Creating and Using Post-Stratification Weighting for District-level Estimates

Shiro Kuriwaki*

October 2019

In this memo I explore how representative CCES samples are at subnational geographies, and whether post-stratification weights ameliorate the problem. Although post-stratification weights are designed to coerce samples to be representative, one set of weights makes a sample representative to only one target, in CCES' case the entire U.S. It is not clear if those same weights *also* make subgroup samples representative of their respective subgroup targets.

Multi-level post-stratification (MRP) methods take another route, whereby the targets for granular joint distributions are computed first so that they can be combined to any preferred level of aggregation for the post-stratification step. However, MRP has several limitations. First, often times analysts only have marginal, not joint distributions at the level they care about. Second, one MRP model is specific to one outcome, so comparing more than one issue question would involve building as many multi-level models as there are outcomes. In contrast, once post-stratification weights are defined, it can be used for any sets of outcomes.

1 Post-stratification Weights and MRP

Each method has its pros and cons, sumarized by the following table:

	Weighting	MRP
Applying to different outcomes	Easy	Hard, need one model per outcome
Applying to different subgroups	Hard, need to create one set of weights for each target geography	Easy
Pools Observations from different geographies?	Usually no	Yes
Key Modeling Assumptions	The selection model	The outcome model

*Ph.D. Candidate, Department of Government and Institute of Quantitative Social Science, Harvard University. Thanks to Soichiro Yamauchi for helpful discussions.

Setup

The goal of the rest of this memo is to document how weighting affects the distribution of covariates at different sub-geographies. I do not discuss MRP here — this is because the joint distributions of the covariates that are available are matched exactly by construction, and the worry for MRP is about the outcome model, not the weighting model.

The ACS and CCES collect the following shared variables about their respondents. There is some difference in question wording or lumping categories, but here for simplicity I use the most common denominator.

Gender	Age	Education
• Male • Female	• 18 to 24 years • 25 to 34 years • 35 to 44 years • 45 to 64 years • 65 years and over	• No high school • High school graduate • Some college, no degree • Associate's degree • Bachelor's degree • Graduate / professional

Exploration

I considered three types of estimates:

1. Unweighted sample proportions
2. Weighted proportions with YouGov's national post-stratification weights
3. Weighted proportions with custom state-specific weights.

YouGov's national weights The CCES 2018 Guide reports

> The [matched] cases and the frame were combined and the combined cases were balanced on multiple moment conditions using the 2017 ACS. ... First, for the common content, the completed cases were weighted to the sampling frame using entropy balancing. ... The CCES sample was weighted to match the distributions of the 2017 ACS ...
>
> The moment conditions included age, gender, education, race, plus their interactions. The resultant weights were then post-stratified by age, gender, education, race, "born again" status, voter registration status, and 2016 Presidential vote choice, as needed. Additionally, for the common content, the weights were post-stratified across states and statewide political races (for governor and senator). Weights larger than 15 in the common content were trimmed and the final weights normalized to equal sample size.

Although we do not have access to YouGov's full code, we can partly reproduce this procedure. The ACS provides their own estimates of marginal and some distribution of demographics at the national, state, and congressional district level. We uses those as a source of our target distribution.

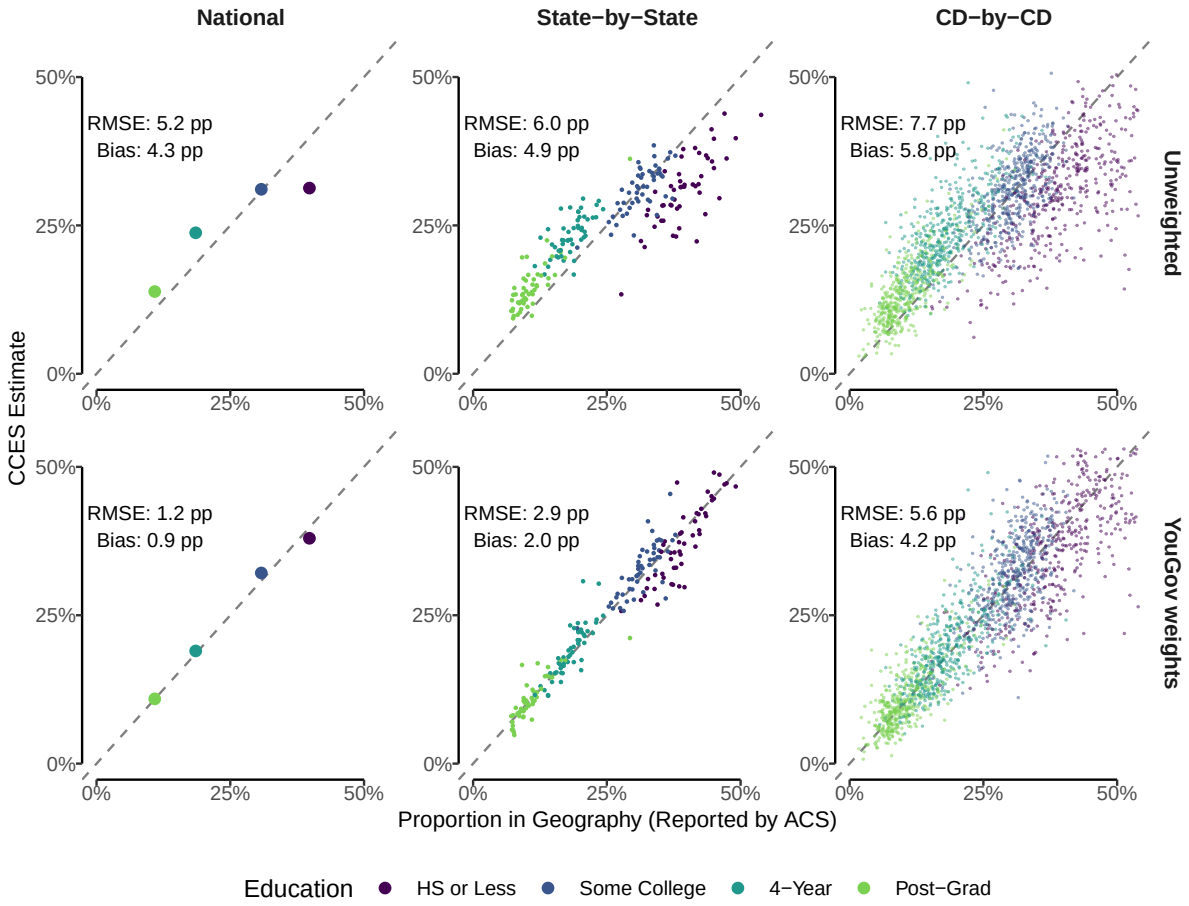
State-specific weights I created simple rim weights by going state sample by state sample, and assigning a set of weights that targeted marginal distribution of gender, age, and education in that state (as reported by the ACS). I did not target the joint distributions because some state samples were too samples were too small and had zero observations for some of the cells.

YouGov Weight Results

We first start by evaluating one metric, education, in Figure 1. We notice several things from the figure:

1. The first set of plots in the first row show that small-samples are on average less representatives than larger ones.
2. The second row, by comparison, shows that YouGov's weights make the estimates more representative. Although the weights primarily target the national distribution, (a) the weighted average for national estimates are not perfect, and (b) state and district estimates are improved as well.
3. Most of the improvement in the second comes from a reduction in bias rather than reduction in variance.
4. There is a smaller reduction in bias in the district level estimates.

Figure 1: Representativeness of samples at different levels of geographies, education

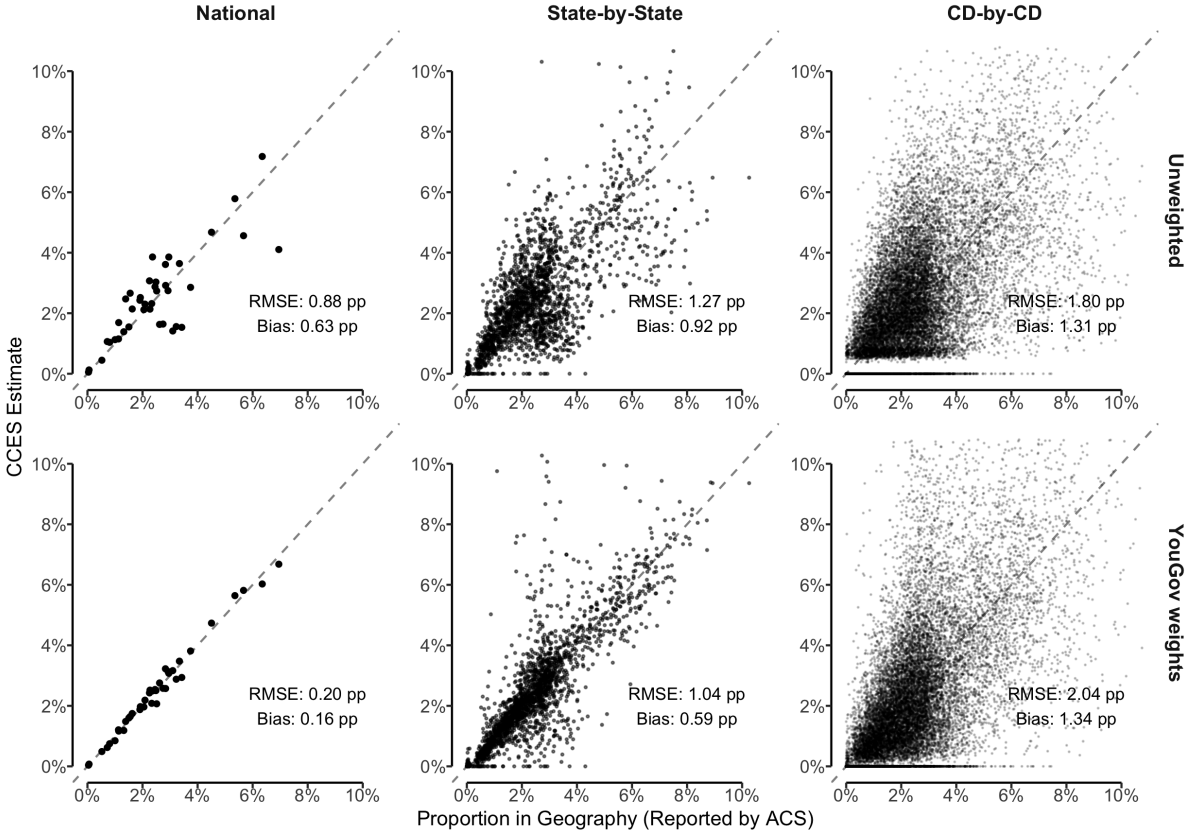


CCES 2018, ACS 1yr 2017. All CCES weighting uses YouGov's national weights, even for state/CD subsets in middle/right panels. CCES/ACS estimate the proportion of an education cell (6 combinations) per geography (1 nation, 50 states, or 435 CDs).

We next take a look not only at education, but the representativeness in terms of education-gender-age joint distributions. YouGov weights on first moments and (presumably two-way) interactions, so their weighted are not guaranteed to hold for three-way joint distributions. Figure 2 shows those quantities of interest, proliferating the number of points to examine. We find:

1. National estimates improve about as much as in Figure 1 in ratio terms.
2. State estimates also improve somewhat, but not by much (8 percent reduction in RMSE, as opposed to 40 percent in the marginal distribution case).
3. District estimates do not improve, and its bias *increases* slightly by 0.08 percentage points. The bias variance decomposition suggests that the variance has increased as well.
4. Some of the outliers in the state estimate suggests that the YouGov national weights up-weight some state-demographic cells in a way that makes them less representative of the state.

Figure 2: Representativeness of samples at different levels of geographies, education × age × gender fraction



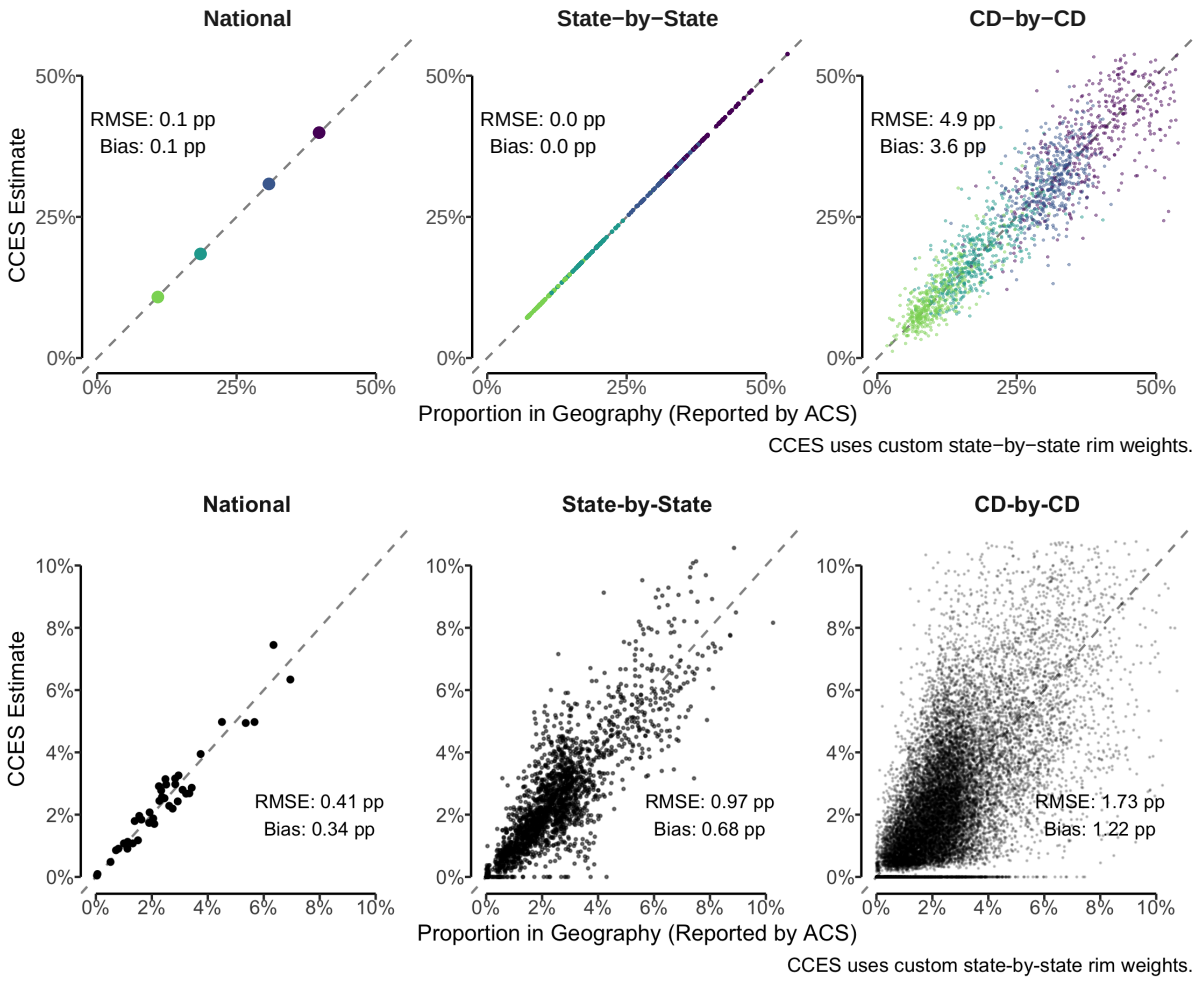
Source: CCES 2018, ACS 1yr 2017. All CCES weighting uses YouGov's national weights, even for state/CD subsets in middle/right panels. CCES/ACS estimate the proportion of a {gender x age bin x education} cell (60 combinations) per geography (1 nation, 50 states, or 435 CDs)

Custom Rim Weights

Do weights that are specifically targeted to match on the state-specific moments improve representativeness? These weights must get the state-by-state marginal distributions exactly right by construction. How does it improve representativeness at larger or lower levels of aggregation? Figure 3 shows figures analogous to the prior figures:

1. The rim weights coerce the marginal estimates to match the population targets, as expected.
2. Although each state subsample's weights is computed separately, its concatenation makes national estimates representatives too.
3. At the smaller district level, the marginal estimates have also improved, and slightly outperform the YouGov weights.
4. However, rim weights do not necessarily improve representativeness of joint distributions. The bottom panel shows that those estimates are about as representative as those with YouGov weights, perhaps by reducing the variance.

Figure 3: Representativeness with custom state-by-state rim weights



Method

We first discuss two traditional approaches: rake weights and propensity score weights. These traditional approaches each have known limitations. After highlighting them, we outline our own proposal, a doubly-robust estimator with balancing properties.

A Introduction to Rake Weights and Iterated Weighting (rim weights)

Post-stratification is the process of generating weights to coerce the weighted proportion of the sample to equal any given target distribution (it is called "post" because it occurs after the survey was fielded, instead of over-sampling low-propensity individuals during the survey). The method is very simple and was outlined in Deming and Stephan (1940).

Index individuals by $i \in \{1, ..., n\}$. Each individual has a scalar weight at iteration t as $w_i^{(t)}$, and $j \in \{1, ..., K\}$ categorical covariates $\boldsymbol{X}_i$. For example, one grouping could be "education", which is a discrete variable with four levels (high school or less, 2-year college, college, or post-graduate). Denote the levels of group j by $\ell \in \{1, ..., L_j\}$.

For each grouping, there is a target distribution

$$\boldsymbol{\pi}_j = \pi_{j1}, ...\pi_{KL_j}, \quad \sum_{\ell=1}^{L_j} \pi_{j\ell} = 1$$

that we would like to hit. In parallel, there is an empirical proportion from the survey:

$$\widehat{\boldsymbol{\pi}}_j = \widehat{\pi}_{j1}, ..., \widehat{\pi}_{jL}, \quad \sum_{\ell=1}^{L_j} \hat{\pi}_{g\ell} = 1.$$

These estimates are weighted averages of binary random variables, where th normalized weights at iteration t are applied as simple weighted means:

$$\widehat{\boldsymbol{\pi}}_{j\ell}^{(t)} = \frac{1}{n} \sum_{i=1}^{n} w_i^{(t)} \mathbf{1}(X_{ji} = \ell)$$

At each group j, we update the weights simply by computing the ratio of target over actual for each level ℓ

$$r_{j\ell}^{(t+1)} = \frac{\pi_{j\ell}}{\widehat{\pi}_{j\ell}^{(t)}}. \tag{A.1}$$

This is the crux of rake weighting, which is the same across various packages.[1] To interpret this ratio as weights, we apply these factors equally to all respondents for which $X_{ji} = \ell$ that the

[1] See, e.g., the relevant lines in the source code of `anesrake`, `iterake`, and `survey`.

sum of weights is n,

$$w_i^{(t+1)} \leftarrow \left(\sum_{\ell=1}^{L_j} \mathbf{1}(X_{ji} = \ell) r_{j\ell}^{(t+1)} \right) \Big/ \underbrace{\left(\frac{1}{n} \sum_{i=1}^{n} \sum_{\ell=1}^{L_j} \mathbf{1}(X_{ji} = \ell) r_{j\ell}^{(t+1)} \right)}_{\text{Normalizing factor}} \tag{A.2}$$

this will always set the proportion to the target because now for all $\ell \in \{1, ..., L_j\}$,

$$\frac{1}{n} \sum_{i=1}^{n} w_i^{(t+1)} \mathbf{1}(X_{ji} = \ell) = \frac{\pi_{j\ell}}{\widehat{\pi}_{j\ell}^{(t+1)}} \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(X_{ji} = 1) = \pi_{k\ell} \tag{A.3}$$

Convergence However, this does not necessarily guarantee that the weight $w^{(t+1)}$ balances the sample moments for other groups k'. The algorithm cycles through each $j \in \{1, ...K\}$. This constitutes one iteration.

At the each of iteration, the algorithm assess total balance on typically a squared loss:

$$SS^{(t+1)} = \sum_{j=1}^{K} \sum_{\ell=1}^{L_j} \left(\frac{1}{n} \sum_{i=1}^{n} (w_i^{(t+1)} \mathbf{1}(X_{ji} = \ell)) - \pi_{j\ell} \right)$$

and stops if the sums of square is below some threshold, e.g. $SS^{(t+1)} < 10^{-10}$.

Limitations There are two limitations to this algorithm. The first is that users typically set a cap on the weights, e.g. $w_i < 10$. The CCES caps weights at 15 for the common content ($n = 60,000$) and at 7 for team modules ($n = 1,000$). The second is easily expected from the fact that the main function has $\widehat{p}_{j\ell}$ in the denominator, meaning it cannot be zero. This comes down to

$$\sum_{i=1}^{n} \mathbf{1}(X_{j\ell} = 1) > 0, \forall j, \ell$$

i.e. that there is at least one observation for each level, in all groupings. This is essentially a limit on the number of dimensions one can match on. For example, if there are no Asian Americans over 65 years old in a sample of Vermont respondents, then the rim weighting cannot weight to race-age bin interactions.

B Implementation of Rim Weights

We use the 2017 ACS ($N = 3,210,525$). This includes $267,971$ non-citizens, but we keep them in for now because they are part of the ACS calibrated counts. The covariates are those mentioned at the beginning of this memo: gender, age, race, and education.

We separate states into three tiers based on a data-availability basis:

1. Large states, where all six pairwise interaction of the categories has no zero-cells in the CCES. These include the top seven largest states in the CCES. These states are California, Texas, Florida, New York, Ohio, Pennsylvania, and Illinois. We calculate rim weights by the *marginals and the interactions*.
2. Medium states, where at least pairwise interaction has at least one zero cell, but where all the marginals are populated in the CCES. We calculate rim weights here by the *marginals only*.
3. Small states, where even some states have missing cells. This happens only in race, and those states are Alaska, Delaware, and North Dakota. We calculate the rim weights here by *marginals only, ignoring the zero cell altogether*.

We conducted this using the ACS with sampling weights and the ACS without sampling weights. The resulting rim weights for each state are called `weight_st_wacs` and `weight_st_uacs`, respectively.