On the Use of Transformations in Empirical Best Prediction of Domain Parameters

Nikos Tzavidis [1]
Joint work with Timo Schmid, Natalia Rojas & Sören Pannier
(Freie Universität Berlin)

Session C8
CESS Conference
Budapest, 21 October 2016

[1]Southampton Statistical Sciences Research Institute, University of Southampton (n.tzavidis@soton.ac.uk)

Estimation of Deprivation Indicators

- Growing needs of statistics agencies for estimates of deprivation-related indicators at fine spatial scales
- Model-based methods have dominated recent literature
- Until recently less attention to validity of model-based inference in real settings

Examples of Indicators

Income-based indicators

- FGT measures (Foster et al.,1984)

$$FGT(\alpha, t) = \sum_{i=1}^{N} \left(\frac{t-y_i}{t}\right)^{\alpha} \mathbb{1}(y_i \leq t)$$

$\alpha = 0$ - Head Count Ratio; $\alpha = 1$ - Poverty Gap

- The Gini coefficient

$$Gini = \frac{N+1}{N} - \frac{2\sum_{i=1}^{N}(N+1-i)y_{(i)}}{N\sum_{i=1}^{N} y_{(i)}}$$

- Quintile Share Ratio

$$QSR_{80/20} = \frac{\sum_{i=1}^{N}[y_i \mathbb{1}(y_i > q_{0.8})]}{\sum_{i=1}^{N}[y_i \mathbb{1}(y_i \leq q_{0.2})]}$$

Recent Methodologies

- The World Bank method
  (Elbers et al., 2003, Econometrica)
- The Empirical Best Predictor (EBP) method
  (Molina & Rao, 2010, CJS)
- EBP based on normal mixtures (Elbers & Van der Weidel, 2014; Lahiri and Gershunskaya, 2011)
- Methods based on M-Quantiles
  (Marchetti et al., 2012, CSDA)
- Semi-parametric estimation of the empirical distribution function
  (Tzavidis et al., 2016)

Empirical Best Prediction

- Point of departure: Nested error regression model Battese, Harter & Fuller (1988, JASA)

  Notation: (k =domain, i =individual)

$$y_{ik} = \boldsymbol{x}_{ik}^T\boldsymbol{\beta} + \boldsymbol{z}_{ik}^T u_k + \epsilon_{ik}, i = 1, ..., n_k, k = 1, ..., D$$

- Use sample data to estimate β, σ_u, σ_ϵ, u_k
- Generate synthetic values for y_{ik}^* using the predictive density $f(y_r|y_s)$, assumed to be Normal
- Generate $u_k^* \sim \ N(0, \hat{\sigma}_u^2 * (1 - \gamma_k))$ & $\epsilon_{ik}^* \sim \ N(0, \hat{\sigma}_\epsilon^2)$

Micro-simulating a synthetic population:

- Generate a synthetic population under the model a large number of times each time estimating the target parameter
- Linear and non-linear indicators can be computed

Motivating Alternative Methods

- EBP relies on Gaussian assumptions :
  - ✓ $u_k \overset{iid}{\sim} N(0, \sigma_u^2)$, the random area-specific effects
  - ✓ $\epsilon_{ik} \overset{iid}{\sim} N(0, \sigma_\epsilon^2)$, the unit-level error terms, $u_k \perp \epsilon_{ik}$
- What if these fail?
  - Option 1: EBP formulation under an alternative distribution (Molina, I et al., 2015)
  - Option 2: Explore the use of transformations
  - Option 3: Explore robust alternatives

Using Transformations in SAE of Deprivation Indicators

Specific problems:

- Highly positive unimodal skewed and leptokurtic data
- Extensions of the transformations to the mixed model
- Invertibility on $\mathbb{R}$
- Appropriate for handling with zero and negative values
- Examples of target parameters
  - Poverty gap, head count ratio
  - Gini coefficient, quantile share ratio

Possible Transformations

- Shifted transformations
  - Log-shift
- Power transformations
  - Box-Cox
  - Exponential
  - Sign power
  - Modulus
  - Dual power
  - Folder power
  - Convex-to-concave
- Multi-parameter transformations
  - Johnson
  - Sinh-arcsinh

Scaled Transformations

Scaled Log-Shift Transformation (λ)

$$T_\lambda(y_{ij}) = \alpha \log(y_{ij} + \lambda),$$

Scaled Box-Cox Transformation (λ) - (Gurka, 2006 JRSS A)

$$T_\lambda(y_{ij}) = \begin{cases} \frac{(y_{ij}+s)^\lambda - 1}{\alpha^{\lambda-1}\lambda}, & \lambda \neq 0 \\ \alpha \log(y_{ij} + s), & \lambda = 0 \end{cases},$$

Scaled Dual Power Transformation (λ)

$$T_\lambda(y_{ij}) = \begin{cases} \frac{2}{\alpha} \frac{(y_{ij}+s)^\lambda - (y_{ij}+s)^{-\lambda}}{2\lambda} & \text{if } \lambda > 0; \\ \alpha \log(y_{ij} + s) & \text{if } \lambda = 0. \end{cases}$$

with α chosen in such a way that the Jacobian of the transformation is 1

Estimation Method (λ)

Residual Maximum Likelihood (REML)

- Using a scaled version of the transformation
- This allows for applying standard maximum likelihood theory

$$
\begin{aligned}
L_{\text{REML}}(T_\lambda, \lambda|\boldsymbol{\theta}) &= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{D}\log|\boldsymbol{V}_i| \\
&- \frac{1}{2}\log\left|\sum_{i=1}^{D}\boldsymbol{X_i}'\boldsymbol{V}_i^{-1}\boldsymbol{X_i}'\right| \\
&- \frac{1}{2}\sum_{i=1}^{D}[T_\lambda(\boldsymbol{y}_i) - \boldsymbol{X_i}\beta]^T\boldsymbol{V}_i^{-1}[T_\lambda(\boldsymbol{y}_i) - \boldsymbol{X_i}\beta]
\end{aligned}
$$

Estimation Algotithm (λ)

REML Algorithm for the EBP Method

1. Choose a transformation
2. Define a parameter interval for λ
3. Set λ to a value inside the interval
4. Maximize the residual log-likelihood function with respect to θ conditional on the fixed λ
5. Repeat 3 and 4 until the 'optimal' $\hat{\lambda}$ is found
6. Apply the EBP method (Montecarlo Approximation)

Design-based simulation: Mexico Case Study

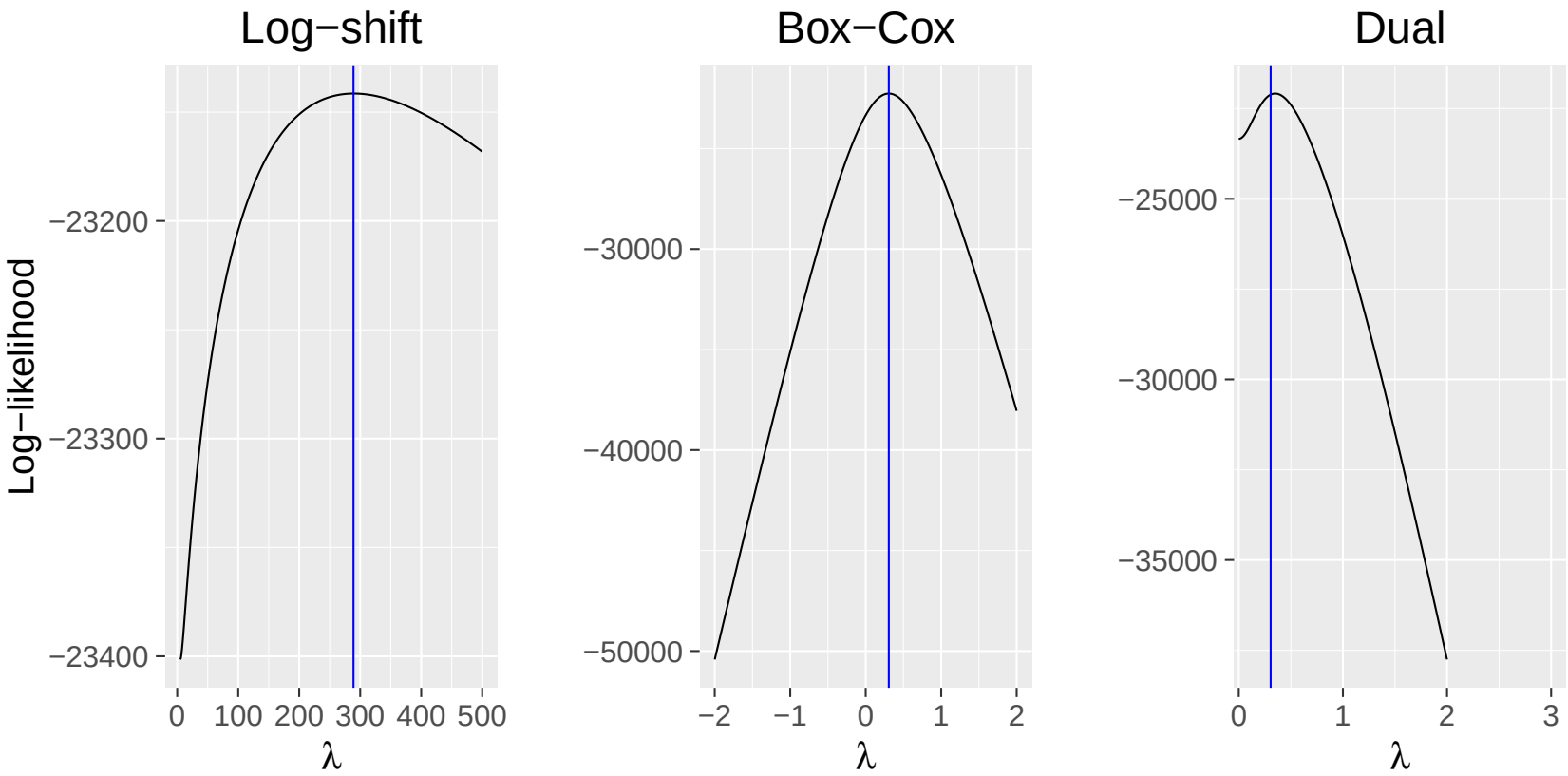
Data

- Data based on the census in EDOMEX
- Outcome is the earned per capita income from work
- Target indicators Gini coefficient & Head Count Ratio

Setup

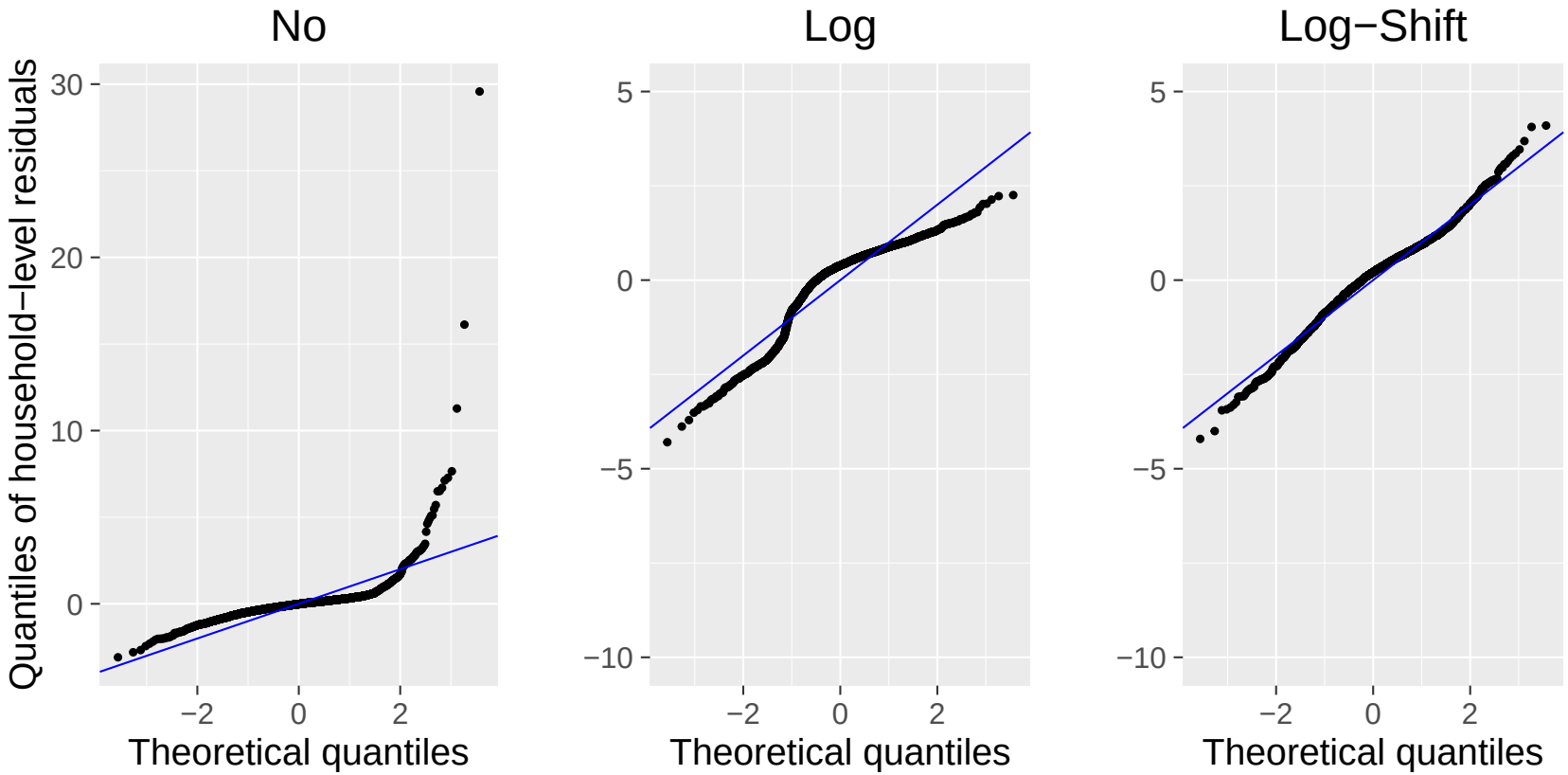
- Design-based simulation with 500 MC-replications repeatedly drawn from EDOMEX Census
- 6 covariates used leading to a R^2 around $40 - 50\%$
- Unbalanced design leading to a sample size of $n = 2195$ ($min = 8$, $mean = 17.6$, $max = 50$)
- Sampling from each municipality

Mexico Case Study: Parameter Estimation



	Log-shift	Box-Cox	Dual
λ	289.46	0.31	0.35

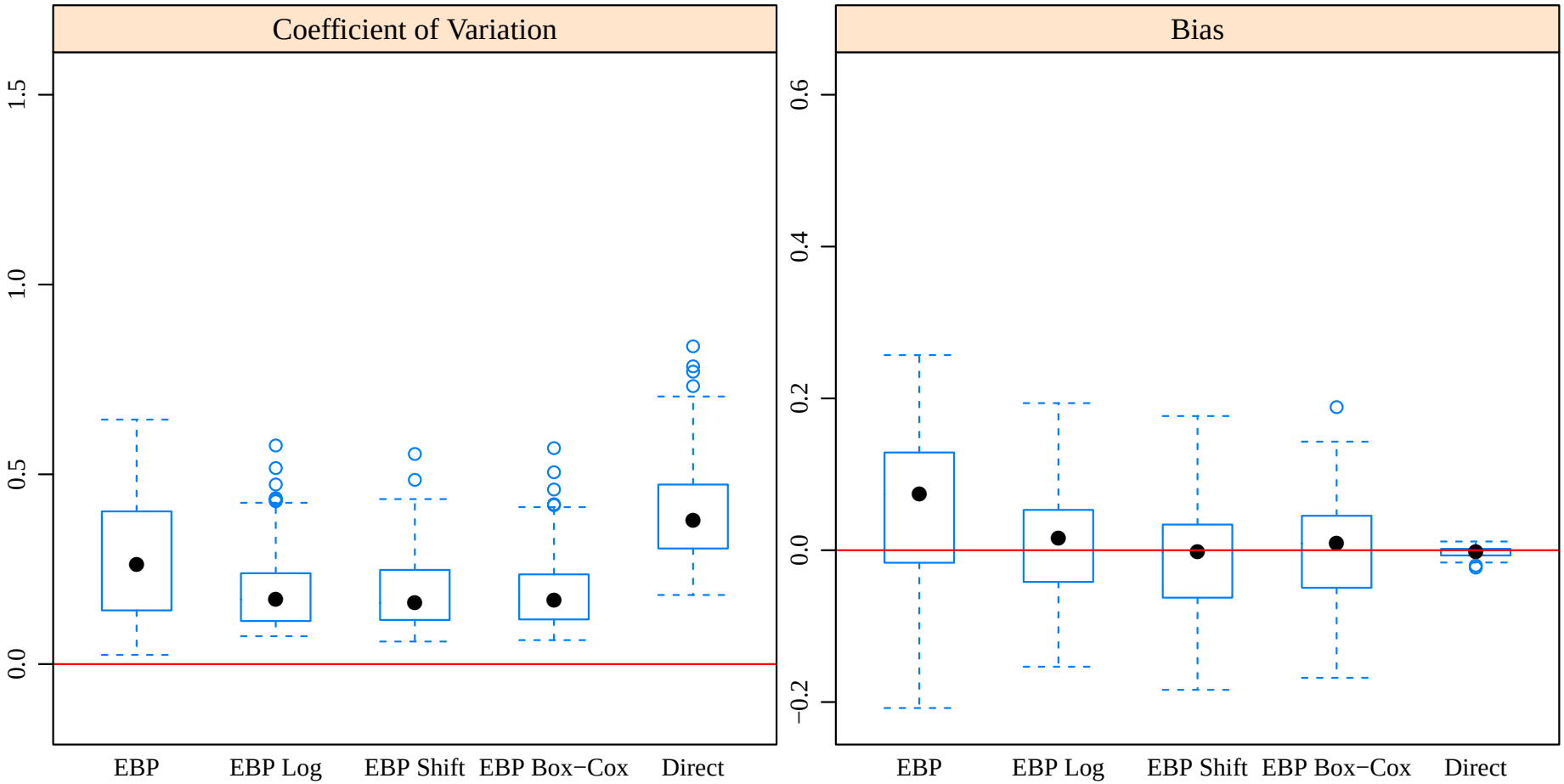
Mexico Case Study: Model and Residual Diagnostics



Transformation	No	Log	Log-Shift	Box-Cox	Dual
R^2	30	40	52	48	48

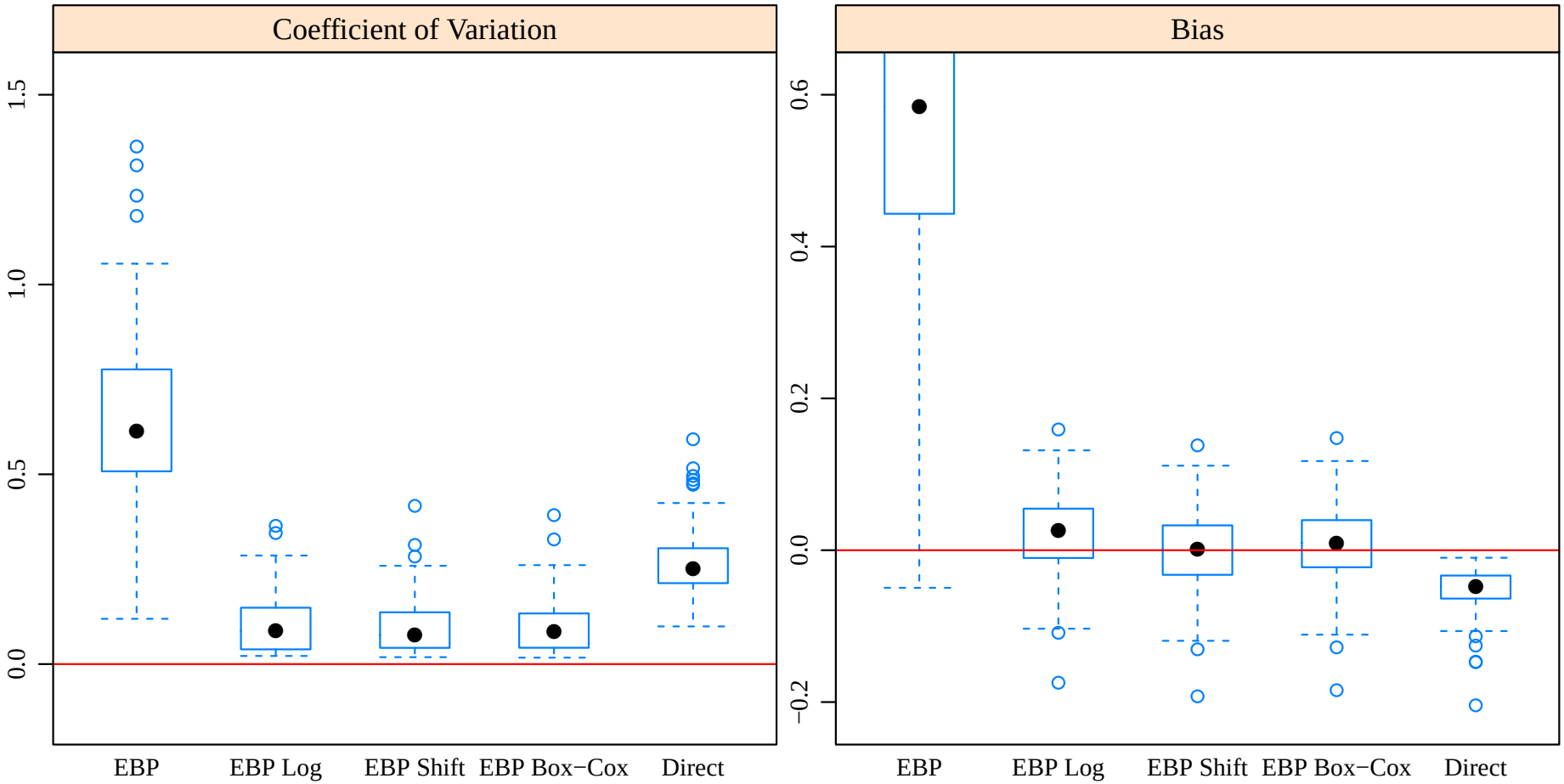
Results - Head Count Ratio

Coefficient of Variation and Bias



Results - Gini Coefficient

Coefficient of Variation and Bias



Remarks and Future Research Directions

Remarks:

- Use of transformations improves the performance of the EBP
- Additionally, it improves and fit of the underling model
- Easy to implement with existing computational tools

Next steps:

- Bootstrap approach for MSE estimation taking into account uncertainty from the transformation parameter estimation
- Use of non-parametric bootstrap for MSE estimation. Avoid impact of small departures from assumed model