svysim:

Creating Realistic Simulations of (Biased) Survey Data

Shiro Kuriwaki

June 15, 2020

https://github.com/kuriwaki/svysim

Motivation: Existing simulations of survey response are too simplistic

Simulated data are important for demonstrating predictive accuracy and proper coverage of estimators, but existing simulations ...

- Use continuous predictors, but almost all survey data is categorical
- Assume no multilevel / clustering structure
- Assume no selection bias

(for exceptions, see Kennedy and Gabry)

Package svysim: Realistic simulation, with control over the sampling scheme

- Population (N = 600, 000): CCES Data, expanded using post-stratification weights
 - This is technically not a census, but it has a natural covariance structure and makes the simulation realistic.
- Sample (n = 1,000): Simple Random Sample (SRS), OR a biased sample where the propensity score for population member $i \in \{1,...,N\}$ is determined by a propensity score p_i .
- Then I get a sample by: sample(1:N, size = n, replace = FALSE, prob = Propensity Score;)

Sampling Functions

$$p_i = \text{invlogit}(bX)$$
 where

"High Education": here bX is:

$$= \left\{ -4 + 2\mathsf{Urban}_i + \begin{bmatrix} 1.0 \\ 0.8 \\ 0.7 \\ 0.6 \\ 0.5 \end{bmatrix}^\top \begin{bmatrix} \mathsf{White}_i \\ \mathsf{Black}_i \\ \mathsf{Hispanic}_i \\ \mathsf{Asian}_i \\ \mathsf{All\ Other}_i \end{bmatrix} + \begin{bmatrix} 4.0 \\ 3.0 \\ 1.2 \\ 0.5 \end{bmatrix}^\top \begin{bmatrix} \mathsf{Post\text{-}Grad}_i \\ 4\text{-}Year_i \\ \mathsf{Some\ College}_i \\ \mathsf{HS\ or\ Less}_i \end{bmatrix} + \begin{bmatrix} 4.0 \\ 1.0 \\ 0.4 \\ 0.3 \end{bmatrix}^\top \begin{bmatrix} \mathsf{Follow\ News}_i \\ \mathsf{Sometimes}_i \\ \mathsf{Now\ and\ Then}_i \\ \mathsf{Hardly}_i \end{bmatrix} \right\}$$

where e.g. White, is an indicator variable for whether respondent i is White.

"High Ed + Partisanship" adds the following partisan component to bX:

$$-2 + \begin{bmatrix} 1.25 \\ 0.75 \\ 1 \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} \mathsf{Dem}_i \\ \mathsf{Indep}_i \\ \mathsf{GOP}_i \end{bmatrix}$$

Simulated Outcome

$$Y_i = \frac{1}{10}(-3 + A_i + 0.2B_i + 0.8C_i + u_i)$$

 $Z_i = \text{invlogit}(Y_i)$

where:

$$A_i = 0.5 \log(\mathsf{Age}_i) - 2\mathbb{I}(\mathsf{White\ Male\ Non\text{-}Postgrad}_i) + 2(\mathsf{Follow\ News}_i) + \begin{bmatrix} 0 \\ 1.5 \\ -0.5 \end{bmatrix}^\top \begin{bmatrix} \mathsf{Dem}_i \\ \mathsf{Indep}_i \\ \mathsf{GOP}_i \end{bmatrix}$$

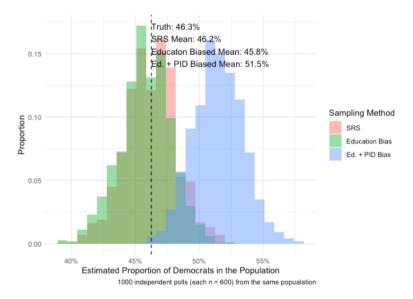
$$B_i \sim \mathsf{Bern}(\pi_{\mathsf{state}[i]}), \ \mathsf{such\ that\ } \mathit{ICC} = 0.15$$

$$C_i \sim \mathsf{Bern}(\pi_{\mathsf{district}[i]}), \ \mathsf{such\ that\ } \mathit{ICC} = 0.3$$

$$u_i \sim \sqcup (0, \mathsf{df} = 5)$$

in which $\pi_{\text{state}[i]}$ is the state-level average of $\text{invlogit}(Y_i)$, and ICC is the intraclass cluster coefficient

Strong error when sampling explicitly is a function of outcome



6 | 6