

---

Subject: Re: Weighting in pooled data

Posted by [Reduced-For\(u\)m](#) on Mon, 16 May 2016 23:39:18 GMT

[View Forum Message](#) <> [Reply to Message](#)

---

Owraza,

I think this comes down to a matter of interpretation, not a specific right/wrong way of doing it.

Lets suppose you have 2 countries and 10 regions each. After collapsing, you have 20 observations, 10 from each country. Each of these is representative of a particular region. You know want to infer some parameter value from the data you have - say, the effect of cluster-aggregated variable X.

As an extreme example, suppose that your two countries were Nigeria (population 180 million) and Burkina Faso (population 17 million). If you wanted "average stunting rate" for children, you would want to weight the Nigerian data to be about 10 times more influential than the BF data. Or, more specifically, you'd want to weight each region by the relevant population (say, children under 5). In that case, you'd just give each region (after collapsing) a new weight that was equal to its population.

Now, suppose you are interested instead if the effect of variable X. If the effect of X is the same everywhere, you don't need to do any further weighting if you don't want to. However, there are two reasons you could still weight. 1) the estimates from larger sample sizes within regions are less variable (remember, your region-level aggregates are really region-level "estimates"), and so weighting-up the higher-sample-size regions can give you some statistical power; 2) we might think that there are variations in the effect of X for different people due to unobserved factors, and what we want to know is the "average treatment effect" of the whole population. In that case, if you can't model the heterogeneity in the effect of X, you might want to weight by population again in order to back out the average effect on the population.

One way to think about it is: am I interested in doing inference on levels/effects for Regions or Individuals? If regions, then once you collapse you are fine (each region is its own observation, and another region from whatever country is adding one new, equally important data point). If you want to know about average effects across individuals in the whole population of the two countries (taken as one population for purposes of inference) then you'd probably want to weight these regions by population in some manner.

But like I said, I don't think this is a "right/wrong" kind of thing. It is an "it depends" kind of thing. Once you collapse to region, you have a region-level-estimate that is weighted to be representative of that region. What you do from there depends on what you want that region to tell you and what you want your analysis to capture. In the world of estimating causal effects, we tend to pretend that our constant-effect model is right, in which case you don't need to weight further. But in worlds interested in estimating population-level characteristics, then you do want to acknowledge that different regions are telling you information about different numbers of people.

Help? Or just obscure things more?

---