Toy simulations

jgershun

January 24, 2022

We start with a relatively simple scheme. The goal here is to allow us to tune models on a small dataset to achieve robust mixing on the real data. In future, we will add complexity to allow assessment of fit performances on the synthetic data.

1 The setup

1.1 Generating a finite population

1.1.1 The population structure

For the current scenario, we let

$$\text{Number of domains: } M = 50$$
$$\text{Number of strata: } H = 20$$

Domains are of equal sizes and strata are equally distributed in each domain, as follows:

$$N_{mh} = c(1, 3, 5, 10, 15, 20, 30, 50, 70, 90, 120, 140, 160, 180, 200, 250, 350, 450, 550, 1000),$$

so that strata sizes are $N_h = MN_{mh}$, which comes out to:

```
> N_h
 [1]     50   150   250   500   750  1000  1500
 2500  3500  4500  6000  7000  8000  9000 10000
 12500 17500 22500 27500 50000
```

$$N_m = \sum_{h=1}^{H} N_{mh} = 3694, N = 184700$$

Each domain belongs to one of $R = 4$ "regions", as follows: region $r = 1$ consists of 20 domains, regions $r = 2, 3, 4$ consist of 10 domains each.

1.1.2 Generating population values

We generate values for population observation i ($i = 1, ..., N_h$) in stratum h ($h = 1, ...H$) and domain m ($m = 1, ...M$) as

$$\begin{aligned}
\log(x_{mhi}) &= a_h + u_{0m} + u_{0mh} + \epsilon_{0hi} \\
\log(y_{mhi,t=1}) &= b_{1h} + u_{1m} + u_{1mh} + \log(x_{mhi}) + \epsilon_{1hi} \\
\log(y_{mhi,t=2}) &= b_{2h} + u_{2m} + u_{2mh} + \log(y_{mhi,t=1}) + \epsilon_{2hi}
\end{aligned}$$

with

$$\begin{aligned}
&\epsilon_{0hi} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2_{0h}), \quad \epsilon_{1hi} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2_{1h}), \quad \epsilon_{2hi} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2_{2h}) \text{ (random errors)}, \\
&\qquad\qquad\qquad\qquad b_{1h} \overset{iid}{\sim} \mathcal{N}(0, \psi^2_1), \quad b_{2h} \overset{iid}{\sim} \mathcal{N}(0, \psi^2_2) \text{ (stratum effect)}, \\
&u_{0m} \overset{iid}{\sim} \mathcal{N}(0, \tau^2_0), \quad u_{1m} \overset{iid}{\sim} \mathcal{N}(0, \tau^2_1), \quad u_{2m} \overset{iid}{\sim} \mathcal{N}(0, \tau^2_2) \text{ (domain effect)}, \\
&u_{0mh} \overset{iid}{\sim} \mathcal{N}(0, \phi^2_0), \quad u_{1mh} \overset{iid}{\sim} \mathcal{N}(0, \phi^2_1), \quad u_{2mh} \overset{iid}{\sim} \mathcal{N}(0, \phi^2_2) \text{ (domain/stratum interaction effect)}
\end{aligned}$$

where we use the following parameter values

$a_h = log(c(2000, 1000, 500, 300, 200, 100, 90, 80, 70, 60, 50, 40, 30, 20, 10, rep(5, 5)))$,

$\sigma_{0h} = \sigma_{1h} = \sigma_{2h} = 0.4$ for all h=1,...H,
$\psi_1 = \psi_2 = 0.3$,

$\tau_0 = \tau_1 = \tau_2 = 0.2$,

$\phi_0 = \phi_1 = \phi_2 = 0.1$.

Note that "regional" effects were not explicitly added to the generating model under the current scenario.

1.2 Select stratified sample

From each stratum h, we draw a simple random sample (with replacement) with probability π_h, where strata selection probabilities are proportional to the standard error of x, i.e., to $\sigma_{xh} = (\frac{1}{N_h-1}\sum_{i=1}^{N_h}(x_i - \overline{x}_h)^2)^{1/2}$; the largest variance stratum is sampled with certainty.

Thus, $\pi_h = C\sigma_{xh}$, where $C = 1/\max_h(\sigma_{xh})$.

For example, a particular realization ($sim = 1$) comes out to the following set.

The total sample size is $n = 1757$; the realized number of sample units is listed below.

Per domain:

```
> print(n_m)
 [1] 27 33 36 44 38 40 33 24 40 45 29 39 33 49 33 32 32
 34 34 38 35 34 37 44 28 25 27 39 49 34 28 33 36 36 42
 34 33 34 40 44 45 37 34 35 27 22 36 36 35
```

Per stratum:

```
> print(n_h)
 [1]  50  64  59  76  82  57  75 105 130 144 154 144 128
94  52  31  46  59  74 133
```

The sampling weights are $w_h = 1/\pi_h$:

```
> print(round(w_h,4))
 [1]   1.0000   2.3368   4.2503   6.5411   9.1542  17.4637
19.8917  23.9068  26.9212  31.2538  38.9355  48.4761  62.5577
95.6410 191.2506 398.4659 376.7961 379.9725 370.1730 376.4017

> Some quick stats:
[01/24/22] from y_countimp_ispbase_01052022.stan

Month 1:
   estnm              bias              mad
1    HT1  624.902743425182 14010.8859551622
2   WLR1  422.368056890071 6540.21523944997
3   Unw1  8726.60550393349 13190.0436936942
4   Pst1 -797.871800888308 7299.82635643497
5 fitted1  1989.8202853117 4730.98067053844

Month 2:
   estnm              bias              mad
1    HT2 -1401.43587433581  16131.353359957
2   WLR2 -1592.76492647959 8228.56250363442
3   Unw2  26511.1396844847 28370.9637027193
4   Pst2 -2711.16156756266 8533.35039726574
5 fitted2 -2348.08287626042 5625.49460548267
```