Validating model-based estimates for small domains (areas) with small or no samples: a simulated population approach

Xingyou Zhang
Statistical Method Group (SMG)
Office of Compensation and Working Conditions (OCWC)
Bureau of Labor Statistics

FCSM Research and Policy Conference
Hyattsville, MD
October 22, 2024

Co-authors

Statistical Methods Group (SMG)

Office of Compensation and Working Conditions (OCWC)

Bureau of Labor Statistics (BLS)

Danny Friel

Erin McNulty

Yu Zhang

The views expressed here are those of the authors and do not necessarily reflect the views or policies of the Bureau of Labor Statistics or any other agency of the U.S. Department of Labor.

Outline

- Small domain estimation (SDE) review
  - Small domains and models
  - SDE validation dilemma
- Small domain estimation validation via simulated samples
  - Small domain estimation with Occupational Requirements Survey (ORS)
    - ORS survey, sample design, and SDE objective
    - Modeling framework and basic steps
    - SDE validation metrics
    - Preliminary Results

Small Domain Estimation (SDE)

- Small domain (area) is defined as an estimation domain where direct survey estimates don't have adequate precision or are not available because of small sample or no sample at all
  - Geographic areas, population groups, industrial sectors, or occupational groups
- Small domain(area) estimation aims to produce statistically reliable estimates of interest for small domains
  - Two basic approaches for model-based SDE
    - Area-level model
    - Unit-level model

SDE validation dilemma

- Internal validation
  - Compare model-based estimates with reliable direct survey estimates
    - Validate model-based estimates for the domains with large sample sizes?
  - How about validating the model-based estimates for those domains with small or no samples?
- External validation
  - Estimates from large surveys
    - American Community Survey (ACS), Occupational Employment and Wage Statistics (OEWS)
  - Estimates from administrative or alternative data sources
  - Limited availability

SDE validation via simulated samples

- Treat ORS721-725 as a population
- Take a stratified random sample from ORS721-725
- Apply a Multilevel Regression and Poststratification (MRP) approach to predict small domain estimates for 6-digit SOCs (occupations)
- Compare model-based estimates with population estimates for 6-digit SOC small domains estimates: bias, coverage accuracy and correlation

Occupational Requirements Survey (ORS)

- A survey conducted by BLS on behalf of the Social Security Administration (SSA) and collects data to measure the requirements of work in the national economy in four areas:
  - 1) Physical demands; 2) Environmental conditions; 3) Education, training, and experience; 4)Mental and cognitive demands
- Sample design
  - A two-stage stratified sample of establishments and occupations within selected establishments
  - Stratum formulation involves ownership, geography, industry and occupation
  - Annual sample size of 10,000 for ORS721-723 and 15,000 for ORS724-725
    - 85% private industry establishments
    - 15% State and Local Government establishments
    - 1st sample group fielded in September 2018

ORS Small Domain Estimation Objective

- Objective: producing reliable estimates for all 844 target occupations (6-digit SOCs)
  - SSA expects populated estimates by SOC for its disability programs

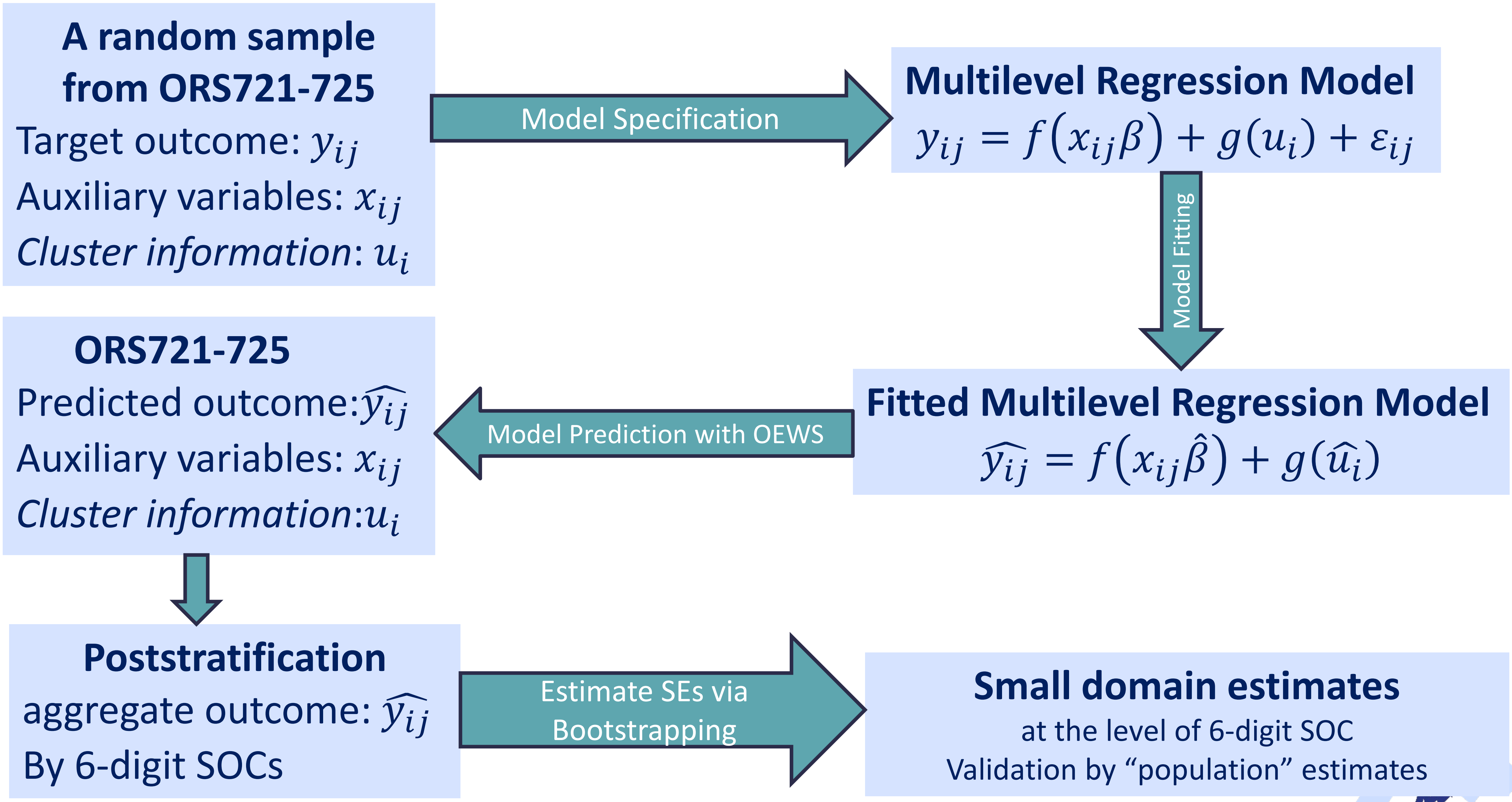
- ORS721-725
  - In total 112,147 job quotes with valid job requirements data
  - **830** out of 844 SOCs were sampled
  - The median job quote sample size is 43 by SOCs
  - 515 SOCs had at least 30 job quotes

- The entire ORS721-725 was treated as a population

Take a stratified random ORS Sample

- Strata
  - Ownership
    - Government (10 strata)
      - Industry (five industrial groups)
      - Occupation (rare vs non-rare)
    - Private (40 strata)
      - Geography (four census regions)
      - Industry (five industrial groups)
      - Occupation (rare vs non-rare)
- Sampling
  - Apply ORS725 stratum sample allocation
  - Take a stratified random sample with a ORS725 sample size of 23,045
  - Strata sample sizes with a median of 156 and a mean of 461
  - 776 out of 830 SOCs were sampled

Modeling framework: Multilevel Regression and Poststratification (MRP) for Small Domain Estimation



A random sample from ORS721-725

Target outcome: y_{ij}

Auxiliary variables: x_{ij}

Cluster information: u_i

Model Specification

Multilevel Regression Model

$$y_{ij} = f\left(x_{ij}\beta\right) + g(u_i) + \varepsilon_{ij}$$

Model Fitting

Fitted Multilevel Regression Model

$$\widehat{y_{ij}} = f\left(x_{ij}\hat{\beta}\right) + g(\widehat{u_i})$$

Model Prediction with OEWS

ORS721-725

Predicted outcome:$\widehat{y_{ij}}$

Auxiliary variables: x_{ij}

Cluster information:u_i

Poststratification

aggregate outcome: $\widehat{y_{ij}}$

By 6-digit SOCs

Estimate SEs via Bootstrapping

Small domain estimates

at the level of 6-digit SOC

Validation by “population” estimates

Step 1: Use the ORS random sample to construct and fit a multilevel model

- This step is to construct multilevel logistic models to borrow information across entire ORS sample to estimate the model parameter estimates: $logit(p_{ij}|y_{ij} = 1) = x_{ij}\beta + z_{ij}\gamma_i$
  - y: the known ORS binary job requirement of interest, such as low posture
  - x: the known fixed effects and β: their unknown regression coefficients
    - 3 establishment ownership (state, local and private)
    - 4 establishment employment size groups (<=49, 50-99, 100-499, and >=500)
    - 24 NCS sampling geographic areas (9 census divisions plus 15 CSAs/MSAs)
    - Industrial groups (2-digit NAICS)
  - z: the known random effects and γ: their unknown regression coefficients
    - detailed occupation groups (6-digit SOCs)
- The model fitting process is to estimate what are the most likely values of model parameters (β and γ), given the known data from ORS (y, x, and z)
  - **This step is the statistical learning process from the ORS random sample**

Step 2: Apply the fitted multilevel model parameters to ORS721-725 to obtain the predicated values for the outcome of interest ($\hat{y}$)

- Model prediction space is the set of small domains based on all possible combinations of model predictors
  - 24 ORS sampling geographic areas, 4 employment size groups, 20 industrial groups, 3 ownership groups, plus 830 6-digit SOC occupation groups, **4,780,800** small domains in total

- We could conveniently have a predicted value for the outcome ($\hat{y}_i$) for each ORS721-725 quote, since X, Z are known, $\hat{\beta}$ and $\hat{\gamma}$ are known after model fitting

$$\hat{p}_{ij} = \frac{e^{X_{ij}\hat{\beta}+Z_{ij}\hat{\gamma}_i}}{1+e^{X_{ij}\hat{\beta}+Z_{ij}\hat{\gamma}_i}} \text{ and } \hat{y}_{ij} \sim Bernouulii\left(\hat{p}_{ij}\right)$$

- The modeling prediction process is to estimate what are the most likely predicted values of outcome of interest ($\hat{y}$), given the known data (X, Z) from ORS721-725 and the known model parameters ($\hat{\beta}$ and $\hat{\gamma}$)

Step 3: Use the predicted $(\hat{y}_{ij})$ in ORS721-725 and we could calculate the populated estimates of interest: $\bar{\hat{y}}_i = \frac{\sum_{j=1}^{n_i} \hat{y}_{ij}}{n_i}$

- $\boldsymbol{n_i}$ is the number of job quotes for a 6-digit SOC_i in ORS721-725
- $\boldsymbol{\hat{y}_{ij}}$ is the predicted value of job requirements for job quote (j)
- $\boldsymbol{\bar{\hat{y}}_i}$ is the model-based estimate for a 6-digit SOC small domain

Step 4: Obtain the variance estimates associated with small domain estimates via bootstrapping

- $\hat{y}_i^B = \frac{\sum_{j=1}^{n_i} \hat{y}_{ij}^B}{n_i}$, where B =1, 2, …, 1,000
- With a sample of 1,000 $\hat{y}_{ij}^B$, we could conveniently to produce 1,000 $\hat{y}_i^B$ and obtain any summary statistics for $\hat{y}_i$
  - Mean, median or any other percentiles: 2.5 percentile and 97.5 percentile to construct 95% Confidence intervals (CIs)

SDE validation metrics

- Estimation bias
  - Mean Absolute Error (MAE): $\mathrm{MAE} = \frac{\sum_{i=1}^{n} |\bar{\hat{y}}_i - \bar{y}_i|}{n}$
    - The mean absolute difference between 6-digit SOC-level population estimates ($\bar{y}_i = \frac{\sum_{j=1}^{n_i} y_{ij}}{n_i}$) and model-based estimate ($\bar{\hat{y}}_i$)
    - n is the number of SOCs for comparison
  - Root Mean Squared Error (RMSE): $\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} [\bar{\hat{y}}_i - \bar{y}_i]^2}{n}}$
- Coverage accuracy (Rate and Count)
  - The proportion/number of 6-digit SOCs that population estimates ($\bar{y}_i$) are within the 95% confidence intervals of model-based estimates ($\bar{\hat{y}}_i$)
- Correlation between $\bar{y}_i$ and $\bar{\hat{y}}_i$
  - Pearson linear correlation coefficient (PCC)
  - Spearman's rank correlation coefficient (SCC)

Apply MRP for ORS Small Domain Estimation

- 4 binary ORS job requirements selected
  - Personal protective equipment use (PPE) (Yes vs No)
  - Telework (TELEWK) (Yes vs No)
  - Low Posture (LOWP) (Yes vs No)
  - Minimum Degree Requirement (MINEDU) (Yes vs No)
- Job requirement prevalence (%) based on ORS721-725

Outcome	Mean	SE	RSE	LCL	UCL
PPE	6.17	0.2540	4.12	5.66	6.67
TELEWK	10.56	0.2442	2.31	10.07	11.04
LOWP	57.63	0.3418	0.59	56.95	58.31
MINEDU	69.83	0.3796	0.54	69.08	70.59

Comparison between model-based estimates and ORS721-725 population estimates for LOWP by SOC Sample Size

Sample	#SOC	Bias		Accuracy		Correlation		Survey
Size		MAE	RMSE	Rate	Count	PCC	SCC	Weights
0~527	830	11.14	15.77	0.73	609	0.89	0.88	NO
0~527	830	11.15	15.78	0.74	612	0.89	0.88	YES
0~0	54	28.71	34.75	0.93	50	0.59	0.56	NO
0~0	54	28.72	34.73	0.93	50	0.59	0.57	YES
1~5	251	17.06	20.25	0.79	198	0.84	0.81	NO
1~5	251	16.99	20.19	0.79	199	0.84	0.81	YES
6~10	153	9.71	11.56	0.80	122	0.94	0.93	NO
6~10	153	9.71	11.70	0.80	123	0.94	0.93	YES
11~15	76	7.90	9.95	0.66	50	0.96	0.96	NO
11~15	76	8.07	10.14	0.64	49	0.96	0.95	YES
16~29	109	6.28	7.73	0.68	74	0.97	0.96	NO
16~29	109	6.35	7.81	0.69	75	0.97	0.96	YES
30+	187	3.44	4.43	0.61	115	0.99	0.99	NO
30+	187	3.46	4.51	0.62	116	0.99	0.99	YES

Preliminary conclusions

- Fitting models without survey weights tends to produce small domain estimates with smaller bias
- Compared to correlation coefficient metrics, bias metrics are more sensitive for validating small domain estimates
- Coverage accuracy metrics, sensitive to standard errors of small domain estimates, could be misguided of small domain estimates
- A small sample increase could reduce a large amount of bias of small domain estimates (e.g. occupation relatedness)

Contact Information

Xingyou Zhang
Division Chief, Statistical Methods Group
Office of Compensation and Working Conditions (OCWC)
www.bls.gov/ors
202-691-6082
Zhang.Xingyou@bls.gov