Nations (2017). Sustainable development goals. `http://www.un.org/sustainabledevelopment/`. Accessed: 2017-07-22.

Yang, Z. (2000). A modified family of power transformations. *Economics Letters*, 87(92):14–19.

Diskussionsbeiträge - Fachbereich Wirtschaftswissenschaft - Freie Universität Berlin
Discussion Paper - School of Business and Economics - Freie Universität Berlin

2017 erschienen:

2017/1 ARONSSON, Thomas und Ronnie SCHÖB
Habit Formation and the Pareto-Efficient Provision of Public Goods
Economics

2017/2 VOGT, Charlotte; Martin GERSCH und Cordelia GERTZ
Governance in integrierten, IT-unterstützten Versorgungskonzepten im Gesundheitswesen : eine Analyse aktueller sowie zukünftig möglicher Governancestrukturen und -mechanismen
Wirtschaftsinformatik

2017/3 VOGT, Charlotte; Martin GERSCH und Hanni KOCH
Geschäftsmodelle und Wertschöpfungsarchitekturen intersektoraler, IT-unterstützter Versorgungskonzepte im Gesundheitswesen
Wirtschaftsinformatik

2017/4 DOMBI, Akos und Theocharis GRIGORIADIS
Ancestry, Diversity & Finance : Evidence from Transition Economies
Economics

2017/5 SCHREIBER, Sven
Weather Adjustment of Economic Output
Economics

2017/6 NACHTIGALL, Daniel
Prices versus Quantities: The Impact of Fracking on the Choice of Climate Policy Instruments in the Presence of OPEC
Economics

2017/7 STOCKHAUSEN, Maximilian
The Distribution of Economic Resources to Children in Germany
Economics

2017/8 HETSCHKO, Clemens; Louisa von REUMONT und Ronnie SCHÖB
Embedding as a Pitfall for Survey-Based Welfare Indicators: Evidence from an Experiment
Economics

2017/9 GAENTZSCH, Anja
Do Conditional Cash Transfers (CCT) Raise Educational Attainment? A Case Study of Juntos in Peru
Economics

2017/10 BACH, Stefan; Martin BEZNOSKA und Viktor STEINER
An Integrated Micro Data Base for Tax Analysis in Germany
Economics

2017/11 NEUGEBAUER, Martin und Felix WEISS
Does a Bachelor's Degree pay off? Labor Market Outcomes of Academic versus Vocational Education after Bologna
Economics

2017/12 HACHULA, Michael und Dieter NAUTZ
The Dynamic Impact of Macroeconomic News on Long-Term Inflation Expectations
Economics

2017/13 CORNEO, Giacomo
Ein Staatsfonds, der eine soziale Dividende finanziert
Economics

2017/14 GERSCH, Martin; Cordelia GERTZ und Charlotte VOGT
Leistungsangebote in integrierten, IT-unterstützten Versorgungskonzepten: eine Konzeption (re-) konfigurierbarer Servicemodule im Gesundheitswesen
Wirtschaftsinformatik

2017/15 KREUTZMANN, Ann-Kristin; Sören PANNIER; Natalia ROJAS-PERILLA; Timo SCHMID; Matthias TEMPL und Nikos TZAVIDIS
The R Package emdi for Estimating and Mapping Regionally Disaggregated Indicators
Economics

2017/16 VOGT, Charlotte; Cordelia GERTZ und Martin GERSCH
Ökonomische Evaluation eines integrierten, IT-unterstützten Versorgungskonzepts im Gesundheitswesen: eine ökonomische Analyse von E-Health-unterstützten Versorgungsprozessen
Wirtschaftsinformatik

2017/17 GASTEIGER, Emanuel und Klaus PRETTNER
A Note on Automation, Stagnation, and the Implications of a Robot Tax
Economics

2017/18 HAASE, Michaela
The Changing Basis of Economic Responsibility: zur Bedeutung und Rezeption von John Maurice Clarks Artikel zur ökonomischen Verantwortung
Marketing

2017/19 FOSSEN, Frank M.; Ray REES; Davud ROSTAM-AFSCHAR und Viktor STEINER
How Do Entrepreneurial Portfolios Respond to Income Taxation?
Economics

2017/20 NEIDHÖFER, Guido; Joaquín SERRANO und Leonardo GASPARINI
Educational Inequality and Intergenerational Mobility in Latin America: A New Database
Economics

2017/21 SCHMITZ, Sebastian: The Effects of Germany's New Minimum Wage on Employment and Welfare Dependency
Economics