Simulated data
(for JOLTS area-level model)

January 21, 2022

1 Setup

First, we generate a finite population stratified by "geography" and a "size class" indicator and construct domain true statistics. Then we draw a stratified simple random sample with replacement from the finite population and obtain direct sample-based and model-based estimates for the population domains. The estimates are evaluated against domains true finite population totals. (In the simulation study, the above steps are repeated a large number of times to capture the variation of population generation and the taking of a sample; then, biases and mean squared errors of competing estimators would be calculated over the simulations.) The details follow.

1.1 Configuring the Finite Population: number of units, domains, and regions

We set a population of size $N_p = 1,000,000$ consisting of $D = 50$ domains. There are 5 types of domains, based on the number of population units each domain type contains; namely, $N_d \in \{39000, 30000, 20000, 10000, 1000\}$, $d = 1, ..., 50$ In this set-up, the 10 largest-sizes domains are constructed with 39,000 population units, whereas the 10 smallest domains are each assigned only 1,000 population units. Each domain belongs to one of $R = 4$ "regions": domains $d = 1, ..., 10$ are in region $r = 1$; $d = 11, ..., 20$ are in region $r = 2$; $d = 21, ..., 30$ are in region $r = 3$; and domains $d = 31, ..., 50$ are region $r = 4$.

1.2 Constructing Population "Employment Size Classes"

The population is subdivided into $S = 6$ "employment size classes" based on variable y_j^{emp} ("employment").
Let size class mean levels be $m_s^{emp} = \{2, 10, 20, 40, 100, 1000\}$.
The total number of population units assigned to each size class is
$N_i = \{700000, 110000, 90000, 70000, 20000, 10000\}$, $\sum_{s=1}^{S} N_s = N_p$. The "employment" level for each business establishment unit in each size class is generated as a Poisson variable $y_j^{emp} \sim \text{Poisson}\,(m_s^{emp})$, $j \in s$, $s = 1, ..., S$.

Each domain includes approximately equal proportion of units from each size class. The true finite population domain employment level that we generate as $Y_d^{emp} = \sum_{j=1}^{N_d} y_j^{emp}$, is assumed to be known for our modeling.

1.3 Generating a Sub-employment Variable

Let $x_d \sim Unif\,(0.02, 0.3)$, $d = 1, ..., D$, be a domain specific predictor that is assumed to be fixed and known. We set parameters $\sigma_\lambda^2 = 0.1$, $\sigma_\epsilon^2 = 1$, $\beta = 0.7$ and generate $\lambda_d \sim N(\beta \log(x_d), \sigma_\lambda^2)$, $d = 1, ..., D$ and $\epsilon_j \sim N(-0.5\sigma_\epsilon^2, \sigma_\epsilon^2)$, $j = 1, ..., N_p$.

Population values for a sub-employment variable (e.g., job openings, hires or separations), y_j, are generated from an overdispersed Poisson distribution, as $y_j \sim Poisson(m_j)$, with means $m_j = y_j^{emp} \exp(\lambda_d + \epsilon_j)$, $j = 1, ..., N_p$.

True finite population targets are domain totals $Y_d = \sum_{j=1}^{N_d} y_j$.

1.4 Sampling Design and Estimation

We use a stratified simple random sampling with replacement design, where strata are defined by intersections of regions and size classes, $h = 1, ..., H, H = RS$. Sampling selection probabilities are defined by employment size strata as $\pi_s = (0.00025, 0.00075, 0.00125, 0.0025, 0.0075, 0.0125)$, which induces respective sampling weights $w_s = (4000, 1333, 800, 400, 133, 80)$. Strata that share the same employment size classes across regions are assigned the same sampling selection probabilities. In designing the sampling scheme for this simulation, we strive to produce sampling weights that would resemble those in the actual real data JOLTS application, where the weights range from about 1 to 5000.

The direct sample weighted domain ratio estimator has the form $\hat{Y}_d^{Dir} = Y_d^{emp} \hat{R}_d$, where

$$\hat{R}_d = \frac{\sum_{j \in S_d} w_j y_j}{\sum_{j \in S_d} w_j y_j^{emp}},$$

and S_d is the set of sampled units in domain d and w_j is a sampling weight of unit $j \in S_d$.

Note that this sampling design does not guarantee that a given domain is represented in the sample. There is a chance that some of the smaller domains would not be included into sample, and thus, their direct estimates are not defined. In this case the model would still provide an estimate for these domains.

The model input data includes variances of direct estimates, which we compute from the sample using the usual linearization formula of the ratio estimator:

$$\hat{V}_d^{Dir} = \sum_h \frac{N_{dh}^2}{n_{dh}} \frac{\sum_{j \in S_{dh}} (u_{dj} - \bar{u}_{dj})^2}{n_{dh} - 1}, u_{dj} = y_j - \hat{R}_d y_j^{emp}, \bar{u}_{dj} = \frac{1}{n_{dh}} \sum_{j \in S_{dh}} u_{dj},$$

where S_{dh} is a set of sampled units belonging to stratum h and domain d. The "regional" and "national" level point and variance estimates are derived using analogous formulas.