Supervised statistical (machine) learning for domain estimation with business survey data

Vasilis Chasiotis[1], Nikos Tzavidis[2], Chiara Bocci[3], Paul Smith[2]

[1]Department of Statistics, AUEB, GR
[2]Department of Social Statistics and Demography, and Southampton Statistical Sciences Research Institute, University of Southampton, UK
[3]Department of Statistics, Computer Science, Applications "G. Parenti", University of Florence, IT

12th International Conference:
The Use of R in Official Statistics (uRos2024)

Piraeus, November 27-29, 2024

Introduction

- Small area typically denotes any domain for which the specific sample is not large enough to support precise direct estimates.
- Small Area Estimation (SAE) is a series of methods to estimate indicators of small domains with applications in many different types of surveys.
- Data requirements:
  - Survey data: available for the target variable Y and for the auxiliary variable X, related to Y.
  - Census/Administrative data: available for X but not for Y.
- Challenges in business surveys:
  - Likely to include outliers.
  - Skewness and variability of variables.
  - Estimates of the quality of point estimates (MSE) for proper inference.

Research questions

- What is the role of random effects in machine learning?
- What is the role of data transformations?
- Can machine learning improve SAE compared to linear models?
- Can machine learning algorithms offer protection under misspecification of linear-type models?

Mixed Effects Random Forest (MERF)

Notation

- Sample-size n; areas $i = 1, ..., D$; individuals j; covariates $\mathbf{x}_{ij}$.
- We assume that y_{ij} follows the following general model:

$$y_{ij} = f(\mathbf{x}_{ij}) + v_i + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma^2_\epsilon) \quad \text{and} \quad v_i \sim N(0, \sigma^2_v)$$

- $f(\mathbf{x}_{ij})$ is a random forest.
- Fitting algorithm similar to EM algorithm.

Krennmair, P. & Schmid, T. (2023). Flexible domain prediction using mixed effects random forests. *Journal of the Royal Statistical Society Series C*, 71(5):1865-1894.

Krennmair, P., Tzavidis, N., Schmid, T. & Wurz, N. (2024). On the use of random forests and mixed effects random forests for small area estimation of general parameters (working paper).

Fitting algorithm ("naive" version)

1. Set $l = 0$ and the random effects $\hat{v}_{(0)}$ to zero.
2. Update $f(\mathbf{x}_{ij})_{(l)}$.
   - 2.1 Set $l = l + 1$.
   - 2.2 $\tilde{y}_{ij,(l)} = y_{ij} - \hat{v}_{i,(l-1)}$
   - 2.3 Estimate $\hat{f}(\mathbf{x}_{ij})_{(l)}$ using a random forest with dependent variable $\tilde{y}_{ij,(l)}$ and covariates $\mathbf{x}_{ij}$.
   - 2.4 Compute OOB predictions, $\hat{f}(\mathbf{x}_{ij})_{(l)}$.
   - 2.5 Compute OOB residuals $r_{ij} = y_{ij,(l)} - \hat{f}(\mathbf{x}_{ij})_{(l)}$.
3. A naive estimator of the residual variance is given by

$$\hat{\sigma}^2_{Naive} = n^{-1} \sum_i \sum_j \left(y_{ij} - \hat{f}(\mathbf{x}_{ij})_{(l)} \right)^2 .$$

4. $\hat{\sigma}^2_{Naive}$ can be decomposed into different sources under the assumed model.
5. Update variance components and v by fitting an empty random intercepts model with r_{ij} as the outcome.
6. Repeat steps 2-5 until convergence.

Fitting algorithm (modified version)

1. $\hat{\sigma}^2_{Naive}$ overestimates the residual variance (Mendez & Lohr, 2011).
2. Implement step 2 as before.
3. Correct the bias of $\hat{\sigma}^2_{Naive}$ by using a residual-based non-parametric bootstrap bias correction. Compute the bias correction term $\widehat{K}$. Compute $\hat{\sigma}^2_{bc} = \hat{\sigma}^2_{Naive} - \widehat{K}$.
4. Compute rescaled residuals $r^{rs}_{ij} = \frac{r_{ij}}{\hat{\sigma}_{Naive}}\hat{\sigma}_{bc}$ to match the corrected variance.
5. Update variance components and v by estimating an empty random intercepts model with r^{rs}_{ij} as the outcome.
6. Repeat steps 3-6 until convergence.

Mendez, G., & Lohr, S. (2011). Estimating residual variance in random forest regression. *Computational Statistics and Data Analysis*, 55:2937-2950.

Application

- Retail businesses in Italy (excluding petrol stations).
- Population size $N = 71,568$.
- `Y`: revenue of 2020.
- `X`: revenue of 2018.
- `sc`: size class, based on the working persons (1, 2-4, 5-9, 10-19, 20-49).
- `wp`: the number of working persons.
- `ind`: industrial classification (36 industry groups).
- Stratified design with Neyman allocation.
- Sample size $n = 5,000$.
- Sample sizes per industry group varying from 12 to 905.

Results - Mean revenue of 2020

$$\text{EBLUP: } \mathtt{Y} = \mathtt{X} + \mathtt{wp} + \mathtt{wp} \times \mathtt{X} \quad [\text{var} = f(\mathtt{wp})]$$

$$\text{EBP: } \log(\mathtt{Y}) = \log(\mathtt{X}) + \mathtt{wp} + \mathtt{sc} + \mathtt{wp} \times \log(\mathtt{X})$$

$$\text{MERF: } \log(\mathtt{Y}) = \mathtt{X} + \mathtt{wp} + \mathtt{sc}$$

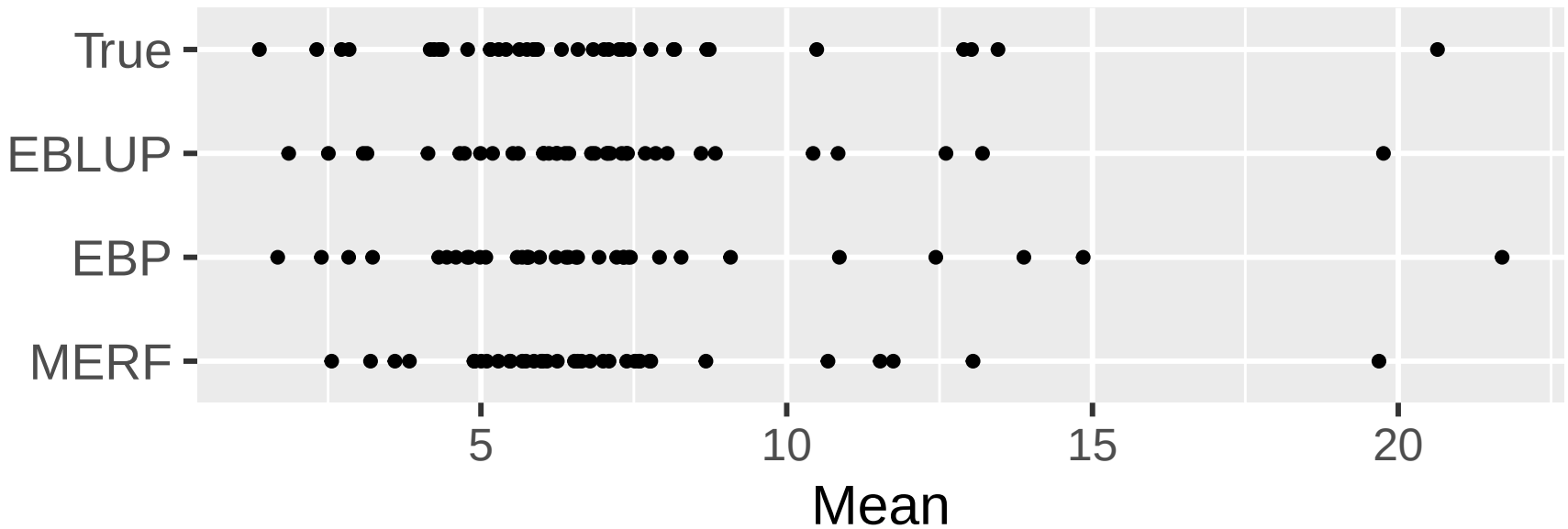

Figure: The true values and the values of estimators by industry group.

Results - Absolute bias & MSE

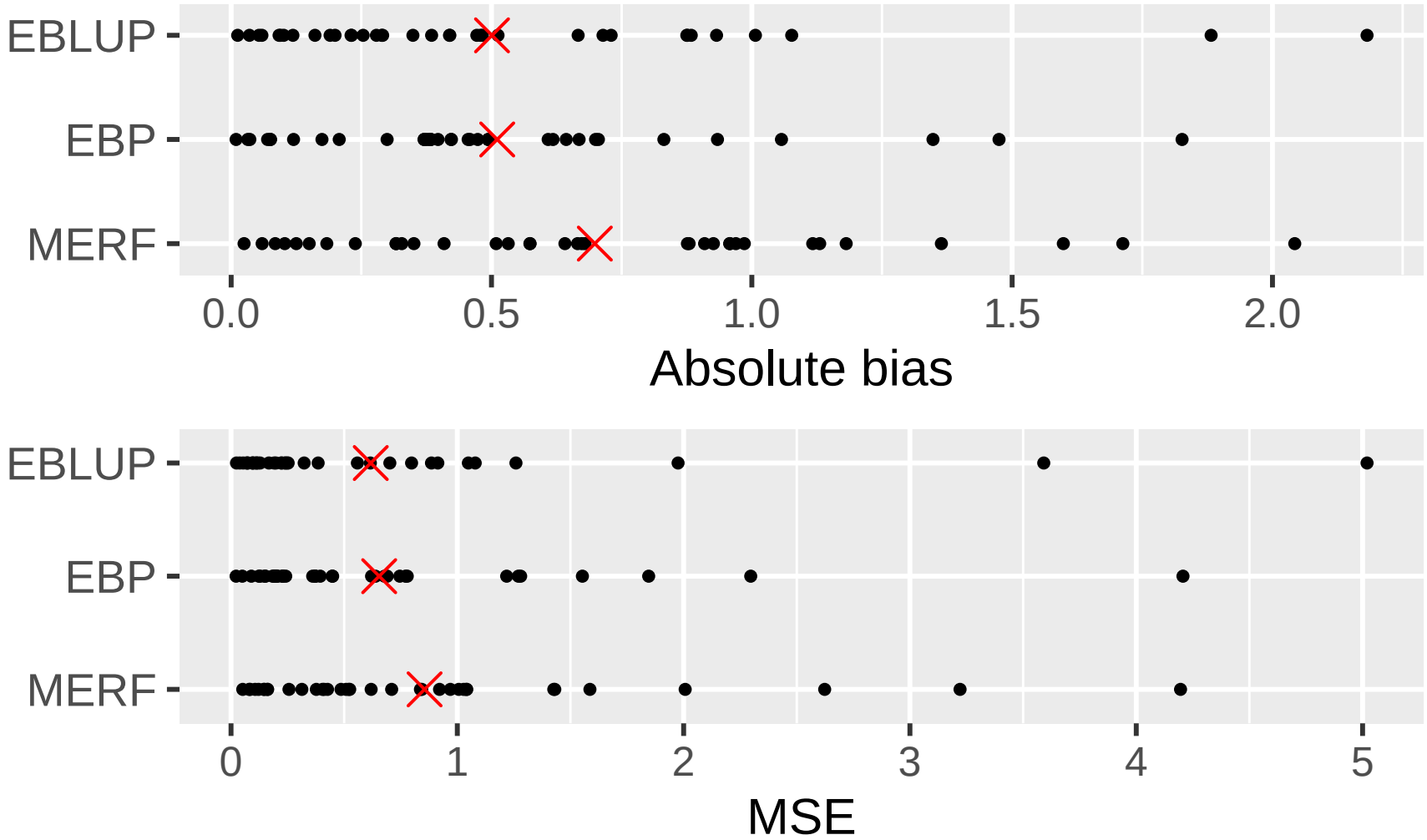

Figure: Absolute bias and MSE by industry group; the cross shows the mean values.

Uncertainty estimation

1. Compute marginal OOB residuals $r_{ij} = y_{ij} - \hat{f}(\mathbf{x}_{ij})$.
2. Compute level 2 residuals scaled to estimated level 2 variance

$$\bar{r}_i^{c(2)} = \frac{1}{n_i} \sum_{j=1}^{n_i} r_{ij}$$

3. Compute level 1 residuals $r_{ij}^{c(1)} = r_{ij} - \bar{r}_i^{c(2)}$, scaled to level 1 estimated variance.
4. For $b = 1, ..., B$:
   - 4.1 Sample from the scaled and centred level 1 and level 2 residuals:

$$r_{ij}^{(b)} = \text{srswr}(r_{ij}^{c(1)}, N) \quad \text{and} \quad r_i^{(b)} = \text{srswr}(\bar{r}_i^{c(2)}, D).$$

   - 4.2 Construct the bootstrap population under the MERF $y_{ij}^{(b)} = \hat{f}(\mathbf{x}_{ij}) + r_i^{(b)} + r_{ij}^{(b)}$.
   - 4.3 Compute the bootstrap population parameter of interest $\theta_i^{(b)}$.
   - 4.4 From each bootstrap population, draw a bootstrap sample, obtain $\hat{\theta}_i^{(b)}$.
5. Compute MSE estimator

$$\widehat{MSE}_i = B^{-1} \sum_{b=1}^{B} \left(\hat{\theta}_i^{(b)} - \theta_i^{(b)} \right)^2.$$

Results - MSE for EBP

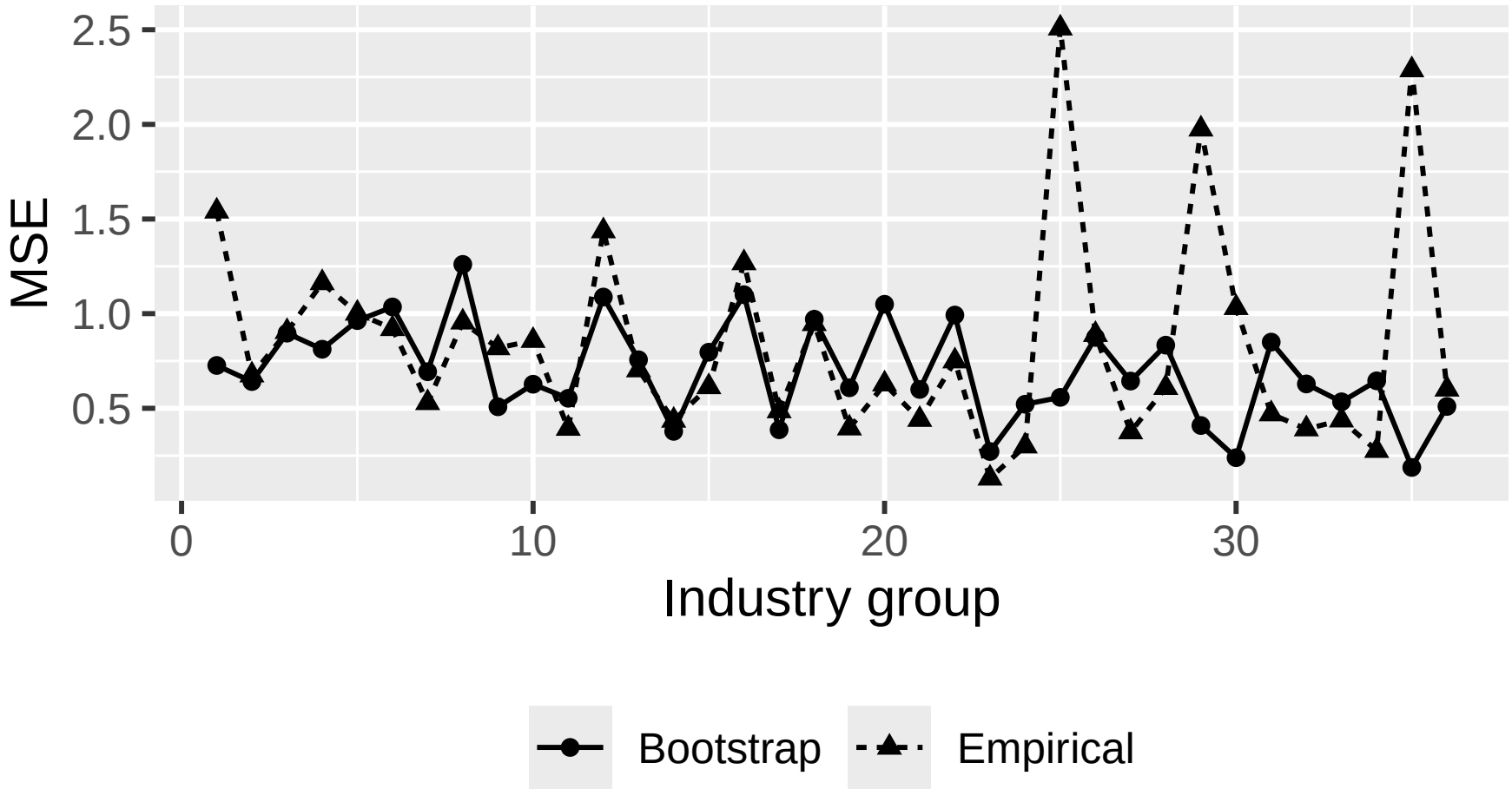

Figure: Bootstrap MSE and empirical MSE for EBP with 100 Monte Carlo simulations and 100 bootstrap samples.

Results - MSE for MERF

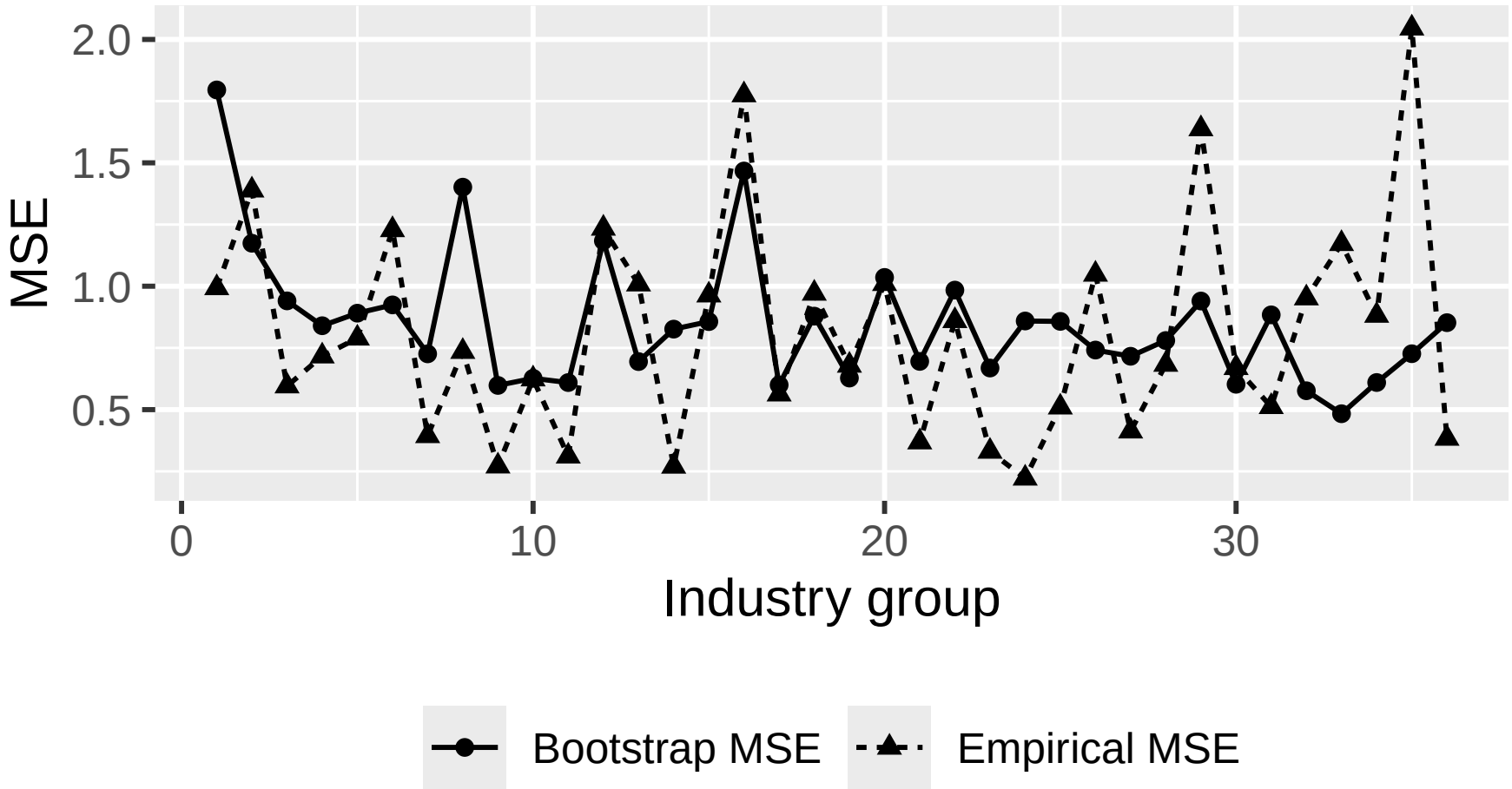

Figure: Bootstrap MSE and empirical MSE for MERF with 100 Monte Carlo simulations and 100 bootstrap samples.

Discussion

- Random effects are central to SAE and play an important role in machine learning.
- The usefullness of data transformations.
- MERFs are competitive compared to linear models, offering protection under misspecification.

- Alternatives to random effects specification?
- Consider alternative estimation strategies for random effects, more in line with algorithmic culture (Breiman, 2001).

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3):199-231.

Thank you for your attention.