Performance of Empirical Best Predictor in Informative Samples - A Monte Carlo Simulation

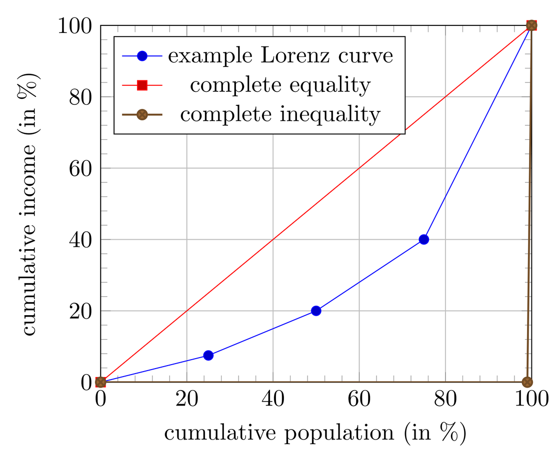
Felix Skarke, Enno Tammena, Christian Koopmann

Freie Universität Berlin, Humboldt Universität zu Berlin

Idea of simulation study

- Dealing with domains of interest, that do not have a sufficiently large sample size can lead to unsatisfactory results, when using direct estimators
  $\Rightarrow$ reasoning behind using small area methods like EBP (Empirical Best Predictor) approach
- When complex sampling designs are used to generate a sample (practical reasons, special interest in small subpopulation), not using sampling weights can lead to biased estimators
  $\Rightarrow$ use of direct estimators like weighted Gini $\Rightarrow$ In model-based inference normally the sampling design is assumed to be uninformative (like SRS): $P(s|y) = P(s), \quad \forall y \in \mathbb{R}^N, \ \forall s$
- Since sampling weights cannot be used directly in the EBP approach the question of this study is how to deal with a sample that has a complex sampling design and also might not have enough subjects in every domain of interest for a direct estimator to deliver good results

The Gini: a meassure for inequality



- The Gini coefficient is used to meassure inequality of distribution (e.g. income, wealth) in a society
- It is defined as the area between the Lorenz curve and the 45Â° line (=A) in relation to the area beneath the 45Â° line (=A+B), where B is the area under the Lorenz curve

Therefore it can be expressed as:

$G = A/(A+B) = 2A = 1 - 2B$

The Gini: unweighted and weighted

The Gini coefficient can be expressed without a direct reference to the Lorenz curve:

unweighted version

$$\hat{G} = \frac{2\sum_{i=1}^{n} iy_i}{n\sum_{i=1}^{n} y_i} - \frac{n+1}{n}$$

weighted version

$$\hat{G} = 100\left[\frac{2\sum_{i=1}^{n}\left(w_i y_i \sum_{j=1}^{i} w_j\right) - \sum_{i=1}^{n} w_i^2 y_i}{\left(\sum_{i=1}^{n} w_i\right)\sum_{i=1}^{n}(w_i y_i)} - 1\right]$$

⚠ Using weights in direct estimators is important, if a complex sampling design is used. Not weighting the observations leads to (sometimes severely) biased estimates.

Empirical Best Predictor

random effects model:

$$y_{ij} = x_{ij}\prime\beta + u_i + e_{ij}, \quad j = 1, \dots, n_i, \quad i = 1, \dots, D$$

,where $u_i \sim \mathcal{N}(0, \sigma_u^2)$ and $e_{ij} \sim \mathcal{N}(0, \sigma_e^2)$

estimation of model:

1. estimate $\hat{\beta}, \hat{\sigma}_u^2, \hat{\sigma}_e^2, \hat{u}_i, \hat{\gamma}_i = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \frac{\hat{\sigma}_e^2}{n_i}}$ from sample
2. generate $e_{ij}^* \sim \mathcal{N}(0, \hat{\sigma}_e^2)$ and $u_i^* \sim \mathcal{N}(0, \hat{\sigma}_u^2(1-\hat{\gamma}_i))$ for L pseudo-populations:
   $y_{ij}^{*(l)} = x_{ij}\prime\hat{\beta} + \hat{u}_i + u_i^* + e_{ij}^*$
   $\Rightarrow$ obtain an estimator of interest in each domain for every pseudo-population
3. calculate $\hat{\theta}_i^{EBP} = \frac{1}{L}\sum_{l=1}^{L}\hat{\theta}_i^{(l)}$ for each domain

Umsetzung

Hier wird das Vorgehen erklï¿½rt:

- ...
- ...
- ...

Implementation

- Dataset: EUSILC Data as provided by the emdi package
- to achieve a sufficient population size N, randomly dublicate observations between 1 and 8 times
- add ε to the dependent variable of duplicated observations
- take a random sample of g observations from 5 income groups to get a sample of size n
- calculate *frequencyweights* $= N_{SMA,incomegroup}/n_{SMA}/5$
- take random sample of c observations for second level EBP data
- for 1:s {
  1. split population by SMA, take a sample of $n_{sma}/5$ from each income group
  2. estimate $Gini_{direct}$, $Gini_{weighted}$
  3. estimate $Gini_{EBP}$ based on l pseudo-populations
  4. estimate $MSE_{Gini_{EBP}}$ based on b bootstraps
  5. expand the sample by frequency weights
  6. estimate $Gini_{weightedEBP}$ based on l pseudo-populations
  7. save the results per SMA

  }
- calculate quality measures per SMA

Parameters: $N = 112644$, $s = 250$, $\varepsilon \sim N(0, 5000)$, $g = 2000$, $n = 10000$, $c = 25000$, $l = 50$, $b = 10$, $SMA = District$

RMSE of weighted and unweighted EBP per Domain

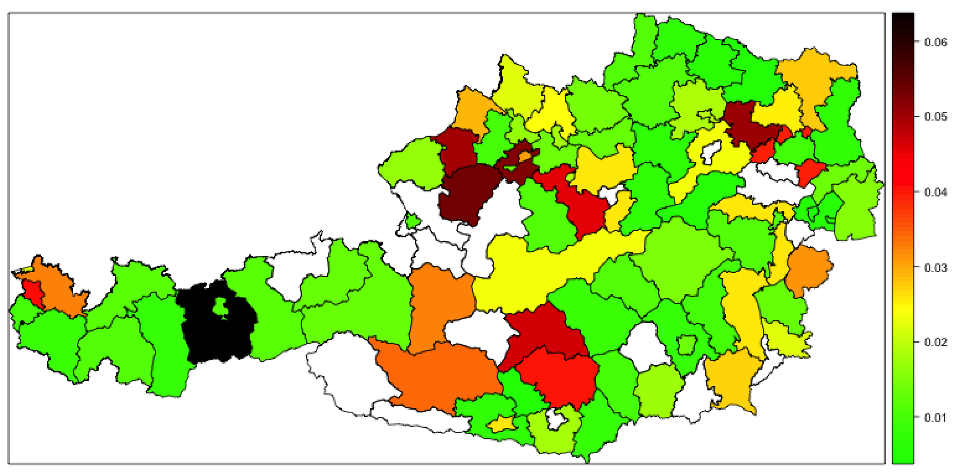


Figure: RMSE of Weighted EBP per Domain

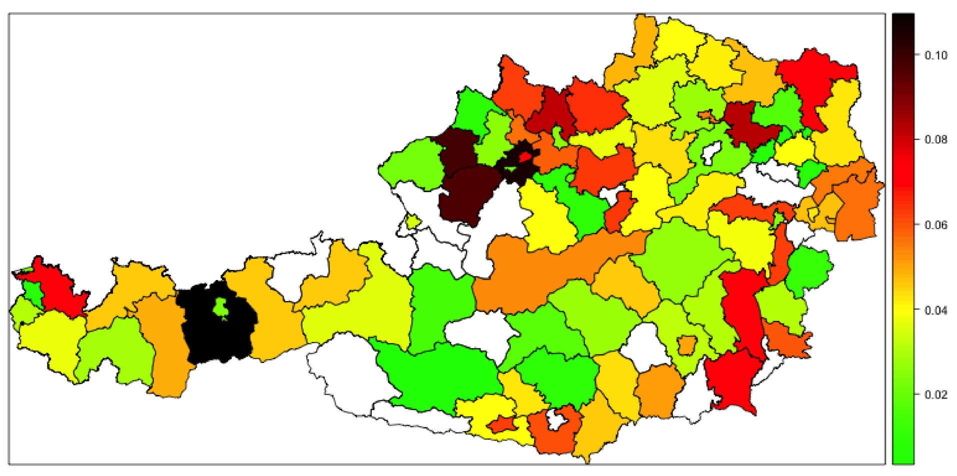


Figure: RMSE of Unweighted EBP per Domain

Domain Level MRE and RMSE of Estimators

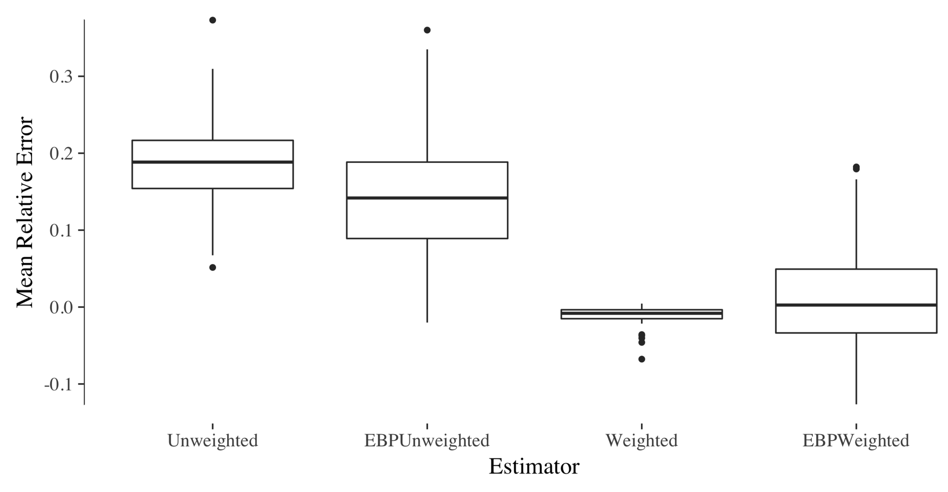


Figure: Boxplot of Mean Relative Error on Domain Level

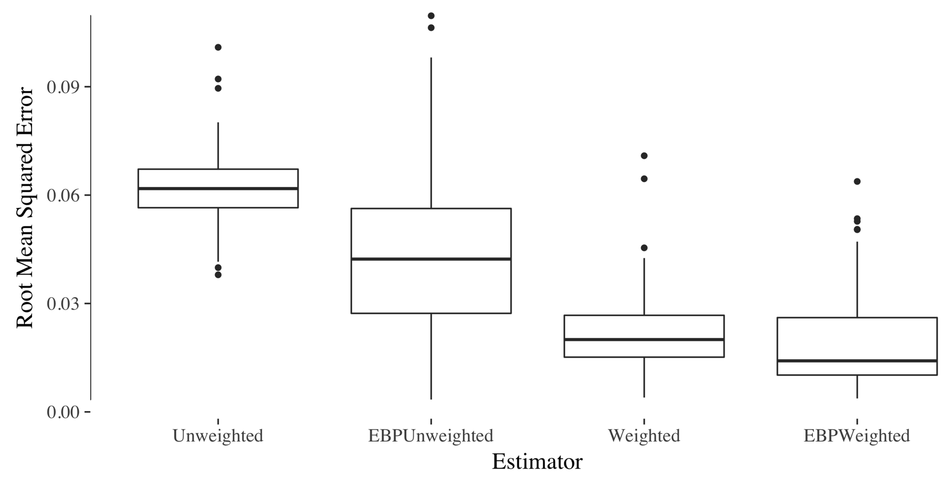


Figure: Boxplot of RMSE on Domain Level

Accuracy of inbuilt MSE estimator

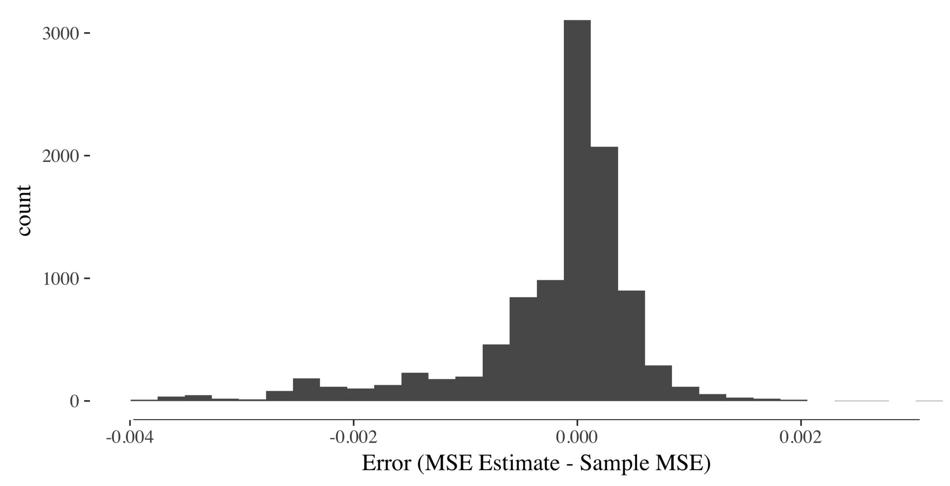


Figure: Histogram of Difference between EBP MSE estimator and the sample MSE across iterations

Observations:

- Bootstrap estimator agrees quite well with sample MSE
- Slightly more likely to underestimate sample MSE

Für weitere Informationen

Kontakt:

Name	Matrikel N.	e-mail
Christian Koopmann	572485	c.k.e.koopmann@gmail.com
Felix Skarke		
Enno Tammena		