Calibration procedure for estimates obtained from posterior approximation algorithms, with application to domain-level modeling

Julie Gershunskaya [1]

[1]U.S. Bureau of Labor Statistics (OEUS Statistical Methods Staff)

FCSM Conference - October 2023

Outline

Motivation

Resampling based approach

Simulations

Small domain model used in CES

Summary and Recommendations

Motivation

- Current Employment Statistics (CES) survey publishes monthly estimates of employment at **detailed levels by industry and geography**
- In small samples, **direct sample-based estimates can be unstable.** We use Small Area Estimation (SAE) modeling techniques to produce better estimates for small domains
- To formulate SAE models, we choose a Bayesian approach for its **flexibility and the ability to handle complicated models**
- CES produces estimates monthly and has **tight production schedule:** It is essential to use **fast and efficient** model fitting algorithm.

Motivation (cont'd)

- We use Automatic Differentiation Variational Inference (ADVI) algorithm implemented in Stan modeling language. The mean field approximation of the ADVI is relatively **fast and efficient**.
- Small domain model was tested on historical CES series: estimates **showed better performance**, compared to alternative models
- However, it has been reported in the literature (Yao et al. 2018) that ADVI (in particular, the mean field approximation) may produce **inaccurate uncertainty measures** of point estimates
- **The goal of the current research:** to develop a practical tool for the evaluation of the fit and correction of a bias in the posterior variance of the ADVI-based point estimator

Outline

Motivation

Resampling based approach

Simulations

Small domain model used in CES

Summary and Recommendations

Setting the stage: Fay-Herriot Model

Available data for each of $i = 1, \ldots, N$ domains of interest:

- y_i : direct sample-based estimates
- v_i : variances of direct estimates y_i
- $\mathbf{x}_i$: vector of covariates

Assume the following two-level model:

$$y_i \overset{\text{ind}}{\sim} N\left(\theta_i, v_i\right) \qquad \textit{sampling model}$$

$$\theta_i \overset{\text{ind}}{\sim} N\left(\mathbf{x}_i^T \beta, \tau_u^2\right) \qquad \textit{linking model}$$

β and τ_u^2 are unknown model parameters.
Fit the model and obtain θ_i using ADVI algorithm.

Resampling based approach, general outline

Assume, the model is correct but the fitting algorithm may produce systematic errors in uncertainty measures.
The goal is to evaluate and potentially correct the fit.
We focus on the *first two moments* of the distribution of model fitted parameters

Steps:

1. Fit the model using the original data.
2. Extract multiple samples from the posterior predictive distribution.
3. Refit the model for each of these "bootstrap" samples.
4. Evaluate results against originally fitted parameters.
5. Adjust (if needed) original estimates.

Resampling based approach, details

Example: FH model

$$y_i \overset{\text{ind}}{\sim} N(\theta_i, v_i) \qquad \textit{sampling model}$$

$$\theta_i \overset{\text{ind}}{\sim} N\left(\mathbf{x}_i^T \beta, \tau_u^2\right) \qquad \textit{linking model}$$

Steps:

1. Fit the model using ADVI algorithm on the original data.
2. Extract random draws $(\theta_i^{(\alpha)}, y_i^{(\alpha)})$ from the posterior distribution of θ_i and the posterior predictive distribution of y_i, $\alpha = 1, \ldots, A$
3. Refit the model for each re-sampled dataset $\alpha = 1, \ldots, A$ using the same ADVI algorithm. Obtain posterior means $m(\theta_i^{(\alpha)})$ and variances $v(\theta_i^{(\alpha)})$ for respective parameters.

Pivotal quantity

- Form pivotal quantity

$$T_i^{(\alpha)} = \frac{m(\theta_i^{(\alpha)}) - \theta_i^{(\alpha)}}{\sqrt{v(\theta_i^{(\alpha)})}}.$$

- If $m(\theta_i^{(\alpha)})$ is unbiased for $\theta_i^{(\alpha)}$ and $v(\theta_i^{(\alpha)})$ is consistent estimate of its variance, then $T_i^{(\alpha)} \sim (0, 1)$.

- However, we assume our model parameters are estimated with an error, so the moments of $T_i^{(\alpha)}$ would have to be corrected.

How to adjust the pivot?

Suppose, true posterior variance of θ_i is

$$\mathrm{Var}(\theta_i) = v(\theta_i)c_i,$$

where

$v(\theta_i)$ is the posterior variance of θ_i obtained from the original run using some approximation algorithm;

c_i is a shift in scale due to the approximation algorithm.

1. Based on our assumption, in the "bootstrap world", $v(\theta_i) = v(\theta_i^{\alpha})c_i$. Thus, to correct the variance, we would have to divide pivot T_i^{α} by $\sqrt{c_i}$
2. Since we draw "bootstrap" samples from a posterior distribution with *biased* variance $v(\theta_i)$, we would have to divide pivot T_i^{α} by $\sqrt{c_i}$ (again!), to bring it up to *true* posterior variance $\mathrm{Var}(\theta_i)$.

Estimation of the adjustment

Variance of thus adjusted pivot is

$$\mathrm{Var}(c_i^{-1} T_i^{(\alpha)}) = 1 \qquad \Rightarrow \qquad c_i^2 = \mathrm{Var}(T_i^{(\alpha)})$$

Thus, we can estimate c_i from the bootstrap as

$$c_i = \sqrt{A^{-1} \sum_{\alpha=1}^{A} (T_i^{(\alpha)} - \bar{T}_i)^2},$$

where

$$\bar{T}_i = A^{-1} \sum_{\alpha=1}^{A} T_i^{(\alpha)}.$$

I. Pivot-based Confidence Intervals

The adjusted pivot is

$$\tilde{T}_i^{(\alpha)} = \frac{T_i^{(\alpha)} - \bar{T}_i}{c_i} \sim (0, 1).$$

The calibrated CI for area i is

$$C_i(\gamma) = \left[m(\theta_i) + \sqrt{v(\theta_i) c_i} \tilde{t}_{i,\gamma_l}, m(\theta_i) + \sqrt{v(\theta_i) c_i} \tilde{t}_{i,\gamma_r} \right],$$

where $\tilde{t}_{i,\gamma_l}$ and $\tilde{t}_{i,\gamma_r}$ are quantiles of $\tilde{T}_i^{(\alpha)}$ over the bootstrap distribution, for a given nominal level γ.

II. Re-scaled Confidence Intervals

Adjust each draw θ_i^* from the original posterior distribution of θ_i:

$$\tilde{\theta}_i^* = \frac{\theta_i^* - m(\theta_i)}{\sqrt{v(\theta_i)}} \sqrt{v(\theta_i) c_i} + \tilde{m}(\theta_i),$$

The adjusted CIs are obtained by computing percentiles over the adjusted draws $\tilde{\theta}_i^*$

Outline

Motivation

Resampling based approach

Simulations

Small domain model used in CES

Summary and Recommendations

Simulation setup: FH model

- Consider $N = 150$ domains. Set: $\beta = 1$, $\tau_u^2 = 1$, $\sigma_i^2 = 1$.
- Generate:

$$x_i \overset{\text{ind}}{\sim} Unif(0,2) \qquad \text{covariates}$$
$$u_i \overset{\text{ind}}{\sim} N\left(0, \tau_u^2\right) \qquad \text{random effects}$$
$$\epsilon_i \overset{\text{ind}}{\sim} N\left(0, \sigma_i^2\right) \qquad \text{random errors}$$

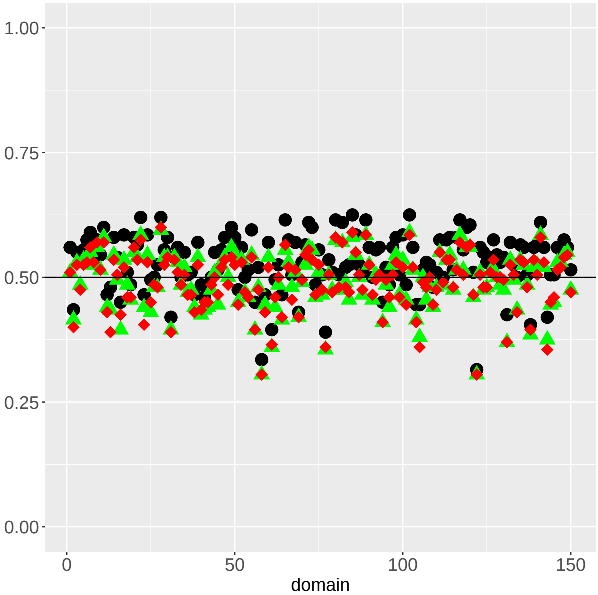
- True domain values: $\theta_i = \mathbf{x}_i^T \beta + u_i$
  "Direct" domain estimates: $y_i = \theta_i + \epsilon_i$
  Suppose variances of y_i are measured exactly $(v_i = \sigma_i)$ and known.
- Run the model and extract $S = 200$ *simulation* datasets from the posterior distribution of θ_i and posterior predictive distribution of y_i.
- Repeat re-sampling algorithm A times on each of S datasets.

Results: FH model

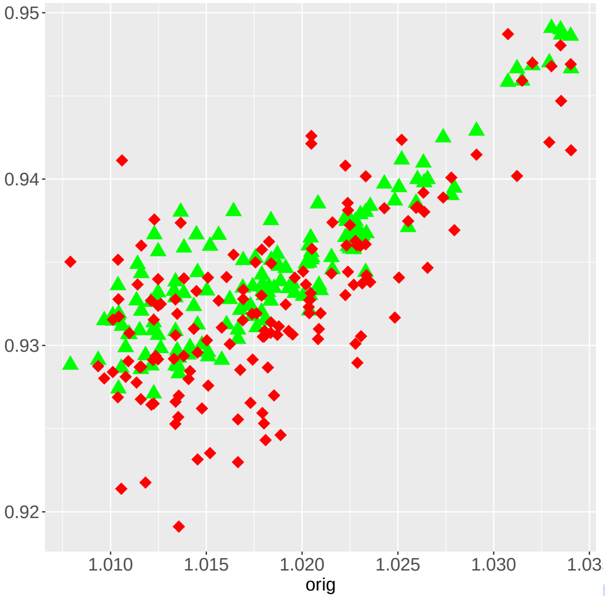
Table: Coverage properties for model fitted $m(\theta_i)$, 50% nominal, over 150 domains and $S = 200$ simulation runs, using $A = 500$ re-samples

	Orig Fitted	Rescaled	Pivot
Coverage	0.531	0.493	0.492
Length	1.019	0.935	0.933

Coverage of Point Estimate, 50% nominal, FH model



Length of 50% nominal CIs, FH model



Outline

Co-modeling of variances and co-clustering model (CCFH)

(y_i, v_i) are observed data, where y_i direct sample-based estimates and v_i are direct sample-based *estimates* of variances of y_i.
(θ_i, σ_i^2) are model parameters

$$y_i \overset{\text{ind}}{\sim} N\left(\theta_i, \sigma_i^2\right) \qquad \textit{sampling model for } \theta_i$$

$$\theta_i \overset{\text{iid}}{\sim} \sum_{k=1}^{K} \pi_k N\left(\mu_k + \mathbf{x}_i^T \beta, \tau_u^2\right) \qquad \textit{linking model for } \theta_i$$

$$v_i \overset{\text{ind}}{\sim} \sum_{k=1}^{K} \pi_k G\left(a n_i, a n_i b_k \sigma_i^{-2}\right) \qquad \textit{sampling model for } \sigma_i^2$$

$$\sigma_i^2 \overset{\text{ind}}{\sim} \sum_{k=1}^{K} \pi_k IG\left(2, \exp\left(z_i^T \gamma_k\right)\right) \qquad \textit{linking model for } \sigma_i^2$$

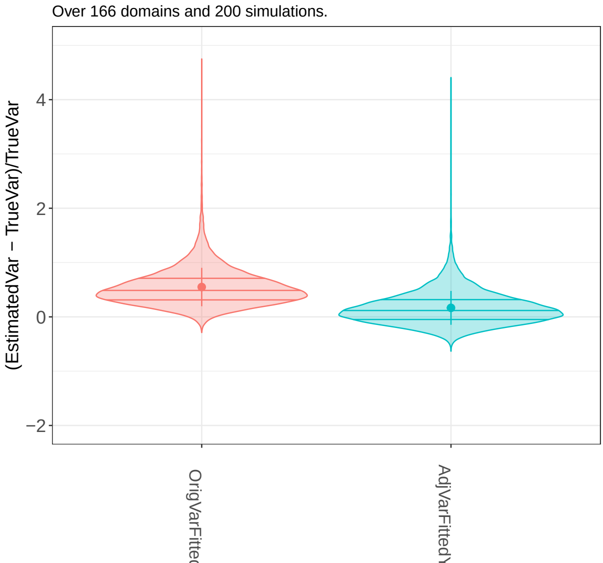
In CCFH, v_i are modeled along with the point estimates y_i.

Simulation setup: CCFH model

- Used real CES data for the initial run.
- Extract $S = 200$ *simulation* datasets from the posterior distribution of θ_i and posterior predictive distributions of y_i and v_i.
- Repeat re-sampling algorithm $A = 500$ times on each of S datasets.

Results: CCFH model

Relative residuals of original and adjusted variances of θ_i

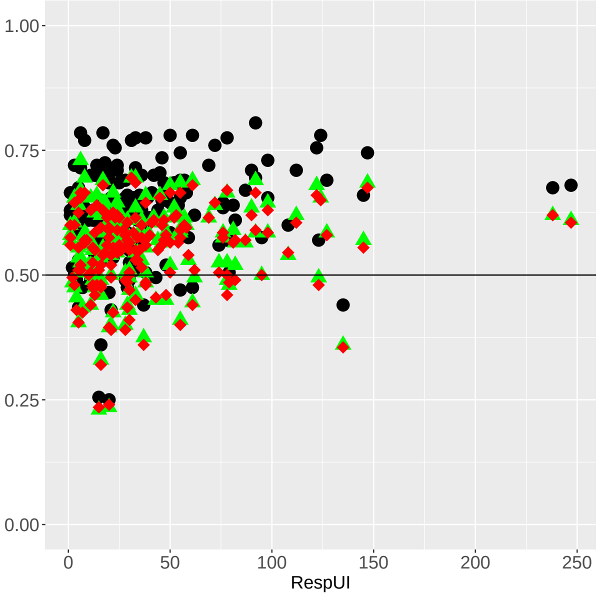


Results: CCFH model

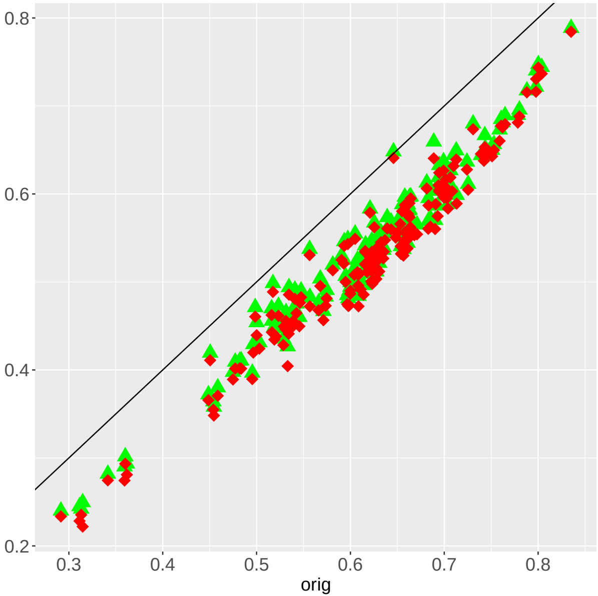
Table: Coverage properties for model fitted $m(\theta_i)$, 50% nominal, over 166 domains and $S = 200$ simulation runs, using $A = 500$ re-samples

	Orig Fitted	Rescaled	Pivot
Coverage	0.625	0.559	0.549
Length	0.618	0.537	0.528

Coverage of Point Estimate, 50% nominal, CCFH model



Length of 50% nominal CIs, CCFH model



Outline

Summary and Recommendations

- Inaccurate uncertainty measurements of some approximation algorithms have been reported in the literature
- We considered a resampling based evaluation and adjustment methods
- The methods rely on the assumption that the model is correct. Hence it is important to conduct thorough model checking
- Bias adjustments may be needed only if there is indications of a significant bias; otherwise, we may only be adding noise
- Although our main target was just the first two moments of the distribution of the fitted parameters, for the considered models, the procedure also lead to CIs with nearly nominal coverage properties
- The pivot-based method gives slightly shorter but more variable CIs. It is also less practical as it requires larger number of bootstrap samples.

Thank you!

- CONTACT INFORMATION:
- Gershunskaya.Julie@bls.gov