#### svysim:

# Creating Realistic Simulations of (Biased) Survey Data

Shiro Kuriwaki

August 1, 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:

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

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:

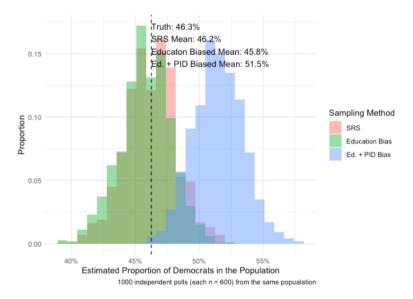
$$A_i = 0.5 \log(Age_i) - 21(White Male Non-Postgrad_i) + 2(Follow News_i)$$

$$+\begin{bmatrix} 0 \\ -0.5 \\ -1.5 \end{bmatrix}^{\top} \begin{bmatrix} \mathsf{Dem}_i \\ \mathsf{Indep}_i \\ \mathsf{GOP}_i \end{bmatrix} + \begin{bmatrix} 0 \\ -1 \\ -3 \\ -3 \\ -4 \\ -5 \end{bmatrix}^{\top} \begin{bmatrix} \mathsf{V. \, Liberal} \\ \mathsf{Liberal} \\ \mathsf{Not \, Sure} \\ \mathsf{Moderate} \\ \mathsf{Conservative} \\ \mathsf{V. \, Conservative} \end{bmatrix}$$

$$B_i \sim \text{Bern}(\pi_{\text{state}[i]})$$
, such that  $ICC = 0.15$   
 $C_i \sim \text{Bern}(\pi_{\text{district}[i]})$ , such that  $ICC = 0.3$   
 $u_i \sim t(0, \text{df} = 5)$ 

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

## Strong error when sampling explicitly is a function of outcome



6 | 6