Master Thesis: Poverty estimation based on M-quantile Small Area Methods

Enno Tammena

M.Sc. Statistics

Humboldt–Universität zu Berlin

Advisors:

Prof. Dr. Timo Schmid

Prof. Dr. Ulrich Rendtel

Freie Universität Berlin

Fachbereich Wirtschaftswissenschaft

Institut für Statistik und Ökonometrie

Motivation

"The diffusion of technological change throughout statistics is closely tied to its embodiment in statistical software." (Koenker and Hallock, 2001, p.153)

Motivation

Current Situation:

- M-quantile regression developed by Breckling and Chambers (1988), MQ-SAE approach by Chambers and Tzavidis (2006), but no R-Packages exist
- Some R-Code available from the "S.A.M.P.L.E. Project", but limited functionality and not very user friendly

Aim of this thesis:

- Provide user friendly functions, that can be integrated in the `emdi` package
- Compare performance to direct estimation and the EBP
- $\Rightarrow$ Facilitate poverty estimation with linear and non-linear indicators based on the MQ approach for future research

Outline

M-Estimation of the Central Point

- Consider the following minimization problem:

$$\min_{\theta} \sum \rho \underbrace{(x_i - \theta)}_{e_i} \tag{1}$$

- θ_{Mean}: Solution to a quadratic loss function $\rho_{L2} = (\cdot)^2$
- θ_{Median}: Solution to an absolute loss function $\rho_{L1} = |(\cdot)|$
- θ_{Huber}: Solution the HP2 loss function

The HP2 Loss function

- The HP2 loss function is given by:

$$\rho_{HP2}(e) = \begin{cases} \frac{1}{2}e^2 & for|e| \leq k \\ k|e| - \frac{1}{2}k^2 & for|e| > k \end{cases}$$

  where k is a tuning parameter.
- for $k \to \infty, \rho_{HP2} \to \rho_{L2}$
- for $k \to 0, \rho_{HP2} \to \rho_{L1}$

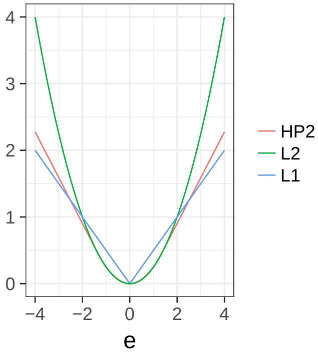

Figure 1: L1, L2, HP2 Loss Functions

M-quantiles

In a finite sample $x_1, \dots, x_n$ the τth M-quantile is obtained by solving

$$\min_{\theta} \sum_{i=1}^{N} \rho_{MQ} \underbrace{(x_i - \theta)}_{e_i} \tag{2}$$

where

$$\rho_{MQ}(e) = \begin{cases} \rho_{HP2}(e)(1-\tau) & \textit{for } e \leq 0 \\ \rho_{HP2}(e)\tau & \textit{for } e > 0 \end{cases}$$

$$\rho_{HP2}(e) = \begin{cases} \frac{1}{2}e^2 & \textit{for} |e| \leq k \\ k|e| - \frac{1}{2}k^2 & \textit{for} |e| > k \end{cases}$$

and $\tau \in [0, 1]$

Plots of M-quantiles, quantiles and expectiles

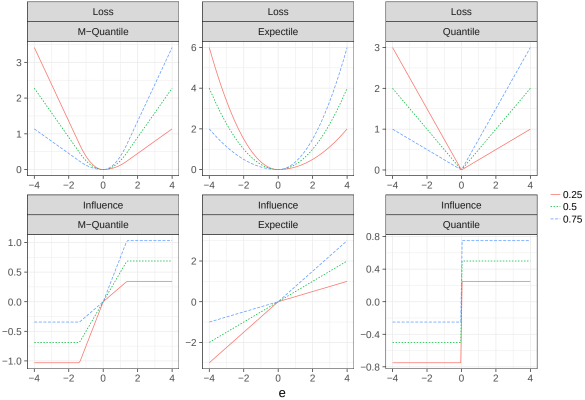

Figure 2: Asymmetric Loss and Influence functions

Regression

- OLS regression line (or hyperplane) models the conditional *mean* of Y given X

$$E[y|X] = X\beta_{OLS} \tag{3}$$

- Quantile regression line (or hyperplane) models the qth conditional *Quantile* of Y given X

$$Q_q[y|X] = X\beta_q \tag{4}$$

- M-quantile regression line (or hyperplane) models the τth conditional *M-quantile* of Y given X

$$MQ_\tau[y|X] = X\beta_\tau \tag{5}$$

M-quantile Regression

If ρ is differentiable:

- influence function $\psi(e) = d\rho(e)/d\theta$
- weight function: $w(e) = \frac{\psi(e)}{e}$
- Use Iteratively re-weighted least squares (IWLS)
  1. Starting solution: e.g. $\hat{\beta}^{(0)} = \hat{\beta}_{OLS}$
  2. Repeat s times until convergence:
     - 2.1 At iteration s calculate $e_i^{(s-1)} = y_i - \hat{y}_i^{(s-1)}$
     - 2.2 Apply the weight function $w_{HP2,\tau}(e)$
     - 2.3 Weighted regression: $\hat{\beta}_\tau^{(s)} = [X'W^{(s-1)}X]^{-1}X'W^{(s-1)}y$, where $W^{(s-1)} = diag(w_1^{(s-1)}, \dots, w_n^{(s-1)})$

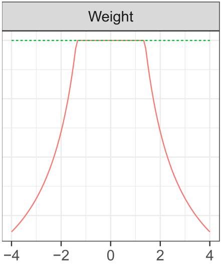

Figure 3: HP2 Weight function for $\tau_{0.5}$

Pseudo-Random Effects with M-quantiles

Assume observations y_i can be attributed to d clusters/domains.
Note that $\forall y_i \exists \tau_i$ such that $y_i = MQ_{\tau_i}(y|x)$

1. Run M-quantile regressions for a fine grid of τ values
2. For each y_i find the corresponding τ_i, using an interpolation procedure
3. Average the τ for each domain j using the mean $\bar{\tau}_j = \frac{1}{N_j} \sum_{i \in j} \tau_i$
4. Run M-quantile regressions for the τ_j values resulting from (3) to obtain the pseudo random effects

Pseudo Random Effect of Area j: $\hat{\beta}_{\hat{\bar{\tau}}_j}$

Pseudo-Random Effects Example

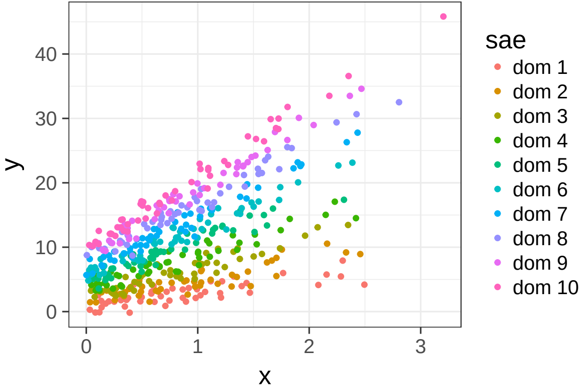

Figure 4: Scatterplot of Example Scenario

Pseudo-Random Effects Example 2

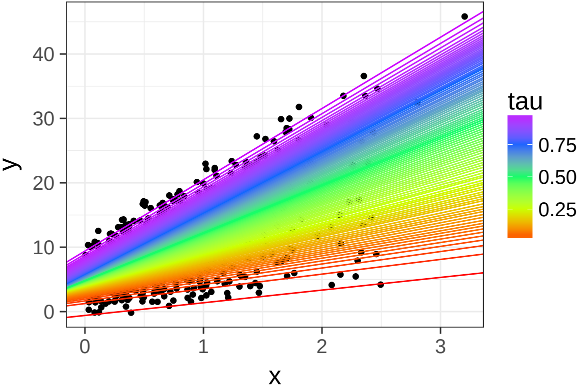

Figure 5: Fine grid of M-quantile Regression lines

Pseudo-Random Effects Example 3

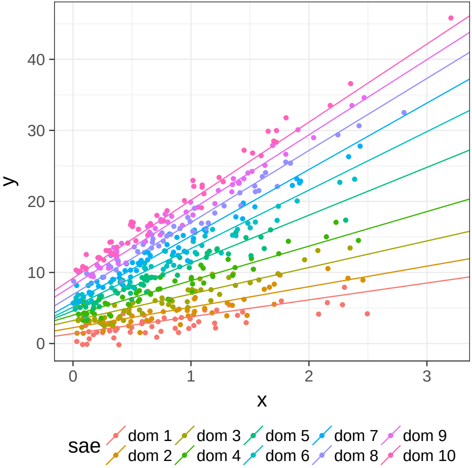

Figure 6: Plot of Pseudo Random Effects

Naive Estimator of the Mean

- Assume the basic unit level model

$$y_{ij} = x_{ij}^T \beta + \vartheta_j + e_{ij} \tag{6}$$

- With units $i = 1, \ldots, n_j$ in domains $j = 1, \ldots, d$
- Consider e.g. the EBLUP estimator for the mean (under assumptions):

$$\hat{\theta}_j^{EBLUP} = \frac{1}{N_j} \left(\sum_{i \in s_j} y_{ij} + \sum_{i \in r_j} \underbrace{(x_{ij}^T \hat{\beta} + \hat{\vartheta}_i)}_{\hat{y}_{ij}} \right) \tag{7}$$

- Naive MQ-estimator given for $\hat{y}_{ij} = x_{ij}^T \hat{\beta}_{\hat{\bar{\tau}}_j}$

EBP Algorithm

1. Using the sampled data, obtain estimators $\hat{\beta}, \hat{\sigma}_u^2, \hat{\sigma}_e^2, \hat{u}_j = E[u_j|y_j], \hat{\gamma}_j = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \frac{\hat{\sigma}_e^2}{n_j}}$
2. For $l = 1, .., L$
   2.1 Generate pseudo-populations: $y_{ij}^{*(l)} = x_{ij}^T \hat{\beta} + \hat{\vartheta}_j + \hat{u}_j + u_j^* + e_{ij}^*$ where
      - For sampled domains: $e_{ij}^* \sim \mathcal{N}(0, \hat{\sigma}_e^2)$ and $u_j^* \sim \mathcal{N}(0, \hat{\sigma}_u^2(1 - \hat{\gamma}_i))$
      - For non-sampled domains: $e_{ij}^* \sim \mathcal{N}(0, \hat{\sigma}_e^2)$ and $u_j^* \sim \mathcal{N}(0, \hat{\sigma}_u^2), \hat{u}_j = 0$

   2.2 Calculate linear or non-linear target indicator $\hat{\theta}_j^{(l)}$
3. Calculate $\hat{\theta}_j^{EBP} = \frac{1}{L} \sum_{l=1}^{L} \hat{\theta}_j^{(l)}$ for each domain

Motivation for the MQ (SAE)-Approach

$$\hat{F}_j^{Naive}(t) = \frac{1}{N_j}[\sum_{i \in s_j} I(y_{ij} \leq t) + \sum_{k \in r_j} I(\hat{y}_{kj} \leq t)] \tag{8}$$

$$\hat{F}_j^{CD}(t) = \frac{1}{N_j}[\sum_{i \in s_j} I(y_{ij} \leq t) + \frac{1}{n_j} \sum_{i \in s_j} \sum_{k \in r_j} I(\hat{y}_{kj} + e_{ij}^{*} \leq t)] \tag{9}$$

The MQ (SAE) Algorithm

1. Using the sampled data, obtain the M-quantile pseudo random effects. Calculate vector of residuals $e_r = (e_{11}, \ldots, e_{n_j j_d})^T$ with $e_{ij} = y_{ij} - x_{ij}^T \hat{\beta}_{\hat{\tau}_j}$
2. For $l = 1, .., L$
   2.1 Generate pseudo population values $y_{ij}^* \in r_j$
      - For $e_{11}^*, \ldots, e_{ij}^*, i \in r_j$, take a random sample with replacement from e_r of length $N_j - n_j$
      - Generate $y_{ij}* \in r_j$ of length $N_j - n_j$ with $y_{ij}^{*(l)} = x_{ij}^T \hat{\beta}_{\hat{\tau}_j} + e_{ij}^*$

   2.2 Combine vector of sampled values $y_{ij} \in s_j$ and $y_{ij}^* \in r_j$

   2.3 Calculate linear or non-linear target indicator $\hat{\theta}_j^{(l)}$
3. Calculate $\hat{\theta}_j^{MQ} = \frac{1}{L} \sum_{l=1}^{L} \hat{\theta}_j^{(l)}$ for each domain

Non-Parametric Bootstrapping for MQ (SAE)

1. Using the sampled data, obtain the M-quantile pseudo random effects. Calculate centered vector of residuals e_r based on $e_{ij} = y_{ij} - x_{ij}^T \hat{\beta}_{\hat{\hat{\tau}}_j}$
2. for $b = 1, \ldots, B$
   2.1 Generate superpopulation, maintaining N_j: $y_{ij}^{*(b)} = x_{ij}^T \hat{\beta}_{\hat{\hat{\tau}}_j} + e_{ij}^*$, where e_{ij}^* results from random sample with replacement from e_r
   2.2 for $s = 1 \ldots S$
      - Take a stratified random sample without replacement from superpopulation such that $n_j^* = n_j$
      - Use MQ (SAE) Algorithm to calculate measure of interest
3. Calculate bias and variance

MSE Calculation

$$\hat{Bias}(\hat{\nu}_j) = \frac{1}{B}\frac{1}{S}\sum_{b=1}^{B}\sum_{s=1}^{S}\left[\hat{\nu}_j^{*bs} - \nu_j^{*b}\right] \qquad (10)$$

$$\hat{Var}(\hat{\nu}_j) = \frac{1}{B}\frac{1}{S}\sum_{b=1}^{B}\sum_{s=1}^{S}\left[\hat{\nu}_j^{*bs} - \bar{\hat{\nu}}_j^{*bs}\right]^2 \qquad (11)$$

$$\hat{MSE}(\hat{\nu}_j) = \hat{Var}(\hat{\nu}_j) + \hat{Bias}(\hat{\nu}_j)^2 \qquad (12)$$

Implementation

Package `mquantreg`

- `mquantile` M-quantiles of a variable
- `mq` M-quantile regression
- `mmqm` Mixed M-quantile models
- `mq_sae` M-quantile small area models incl. non-parametric bootstrapping
- $+$ plot, summary, print, predict functions

Replication: Setup

Scenario

Normal Scenario

$$y_{ij} = 3000 - 150 * x_{ij} + \gamma_j + \epsilon_{ij}$$
$$\gamma_j \sim N(0, 200^2)$$
$$\epsilon_{ij} \sim N(0, 800^2)$$
$$x_{ij} \sim N(\mu_j, 1)$$
$$\mu_j \sim U[4, 10]$$

μ_j *are fixed over all simulations.*

(c.f. Marchetti et al., 2012)

Scenario

χ^2 *Scenario*

$$y_{ij} = 11 - x_{ij} + \gamma_j + \epsilon_{ij}$$
$$\gamma_j \sim \chi(1)$$
$$\epsilon_{ij} \sim \chi(6)$$
$$x_{ij} \sim N(\mu_j, 1)$$
$$\mu_j \sim U[8, 11]$$

μ_j *are fixed over all simulations.*
Error terms are centered.

Replication: Setup 2

Do 500 times:

- Generate population of size N
- Take stratified random sample for 30 areas (without replacement)
- Run MQ (SAE) algorithm with $L = 50$ iterations $B = 1$ bootstrap populations and $S = 400$ bootstrap samples

Fixed $N = 840, 50 \leq N_j \leq 150$, $n = 282$, $5 \leq n_j \leq 15$

$$RB(\hat{\nu}_j) = \frac{1}{H}\sum_{h=1}^{H}\left(\frac{\hat{\nu}_{j,h} - \nu_j}{\nu_j}\right) \tag{13}$$

$$RMSE(\hat{\nu}_j) = \left[\frac{1}{H}\sum_{h=1}^{H}(\hat{\nu}_{j,h} - \nu_j)^2\right]^{0.5} \tag{14}$$

Replication Results: Point Estimation

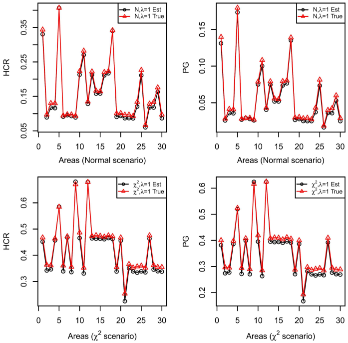

Figure 7: Point-Estimation Results (Marchetti et al., 2012)

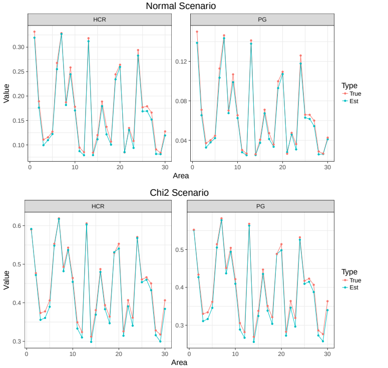

Figure 8: Point-Estimation Results (Replication)

Replication Results: MSE Estimation

χ^2 scenario	HCR Original	HCR Replicated	PG Original	PG Replicated
True	0.096	0.098	0.094	0.097
Estimated	0.096	0.099	0.095	0.099
Rel. Bias (%)	0.19	2.58	0.26	3.49
RMSE	0.019	0.021	0.019	0.022
Normal scenario	**HCR Original**	**HCR Replicated**	**PG Original**	**PG Replicated**
True	0.063	0.065	0.028	0.033
Estimated	0.059	0.063	0.025	0.031
Rel. Bias (%)	-7.66	-7.82	-12.06	-10.47
RMSE	0.022	0.022	0.012	0.014

Table 1: Comparision of Original and Replicated Result

Simulations: General Setup

Do 500 times:

- Generate population according to the defined scenario
- Take stratified random sample without replacement of n_j per area
- Using the sample data, calculate estimators:
  - The direct estimator
  - The EBP estimator, with L=50 Monte-Carlo iterations and Box-Cox transformation
  - The MQ estimator, with L=50 Monte-Carlo iterations
- Calculate true value per bootstrap population

Fixed $N_j = 200$, $8 \leq n_j \leq 29$ and $j = 1 \ldots, 50$

Normal Scenario: Setup

Scenario

Normal Errors

$$
\begin{aligned}
y_{ij} &= max(4500 - 400 * x_{ij} + \vartheta_j + e_{ij}, 0) \\
x_{ij} &\sim N(\mu_j, 3) \\
\vartheta_j &\sim N(0, 500^2) \\
e_{ij} &\sim N(0, 1000^2) \\
\mu_j &\sim U[-3, 3]
\end{aligned}
$$

Normal Scenario: Bias/Rel. Bias

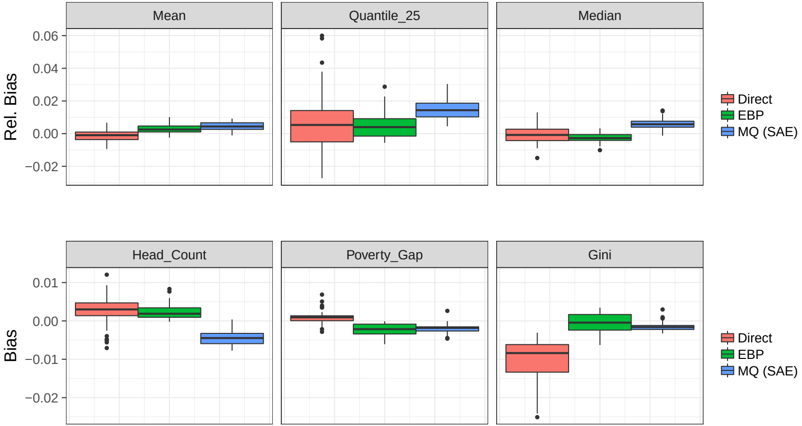

Figure 9: Rel. Bias of Point-Estimation Results for EBP vs. MQ

Normal Scenario: RMSE

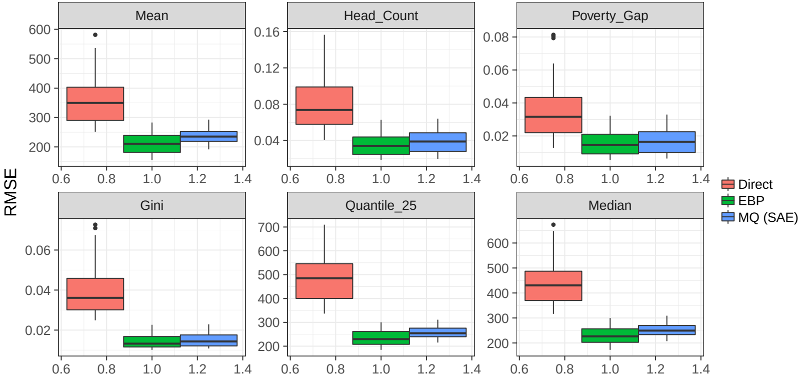

Figure 10: RMSE of Point-Estimation Results for EBP vs. MQ vs. Direct

Log-Outcomes Scenario: Setup

Scenario

Log-Scale Outcomes

$$log(y_{ij}) = 10 - x_{ij} + \vartheta_j + e_{ij}$$
$$x_{ij} \sim N(\mu_j, 0.5)$$
$$e_{ij} \sim N(0, 0.8^2)$$
$$v_j \sim N(0, 0.4^2)$$
$$\mu_j \sim U[2, 3]$$

Log-Outcomes: Rel. Bias

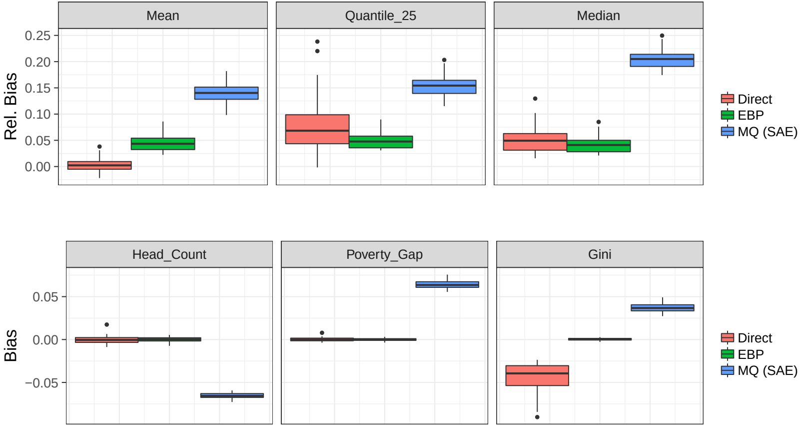

Figure 11: Point-Estimation Results for EBP vs. MQ in scenario

Log-Outcomes: RMSE

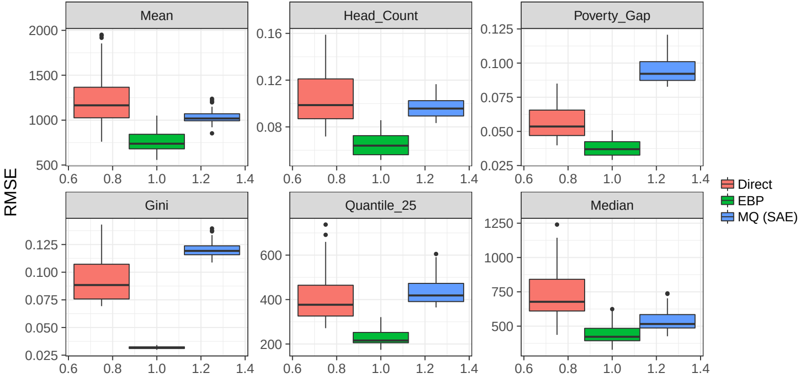

Figure 12: RMSE of Point-Estimation Results for EBP vs. MQ in scenario log-scale

Contaminated Normal Scenario: Setup

Scenario

Contaminated Normal Errors

$$
\begin{aligned}
y_{ij} &= max(4500 - 400 * x_{ij} + \vartheta_j + e_{ij}, 0) \\
x_{ij} &\sim N(\mu_j, 3) \\
\vartheta &\sim N(0, 500^2) \\
e_{ij} &\sim 0.95 * N(0, 1000^2) + 0.05 * N(0, 6000^2) \\
\mu &\sim U[-3, 3]
\end{aligned}
$$

Cont. Normal Scenario: Relative Bias

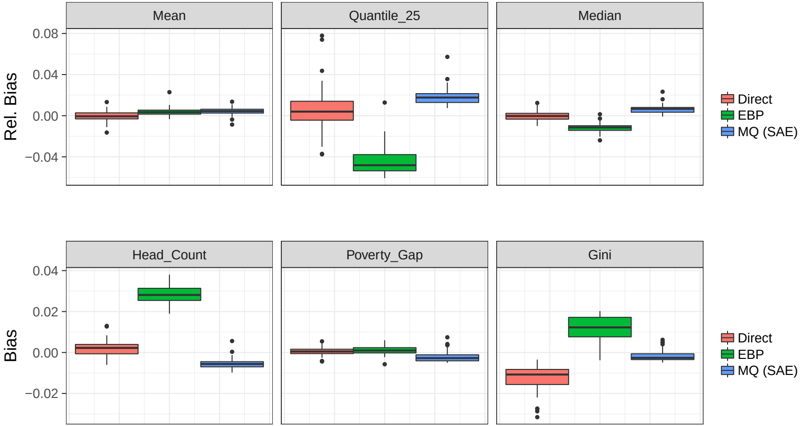

Figure 13: Rel. Bias of Point-Estimation Results for EBP vs. Direct

Cont. Normal Scenario: RMSE

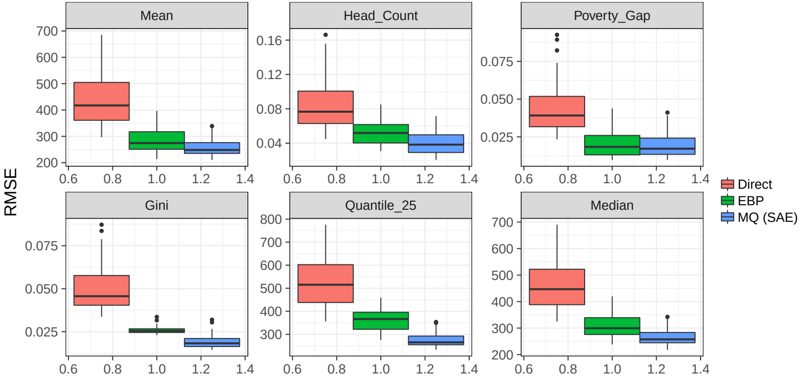

Figure 14: RMSE of Point-Estimation Results for EBP vs. MQ vs. Direct

Pareto Scenario: Setup

Scenario

Pareto Errors

$$
\begin{aligned}
y_{ij} &= 12000 - 400 * x_{ij} + v_j + e_{ij} \\
x_{ij} &\sim N(\mu_j, 7.5) \\
e_{ij} &\sim \sqrt{2} * Pareto(scale = 2000, shape = 3) \\
v_j &\sim N(0, 500^2) \\
\mu_j &\sim U[-3, 3]
\end{aligned}
$$

Note: e is centered to expectation 0

Pareto Scenario: Relative Bias

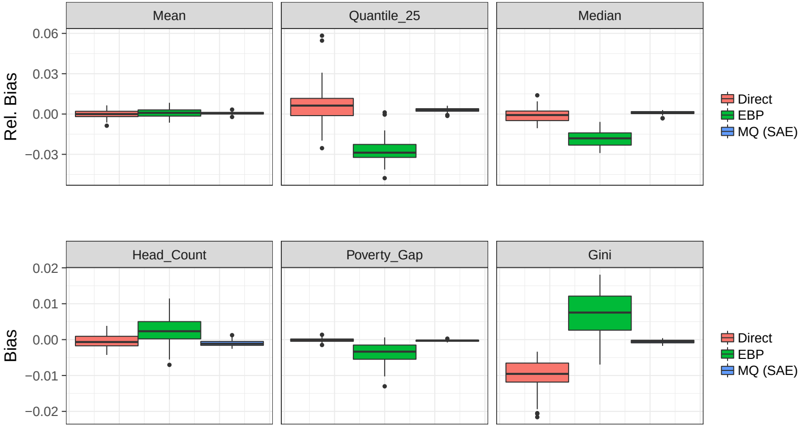

Figure 15: Rel. Bias of Point-Estimation Results for EBP vs. Direct

Pareto Scenario: RMSE

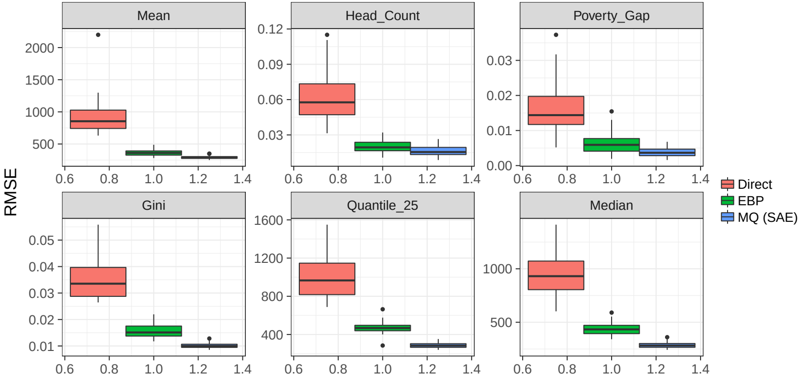

Figure 16: RMSE of Point-Estimation Results for EBP vs. MQ vs. Direct

Simulation Setup for MSE Simulations

Like above, but with

- $N_j = 100$
- 100 runs
- 30 areas
- $B = 50$ Bootstrap Populations
- $S = 50$ Bootstrap Samples (for MQ (SAE))

Normal (MSE) Scenario: Relative Bias of MSE

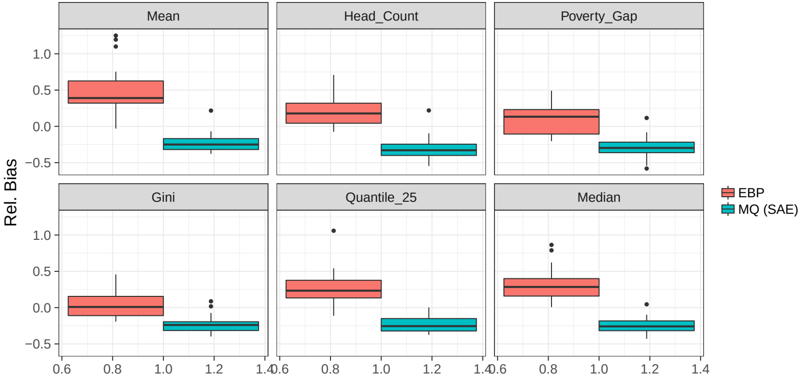

Figure 17: Rel. Bias of MSE-Estimation Results for EBP vs

Normal (MSE) Scenario: RMSE of MSE

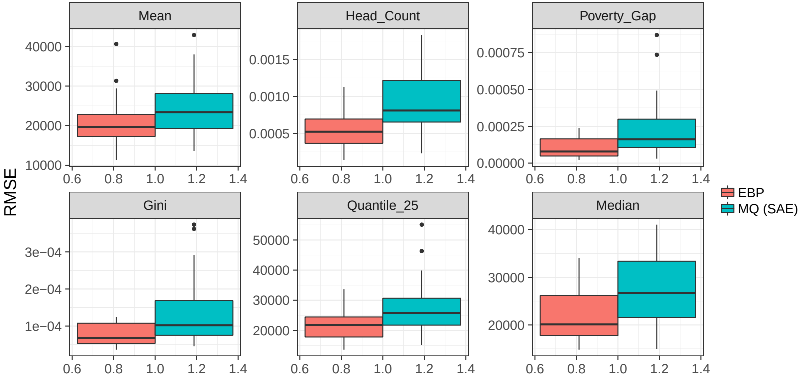

Figure 18: RMSE of MSE-Estimation Results for EBP vs. MQ

Pareto (MSE) Scenario: Rel. Bias of MSE

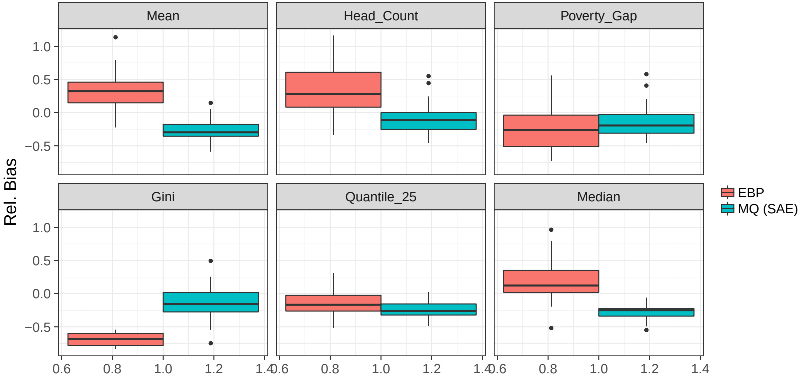

Figure 19: Rel. Bias of MSE-Estimation Results for EBP vs. MQ

Pareto (MSE) Scenario: RMSE of MSE

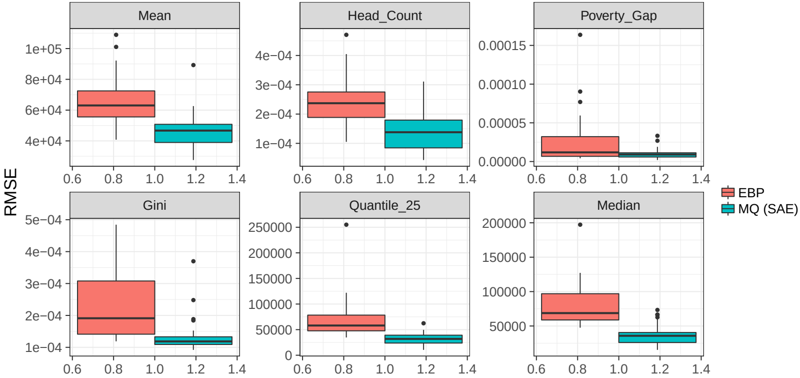

Figure 20: RMSE of MSE-Estimation Results for EBP vs. MQ

Application

- Poverty Mapping for the Austrian EU-SILC dataset
- Data set is provided by the `emdi` package
- 4 numeric variables and 1 factor variable are used to build the models
- Poverty/Inequality measures: Gini, Poverty Gap, Head Count Ratio

Diagnostics

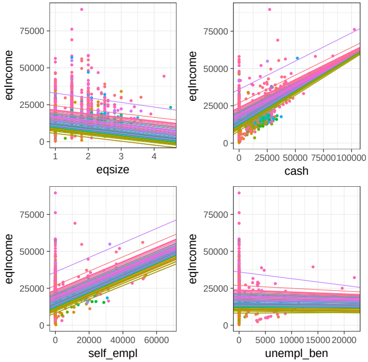

Figure 21: Pseudo Random Effects of the numeric independent variables

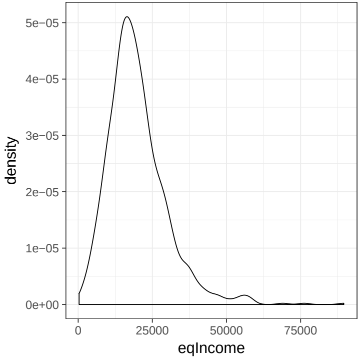

Figure 22: Density Plot of dependent variable

Point Estimation Results

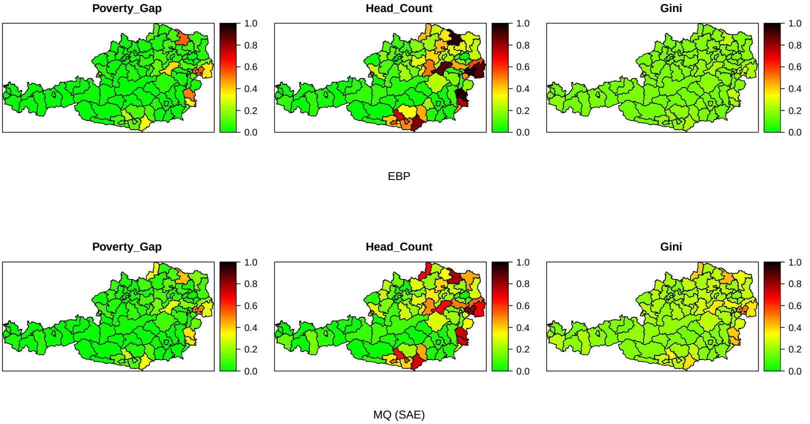

Figure 23: Point Estimation Results for Poverty Mapping in

MSE Estimation Results

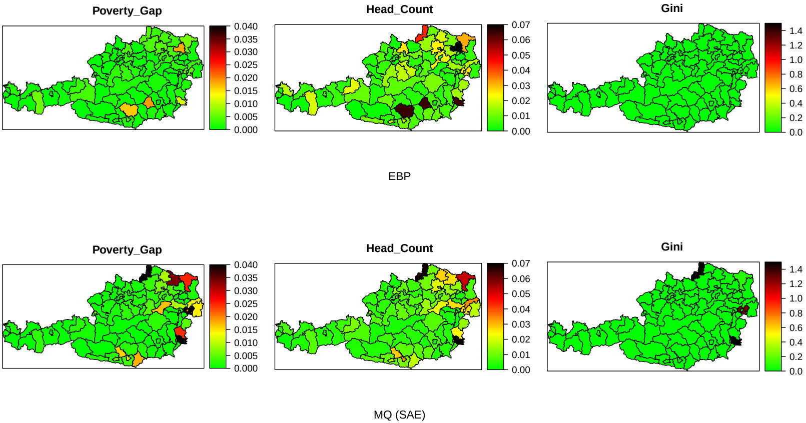

Figure 24: MSE Estimation Results for Poverty Mapping in A

Summary

- The MQ (SAE) can be regarded as beneficial method in some scenarios
  - For normally distributed errors and log-scale outcomes the EBP performs better (Rel. Bias and RMSE)
  - For long-tail distributions the MQ (SAE) approach performs better (Rel. Bias and RMSE)
- Further Research:
  - MQ (SAE) for non-normal ϑ_j
  - Transformations in MQ (SAE)
  - Role of weights in MQ (SAE)

References

- Breckling, J. and R. Chambers (1988): "M-quantiles", Biometrika, 75.
- Chambers, R. and N. Tzavidis (2006): "M-quantile models for small area estimation", Biometrika, 93.
- Huber, P. J. (1963): "Robust Estimation of a Location Parameter", The Annals of Mathematical Statistics, 73.
- Koenker, R., K. Hallock (2001), "Quantile regression", Journal of Economic Perspectives 15, 143?156.
- Marchetti, S., N. Tzavidis, and M. Pratesi (2012a): "Non-parametric bootstrap mean squared error estimation for -quantile estimators of small area averages, quantiles and poverty indicators", Comp. Stat. and Data Analysis, 56.