Insights into Small Area Estimation using the Nested Error Regression Model

Ziyang Lyu[1] & Alan Welsh[2]

[1]School of Mathematics and Statistics
University of New South Wales

[2] Research School of Finance, Actuarial Studies and Statistics, Australian National University

Consumer expenditure on fresh milk products

Population: Dairy Survey component of the 2002 Consumer Expenditure Survey conducted by the U.S. Bureau of the Census for the U.S. Bureau of Labor Statistics; *https://www.bls.gov/cex/pumd_data.htm.*

Data cleaning: Discard 6 states with ≤ 10 observations, leaving $N = 4022$ observations from $g = 34$ states with $36 \leq N_i \leq 397$ observations per state.

Survey variable: Household expenditure on fresh milk products.

Auxiliary variables: Total expenditure on food (FOODTOT), the number of persons under age 18 in the household (PERSLT18) and the total household income before taxes in the last 12 months (FINCBEFX).

Purpose: From a (noninformatively selected) sample of $20 \leq n_i \leq 96$ households from each state, estimate $\bar{y}_i$, the average household expenditure on fresh milk products in each state (small area).

Estimation of average consumer expenditure

Direct estimation: if n_i units are sampled from state i (sample proportion $f_i = n_i/N_i$), $\bar{y}_{i(s)}$ and s_i^2 are the sample mean and sample variance in state i, respectively, an approximate $100(1-\alpha)\%$ CI for $\bar{y}_i$ is

$$[\bar{y}_{i(s)} - 1.96\{(1-f_i)n_i^{-1}s_i^2\}^{1/2},\ \bar{y}_{i(s)} + 1.96\{(1-f_i)n_i^{-1}s_i^2\}^{1/2}].$$

Mixed model estimation: Fit a working linear mixed model

$$y_{ij} = \mathbf{x}_{ij}^T\boldsymbol{\beta} + \sigma_\alpha\alpha_i + \sigma_e e_{ij},$$

where $\{\alpha_i\}$ and $\{e_{ij}\}$ are independent with zero means and unit variances, to the sample data; compute the EBLUPs

$$\hat{M}_i = f_i\bar{y}_{i(s)} + (1-f_i)(\bar{\mathbf{x}}_{i(r)}^T\hat{\boldsymbol{\beta}} + \hat{\alpha}_i),\ \text{where } \hat{\alpha}_i = (\hat{\sigma}_e^2 + n_i\hat{\sigma}_\alpha^2)^{-1}n_i\hat{\sigma}_\alpha^2(\bar{y}_{i(s)} - \bar{\mathbf{x}}_{i(s)}^T\hat{\boldsymbol{\beta}}),$$

a measure of variability $\hat{V}_i$, and then an approximate $100(1-\alpha)\%$ PI for $\bar{y}_i$

$$[\hat{M}_i - 1.96\hat{V}_i^{1/2},\ \hat{M}_i + 1.96\hat{V}_i^{1/2}].$$

Comments

Other estimators: Instead of $\hat{M}_i = f_i \bar{y}_{i(s)} + (1 - f_i)(\bar{\mathbf{x}}_{i(r)}^T \hat{\boldsymbol{\beta}} + \hat{\alpha}_i)$, many use

$$\hat{M}_i^{alt} = f_i(\bar{\mathbf{x}}_{i(s)}^T \hat{\boldsymbol{\beta}} + \hat{\alpha}_i) + (1 - f_i)(\bar{\mathbf{x}}_{i(r)}^T \hat{\boldsymbol{\beta}} + \hat{\alpha}_i) = \bar{\mathbf{x}}_i^T \hat{\boldsymbol{\beta}} + \hat{\alpha}_i,$$

which actually estimates $\eta_i = E(\bar{y}_i | \bar{x}_i, \alpha_i) = \bar{x}_i^T \boldsymbol{\beta} + \alpha_i$.

Model-based measures of variability: Allowing $g \to \infty$ with (i) cluster sizes bounded, we have the Rao-Molina (2015) extension of the Prasad-Rao (1990) estimator, or (ii) with increasing cluster size, the Lyu-Welsh (2023) estimator $\hat{V}_i = (1 - f_i) n_i^{-1} \hat{\sigma}_e^2$.

Model flexibility: We centered the auxiliary variables about their state means and then included the state means as between state variables so that we have 6 auxiliary variables.

Model-based and Design-based properties

Model-based: The population values $\{y_{ij}\}$ are realisations from a stochastic model and inference is made under the population model conditioning on the selected sample.

> **Model-based simulation:** Simulate 1000 different populations and draw the same sample from each population.

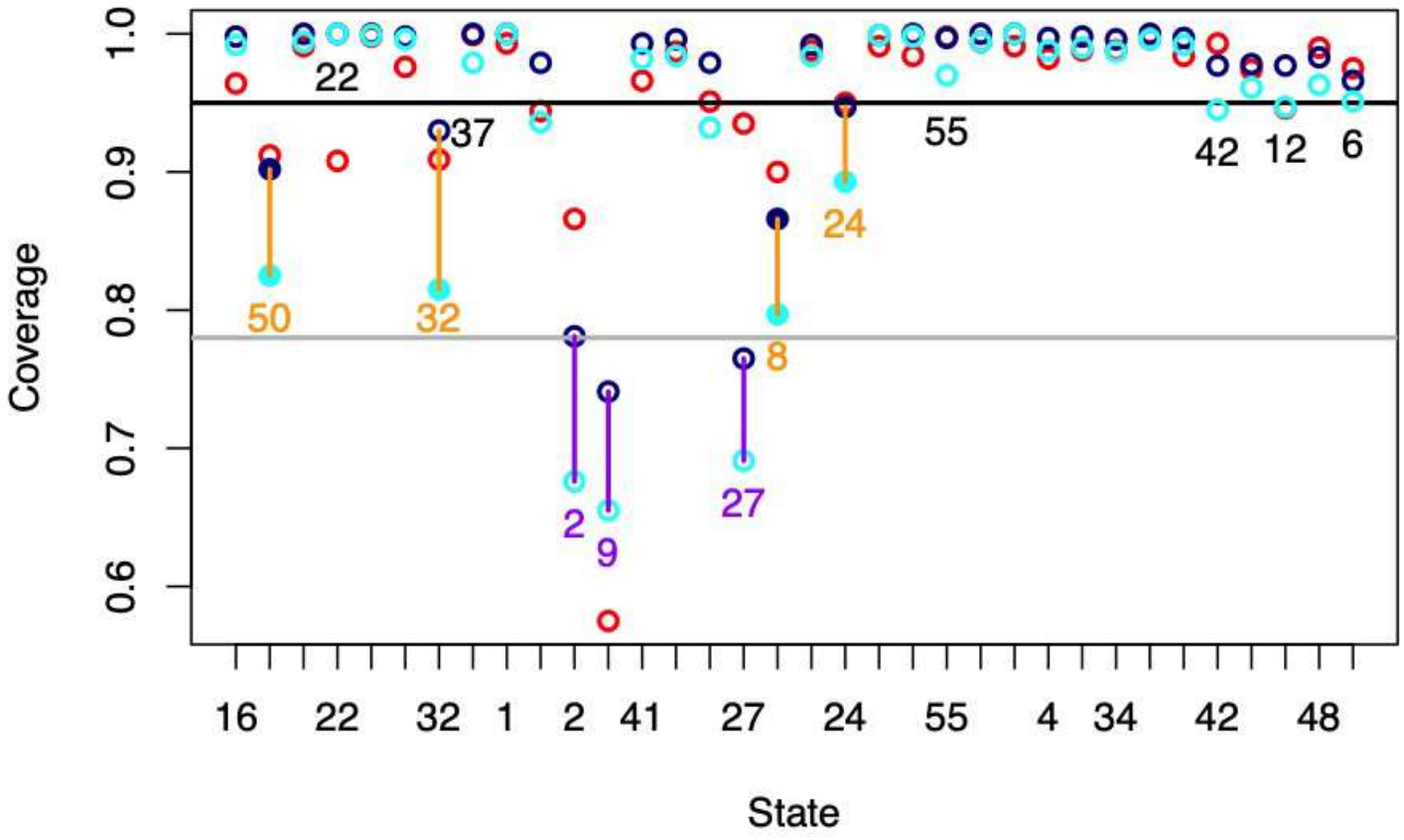
Design-based: Condition on the population values $\{y_{ij}\}$ and make inference under repeated sampling from the fixed population.

> **Design-based simulation:** Simulate a single population and draw 1000 different random samples from the population.

Model-based methods work well in the model-based framework when the assumed model is correct. How do they perform in the design-based framework?

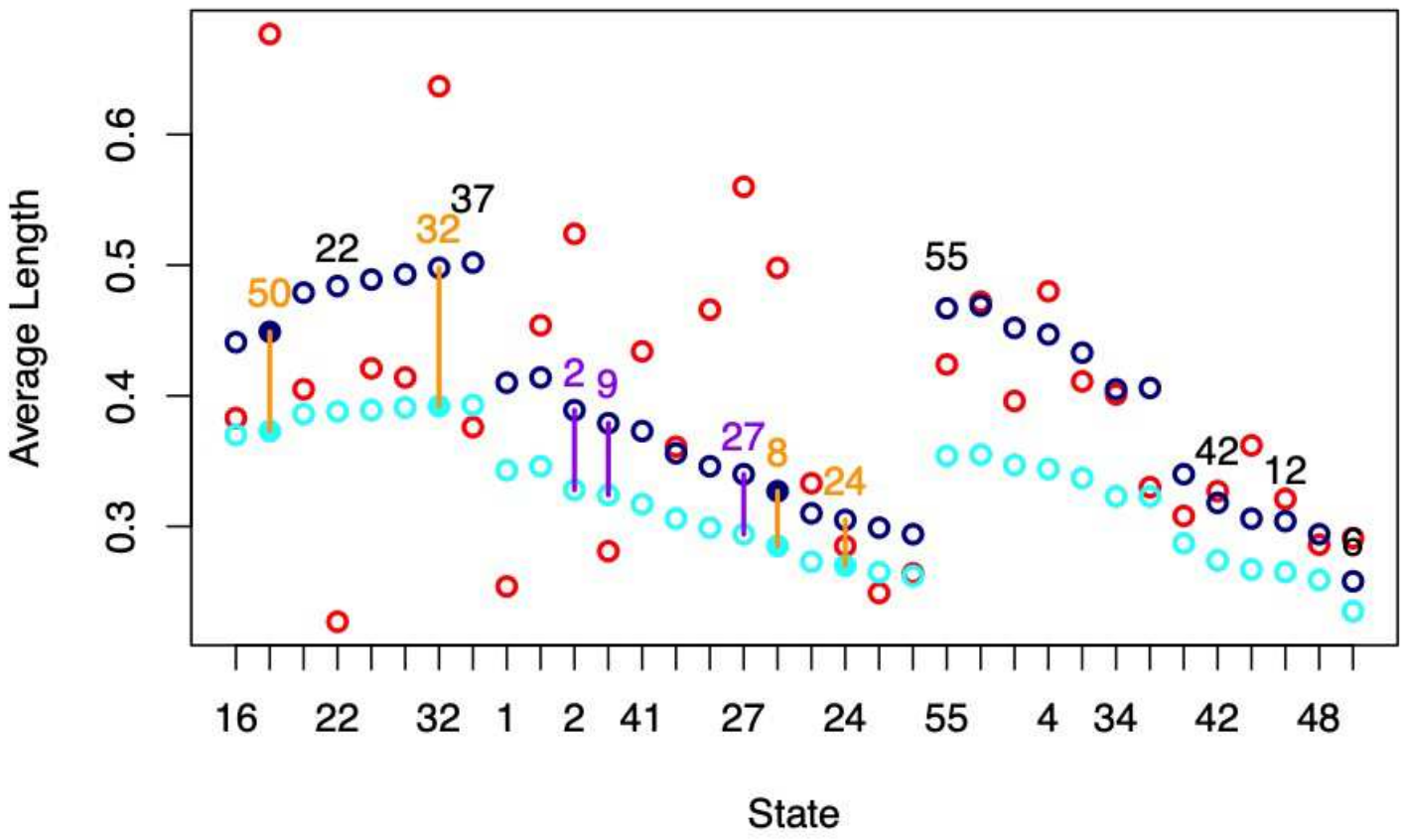
Coverage

red = Direct estimate, blue = Lyu-Welsh, cyan = Rao-Molina



Average length

red = Direct estimate, blue = Lyu-Welsh, cyan = Rao-Molina



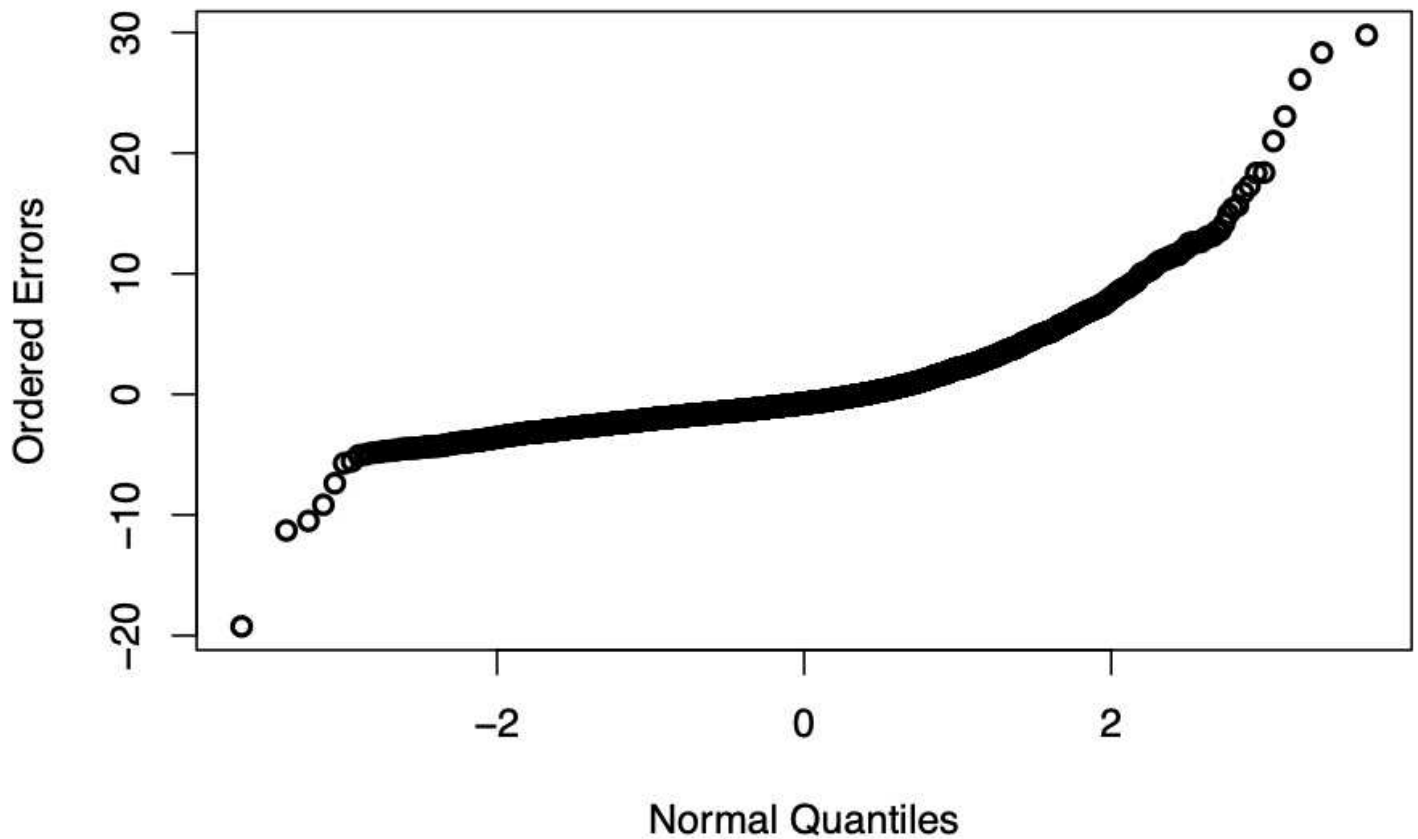
REML parameter estimates for the population model

Effect	Estimate	Std Error
(Intercept)	2.6219	0.7622
FOODTOTavg	1.1617	5.0871
PERSLT18avg	0.9235	0.5526
FINCBEFXavg	0.0155	0.0083
FOODTOTcent	4.7930	0.3656
PERSLT18cent	0.7098	0.0399
FINCBEFXcent	0.0015	0.0010
σ^2_α	0.1757	
σ^2_e	8.5600	

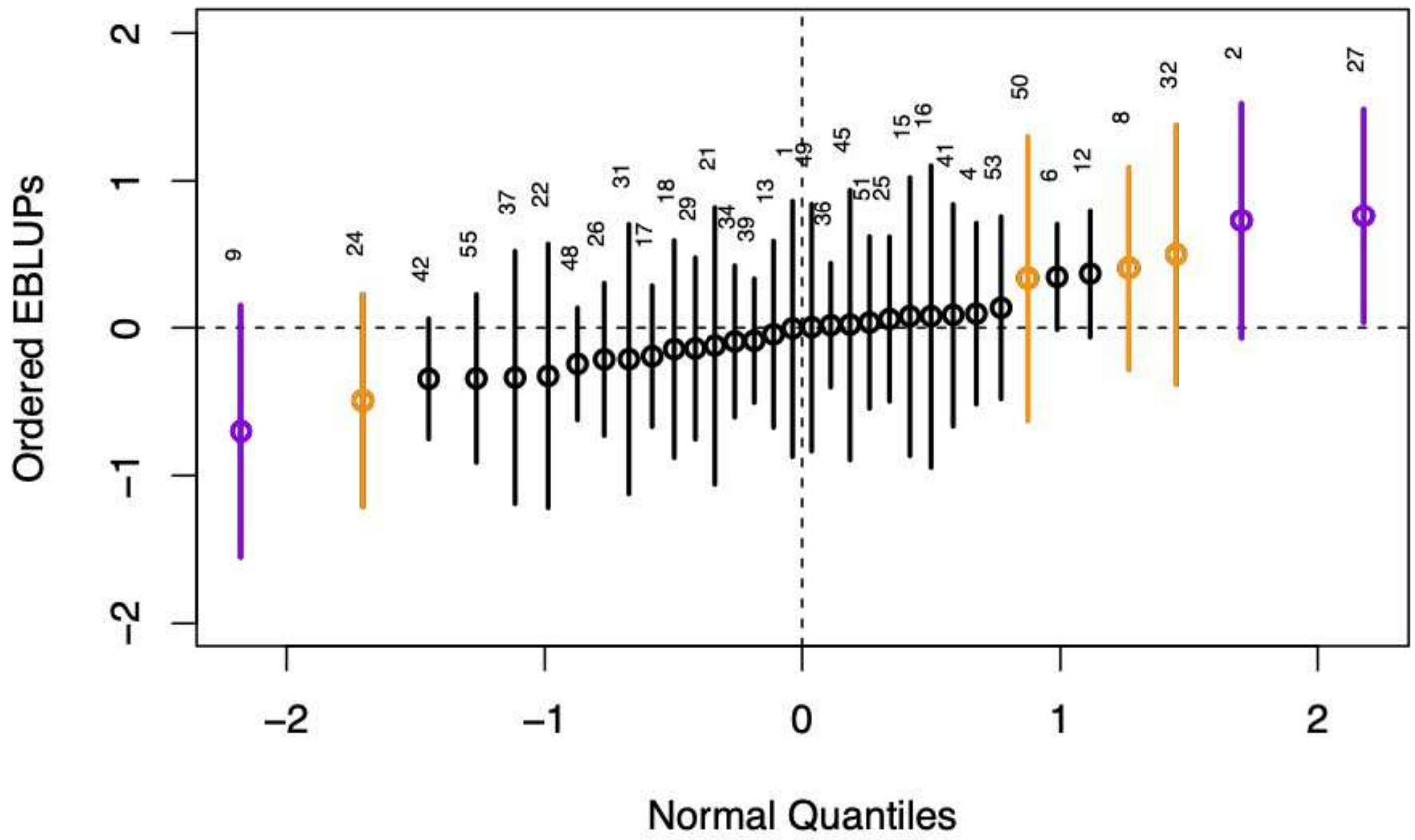
With $\hat{\sigma}^2_e/\hat{\sigma}^2_a = 48.72$, the within cluster correlation is ≈ 0.02.

(`lmer` does not compute standard errors for the variance components.)

QQ-plots

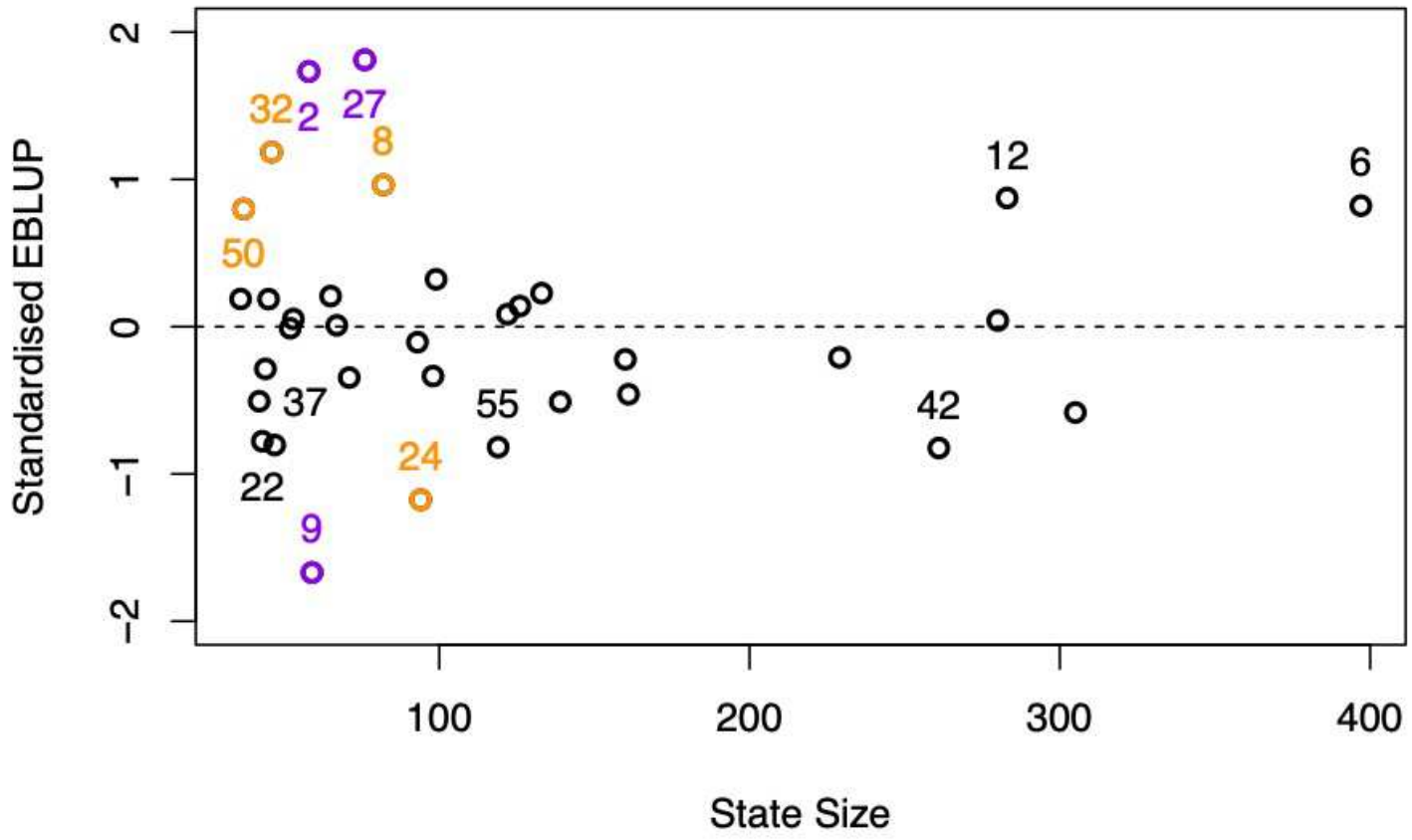


black = Group 1, **orange = Group 2**, **purple = Group 3**



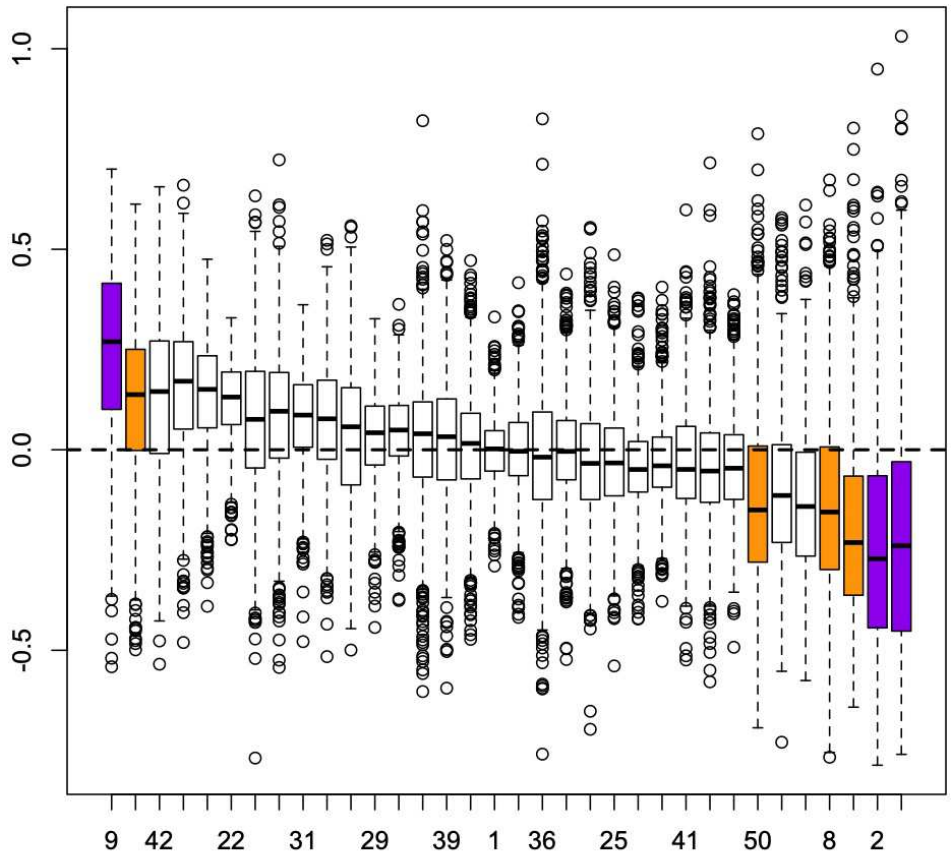
Standardised EBLUPs

black = Group 1, **orange = Group 2**, **purple = Group 3**



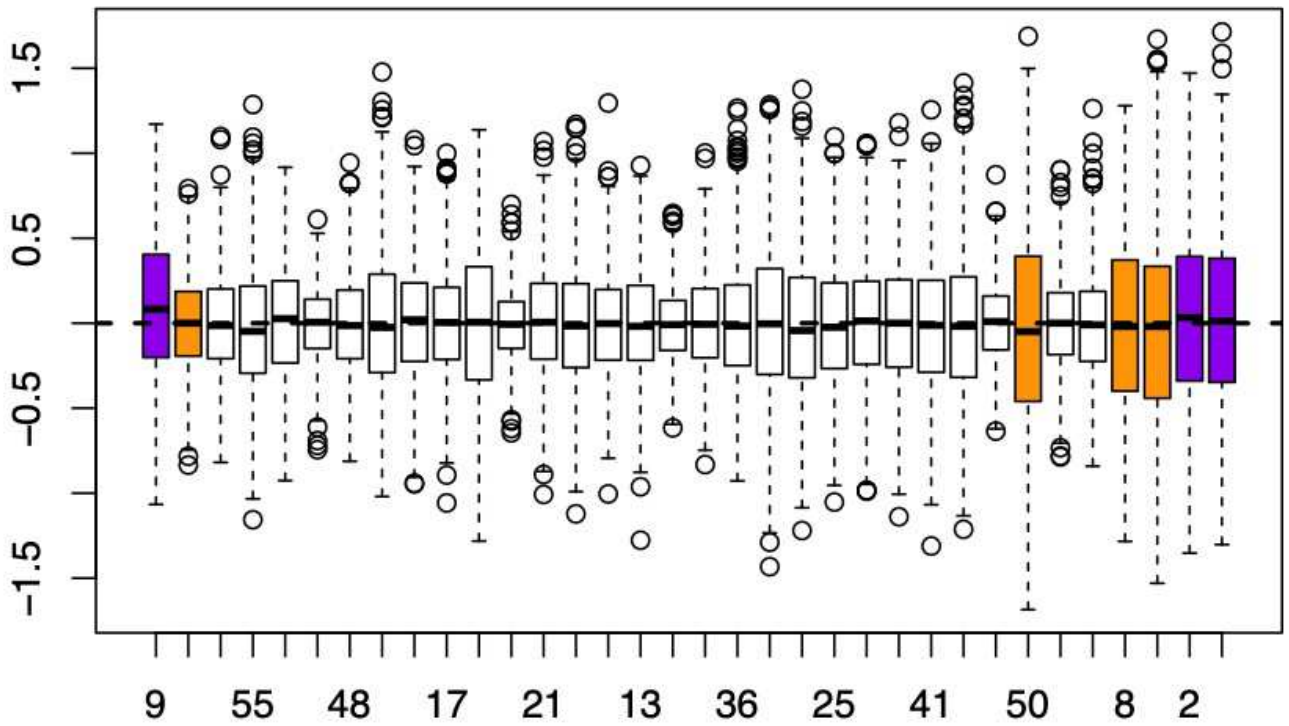
EBLUP Empirical bias

black = Group 1, **orange = Group 2**, **purple = Group 3**



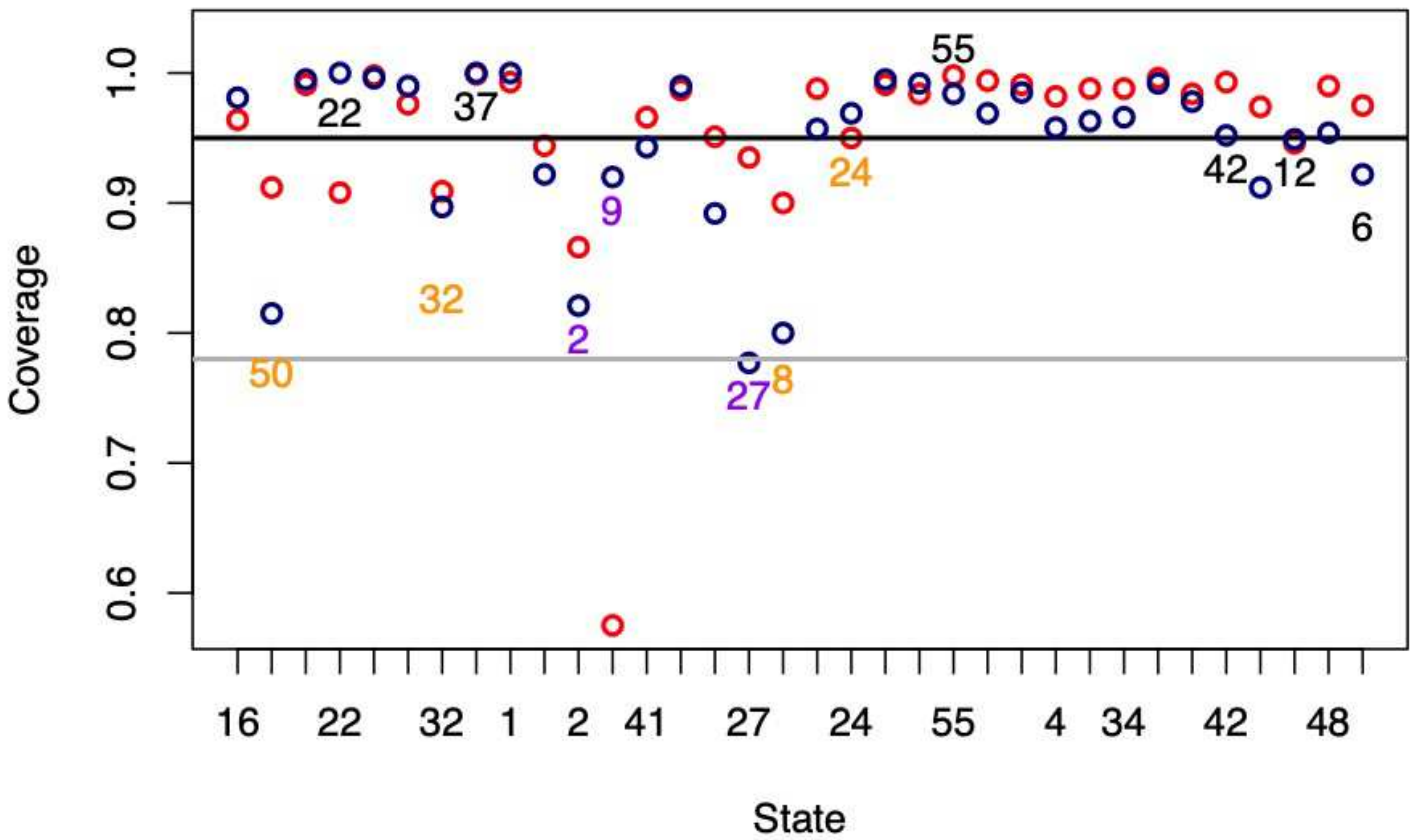
Fixed state effects: Empirical bias

black = Group 1, **orange = Group 2**, **purple = Group 3**



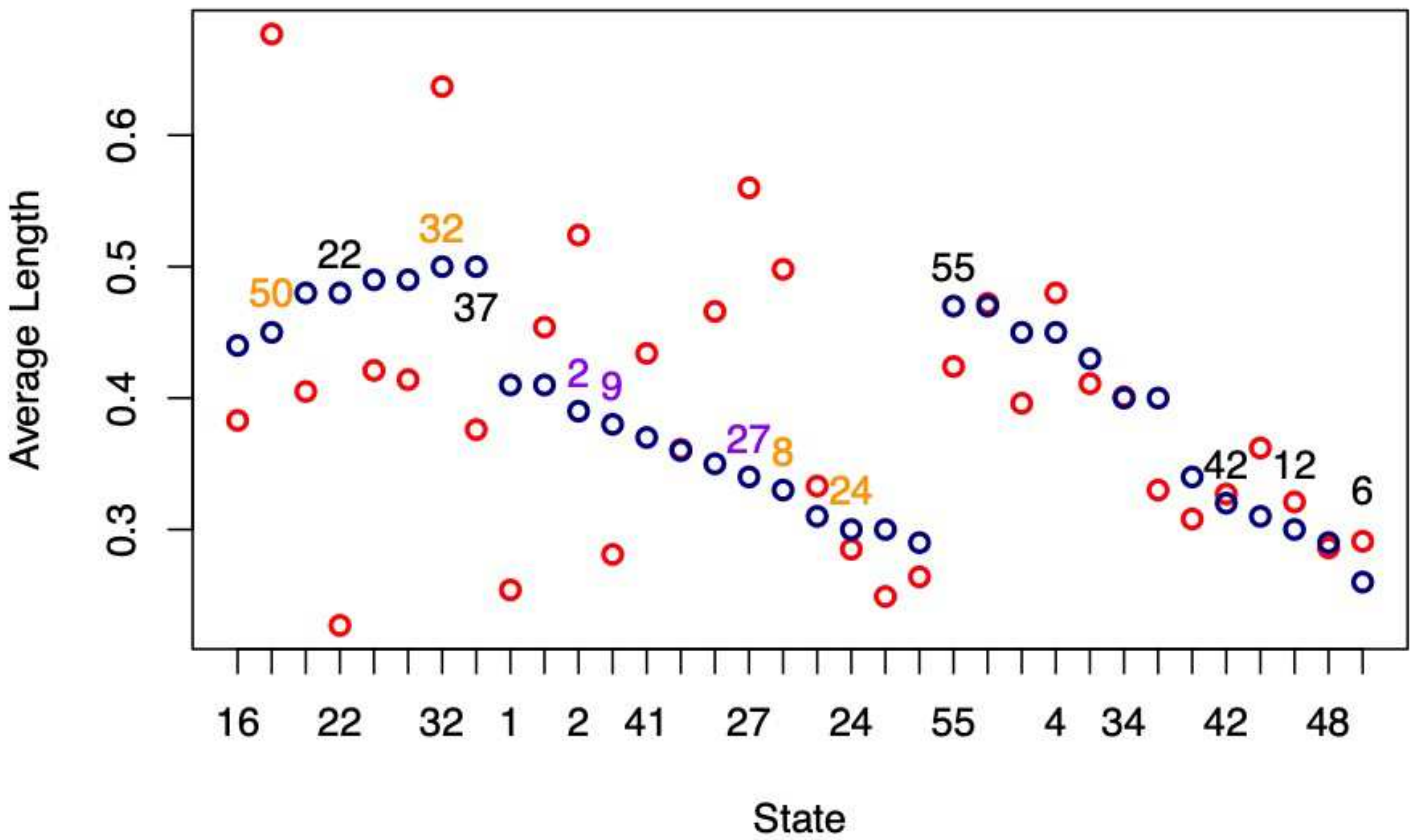
Fixed state effects: Coverage

red = Direct estimate, blue = Model-based regression estimator



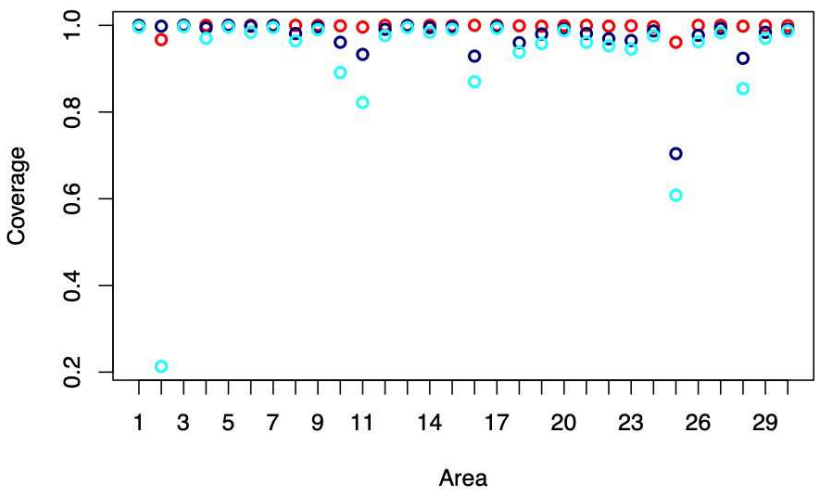
Fixed state effects: Average length

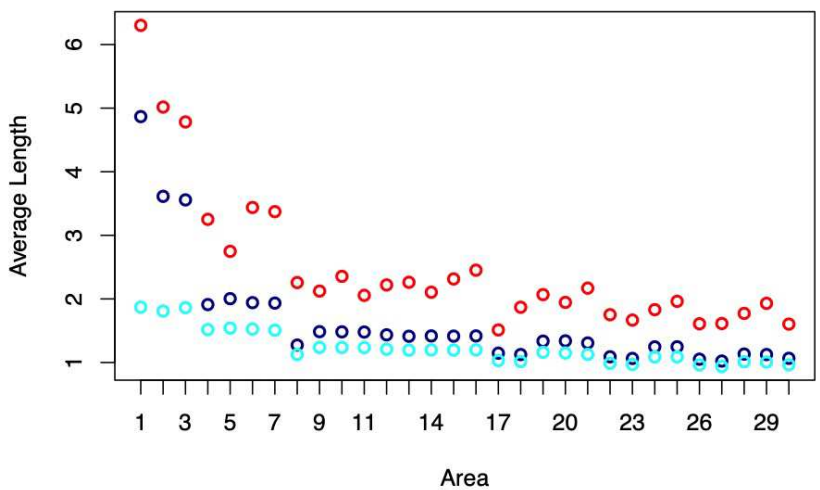
red = Direct estimate, blue = Model-based regression estimator



Design-based simulation: the population model holds

red = Direct estimate, **blue = Lyu-Welsh**, **cyan = Rao-Molina**





Takeaways

- **The mixed model PIs** worked well in the
  - model-based framework when the population model holds.
  - design-based framework, when the population model holds and for the fresh milk products population, where the model probably does not hold, for **Group 1** states with small to moderate effects.

  They do not work as well in the design-based framework for small states with extreme effects (**Groups 2** and **3**).

- Model-based inference treats state effects as random; design-based inference is like conditioning on them and treating them as fixed. **Fitting fixed effects as**
  - **random** introduces bias (through shrinkage) which is larger for more extreme effects in small states.
  - **fixed** removes the bias and improves coverage (but not perfectly because the estimates are more variable and mseps are small).

- For **estimating MSEP**,
  - LW (the number of states and state size diverge) is much simpler than RM (the number of states diverge, state size is bounded).
  - LW is larger than RM.
  - LW and RM are asymptotically equivalent as state size increases.
  - LW and RM are more design-efficient than the direct estimates, even in large states (when survey practitioners argue for just using direct estimates).
- **Non-sample representative outliers are difficult to handle.** In the fresh milk products population, they affected the direct estimate more than the mixed model estimates - perhaps because they are more extreme marginally than conditionally i.e. they are partly explained by the values of their auxiliary variables.

References

Lyu & Welsh (2022a), *JASA*

Lyu & Welsh (2022b), *J. Statist. Plan. Inf.*

Lyu & Welsh (2023), *Statistica Sinica*

Prasad & Rao (1990), *JASA*

Rao & Molina (2015), *Small Area Estimation*