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

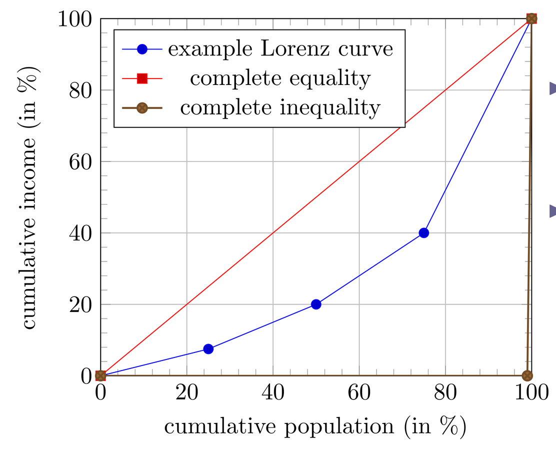
Felix Skarke, Enno Tammena, Christian Koopmann

Freie Universität Berlin, Humboldt Universität zu Berlin

Motivation

- Use of direct estimators in domains of interest with insufficient sample size can lead to unreliable results
  ⇒ Using small area methods like EBP (Empirical Best Predictor) approach can be preferable
- In model-based inference normally the sampling design is assumed to be uninformative (e.g. simple random sampling):
  $P(s|y) = P(s), \ \forall y \in \mathbb{R}^N, \ \forall s$
- When complex designs are used for sampling (practical reasons, special interest in small subpopulation), ignoring weights can lead to biased estimators
  ⇒ use of direct estimators like weighted Gini
- Problem: sampling weights cannot be used directly in the EBP approach because of non-linearity
- The aim of this study is to evaluate the performance of EBP relative to direct estimation under a complex sampling design

The Gini: a measure for inequality



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

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

⚠ Using weights in direct estimators can be important, if a complex sampling design is used.

Empirical Best Predictor

Random Effects model:

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

,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:
   $$y_{ij}^{*(l)} = x_{ij}'\hat{\beta} + \hat{u}_i + u_i^* + e_{ij}^*$$
   ⇒ 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

MSE estimation

Since analytical approximations of the MSE are difficult to derive in the case of nonlinear indicators like the Gini the MSE is approximated using a bootstrap procedure:

1. Fit the model to the sample data and therefore obtain estimates for $\beta, \sigma_u^2 and \sigma_e^2$
2. Generate $u_i^* \sim iid\mathcal{N}(0, \hat{\sigma}_u^2)$ and $e_{ij}^* \sim iid\mathcal{N}(0, \hat{\sigma}_e^2)$ independently for every domain and every person in the population
3. Construct a superpopulation using the generated error terms and the population covariates
   $$y_{ij}^* = x_{ij}'\hat{\beta} + u_i^* + e_{ij}^*$$
4. Draw B bootstrap populations from the superpopulation and calculate θ_i^{*b} in every domain for each of the populations
5. Draw a bootstrap sample from every population, implement the EBP and estimate $\hat{\theta}_i^{*b}$
6. Finally estimate the MSE per domain by:
   $$\hat{MSE} = \frac{1}{B}\sum_{b=1}^{B}(\hat{\theta}_i^{*b} - \theta_i^{*b})^2$$

Data and Sampling

Assume the following scenario:

1. Stratified sample (i income groups nested in SMAs) with data on y and X
   - In each SMA, an equal number of observations per income group is sampled such that:
     - Sample size n_{SMA} per SMA is proportional to N_{SMA}
     - Smaller income groups are oversampled
   - Weights differ between income groups and SMAs
     $\rightarrow frequencyweights_{SMA,incomegroup} = \frac{N_{SMA,incomegroup}}{n_{SMA}/i}$
2. Population data on X on SMA level is available
   - approach is generalizable to other forms of sampling, where certain groups are over- or undersampled in a stratified setting
3. Dataset: EUSILC Data as provided by the emdi package
   - to achieve a sufficient population size N, randomly duplicate observations between 1 and 8 times
   - add ε to the dependent variable of duplicated observations, where $\varepsilon \sim N(0, 5000)$

Implementation

- take an initial random sample of size n with same number (g) of observations from each income group
- calculate n_{SMA}
- calculate *frequencyweights*
- for 1:s {
  1. split population by SMA, take a sample of $n_{SMA}/5$ from each income group
  2. estimate $\widehat{Gini}_{direct\_unweighted}$, $\widehat{Gini}_{direct\_weighted}$ per SMA
  3. estimate $\widehat{Gini}_{EBP\_unweighted}$ based on l pseudo-populations per SMA
  4. estimate $\widehat{MSE}_{Gini_{unweighted\_EBP}}$ based on b bootstraps per SMA
  5. expand the sample by frequency weights
  6. estimate $\widehat{Gini}_{EBP\_weighted}$ based on l pseudo-populations per SMA
  7. save the results per SMA

  } calculate $Gini_{SMA}$ based on population
- calculate quality measures per SMA
  - $MSE_{SMA} = \sum_{i=1}^{s}(\widehat{Gini}_{SMA} - Gini_{SMA})^2$
  - $RelBias_{SMA} = \sum_{i=1}^{s}(\widehat{Gini}_{SMA} - Gini_{SMA})/Gini_{SMA}$
- Parameters: $N = 112644, s = 250, g = 2000, n = 10000, l = 50, b = 10, SMA = district, i = 5$

RMSE of weighted and unweighted EBP

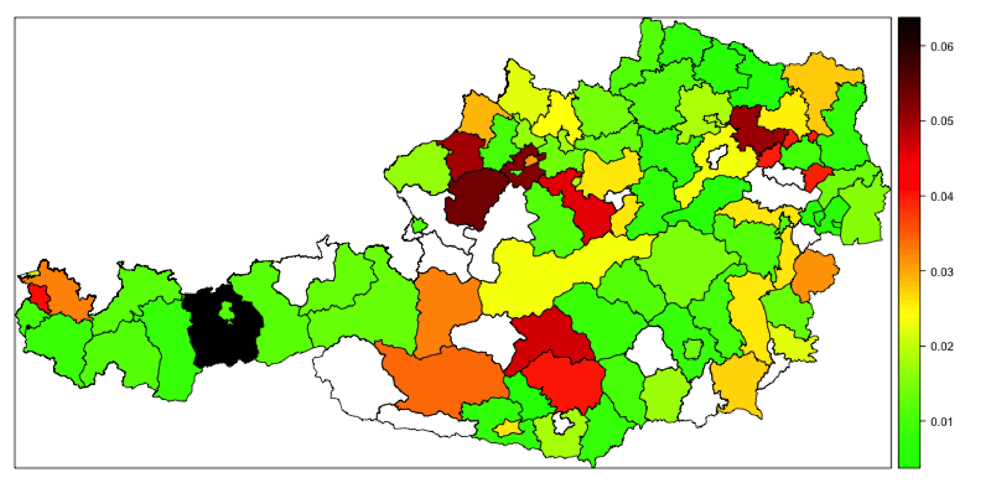


Figure : RMSE of Weighted EBP per Domain

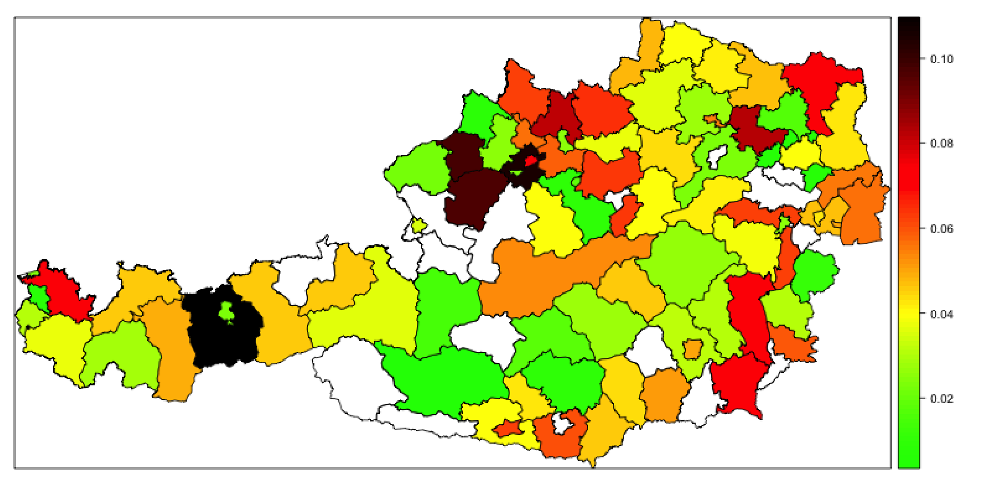


Figure : RMSE of Unweighted EBP per Domain

Domain Level MRE and RMSE of Estimators

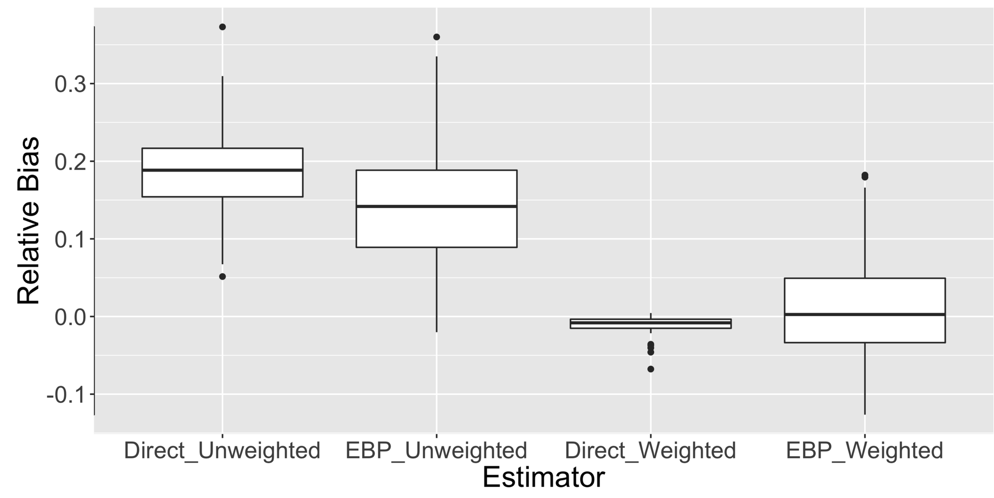


Figure : Boxplot of Relative Bias on Domain Level

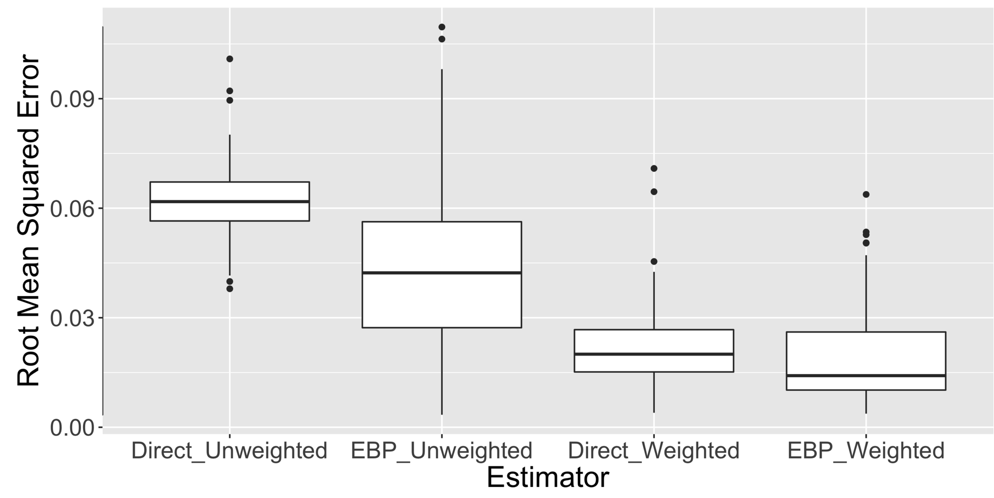


Figure : Boxplot of RMSE on Domain Level

Accuracy of inbuilt MSE estimator

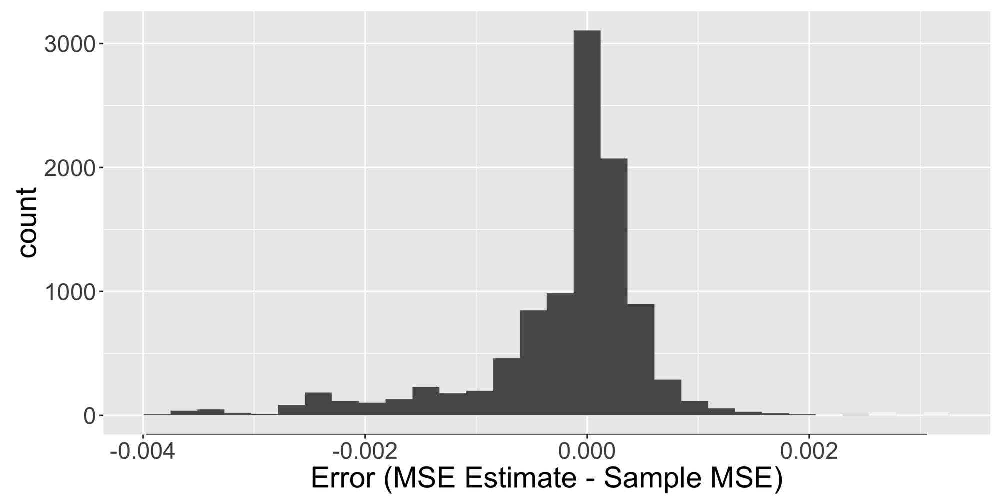


Figure : Histogram of Difference between EBP MSE estimator and the sample MSE across iterations

Observations:

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

Conclusions

Conclusions:

- Unweighted direct estimator and EBP are both biased unlike their weighted versions.
- Weighting substantially increase accuracy for both estimators.
- Relative accuracy advantage of EBP over direct estimation is much larger in the unweighted case.
- Weights can be included in EBP estimation by expanding the sample using frequency weights.

Future Research:

- Explore alternative ways of including weights into EBP estimation.
- Analyse alternative sampling designs.

FÜR WEITERE INFORMATIONEN

Kontakt:

Name	Matrikel N.	e-mail
Christian Koopmann	572485	koopmanc@hu-berlin.de
Felix Skarke	573653	f.skarke@fu-berlin.de
Enno Tammena	572575	tammenae@hu-berlin.de