Generating survey weights using a machine-learning method and the entropy balancing calibration technique

Joseph Kang, Darcy Morris, Patrick Joyce, Isaac Dompreh
US Census Bureau

October 6, 2022

Overview

Raked inverse of probability weighting (IPW) method

Review of IPW

- What is IPW?
- A machine learning-based weights *wgt*

Raking methods based on causal inference

- The entropy balancing technique [Hainmueller, 2012]

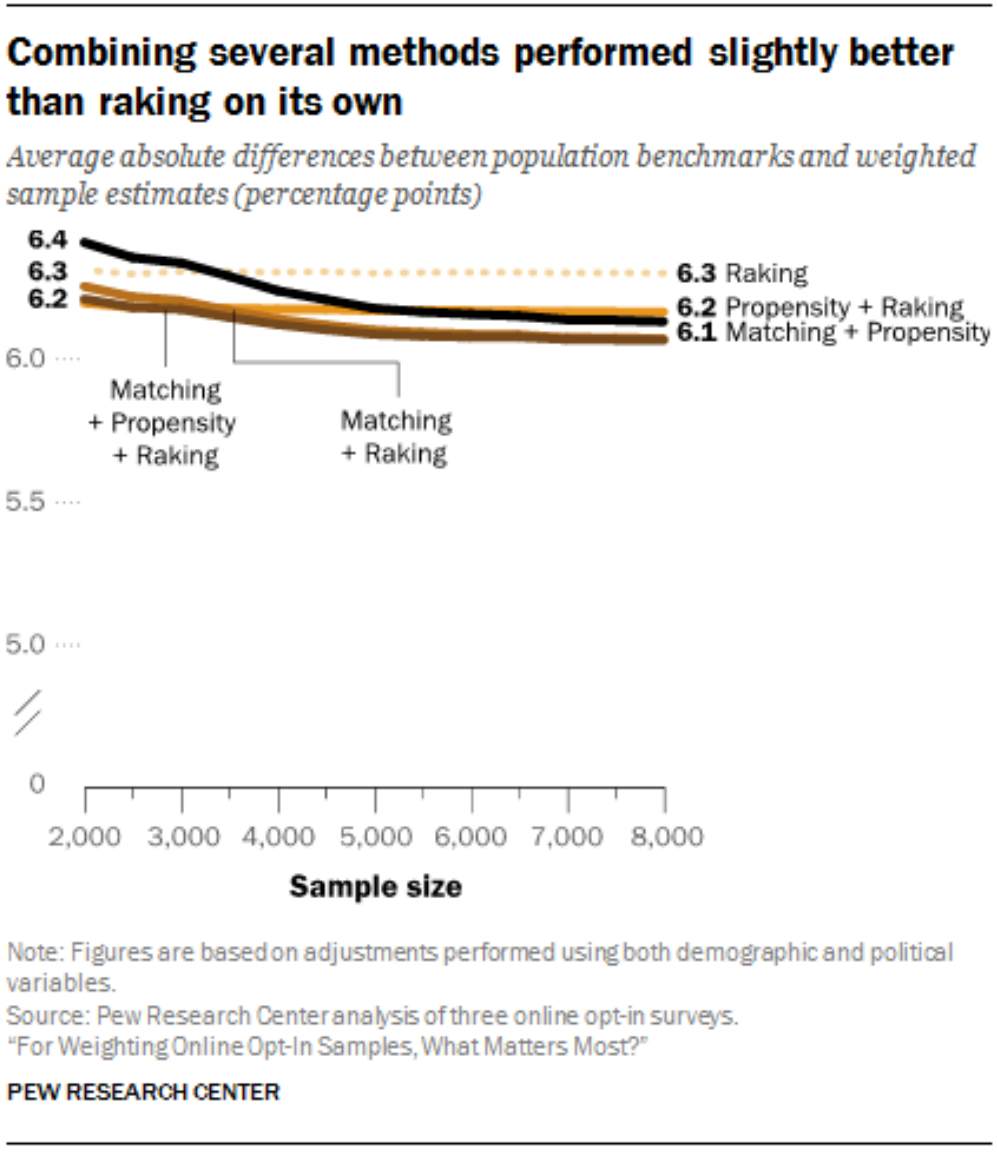
Simulation study

- 1000 simulated samples for ebal
- 1000 simulated samples for gbm
- 1000 simulated samples for balancing results

Summary

Pew Institute's example

- Raking alone was reported to be inferior to combined methods



Review of the IPW weight construction

- Let R denote a subgroup indicator including a binary response condition, X a vector of covariates, and Y a continuous outcome of interest.
- Weights wgt to equate covariate distributions among different populations
$$wgt \times P(X|R=1) = P(X).$$
- IPW $wgt \approx \frac{1}{P(R=1|X)}$.

Interpretation of IPW weights

ID	R	X	Y	$Y1$	$Y0$
1	1	X_1	Y_1	Y_1	??
2	1	X_2	Y_1	Y_1	??
...	...	...	...	...	...
m	1	X_m	Y_m	Y_m	??
$m+1$	0	X_{m+1}	Y_m	??	Y_m
...	...	...	...	...	...
n	0	X_n	Y_n	??	Y_n

n: sample size, m: $\Sigma_{i=1}^{n}(R_i = 1)$, ??: missing Y

- wgt to weight the subgroup $(R = 1)$ data to make it look like the entire data set, i.e., both $R = 1$ and $R = 0$.
- $Y1$: the potential outcome of Y that could have been observed if every unit had $R = 1$.
- $E(Y1) = \frac{\sum wgt \times R \times Y}{\sum wgt}$ and $E(Y0) = \frac{\sum wgt \times (1-R) \times Y}{\sum wgt}$.
- Average causal effect: $E(Y1) - E(Y0)$ under reasonable assumptions [Robins et al., 2000].

IPW for balancing multivariate covariate distributions

- Correct IPW-based weights (*wgt*) are expected to hold:

$$wgt \times P(X|R = 1) = P(X).$$

- X is the p **multivariate** covariates X_1,X_2,...,X_p.
- *wgt* is to weight and make the **multivariate** distribution $P(X|R = 1)$ approximate $P(X)$.
- Causal inference and missing data methodologies that use IPW is based on this notion [Miguel A. Hernán, 2020].
- *wgt* has a clear purpose to make a subset of the data set become representative of a target sample with respect to the **multivariate** distribution of X, *not a univariate* distribution of each variable in $X = X_1$, X_2,...,X_p.

An important condition for *wgt* to correctly balance

- $E(wgt \times R \times X_1) = E(X_1)$.

  The weighted average of the single covariate X_1 should be the average of X_1 for the entire sample ($R = 1$ and $R = 0$).
- *wgt* aims to balance the multivariate covariate distributions between $R = 1$ and the entire sample, but checking this balancing property has been done in an univariate sense [Austin, 2011].
- Even though balancing multivariate distribution of X is a principal task of IPW-based weights *wgt*, it is based on a response propensity model $P(R = 1|X)$, which is a single scalar value and a one-dimensional summary of X.
- Because IPW weights *wgt* require a very 'good' response propensity score, careful modeling is needed.

Machine-learning method: generalized boosted model(GBM)

- IPW weights heavily rely on the correctness of the response propensity model and have been built with machine-learning methods including the GBM.
- R package twang [Cefalu et al., 2021], managed by RAND institute's staff, builds GBM-based responses propensity models in causal inference.
- R package twang automates the process of creating IPW weights with the GBM-based response propensity model.
- GBM can handle high dimensionality, colinearity, nonlinearity, missing data, and complex interactions among covariates.
- GBM does not require an assumption that the logistic regression model needs, and hence it is nonparametric (a black box problem).
- GBM reports the relative influence of each covariates at a percent scale in a nonparametric way.

What if *wgt* is incorrect?

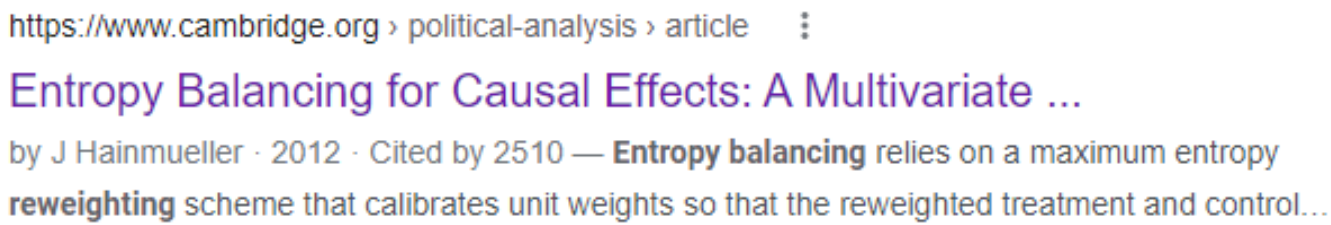
- *wgt* (from IPW) is expected to hold the univariate balancing condition

$$E(wgt \times R \times X_k) = E(X_k),$$

  for k=1, . . . , p.
- If this does not hold, then the response propensity is considered incorrectly specified for *wgt*.
- A calibration is needed for *wgt* to hold the equality above.
- Survey statisticians' raking calibration creates such additional weight to enable the balancing equality.

The entropy balancing (ebal) technique [Hainmueller, 2012]

- Ebal has been extensively used in causal inference



- Ebal aims to attain the p-**univariate** balancing conditions

$$wgt \times P(X_1|R=1) = P(X_1),$$
$$wgt \times P(X_2|R=1) = P(X_2),$$
$$\dots$$
$$wgt \times P(X_p|R=1) = P(X_p),$$

- Recall that the IPW method aims to meet the **multivariate** balancing condition for the multivariate p covariates $X_1, X_2, ..., X_p$

$$wgt \times P(X|R=1) = P(X)$$

The entropy balancing (ebal) technique: continued

- R package ebal [Hainmueller, 2022] minimizes the following objective function to re-weight 'working' IPW weights *wgt* by creating a new weight $wgt_{ebal} = wgt \times a(x)$, where $a(x)$ is an additional factor that helps *wgt* meet univariate balancing properties.

$$\sum_{respondents} \left(wgt_{ebal} log(wgt_{ebal}/wgt) + \sum_{j=1}^{p} \lambda_j \left[wgt_{ebal} X_j - m_j \right] + ... \right)$$

- wgt_{ebal} aims to attain the p univariate balanced distributions, while the IPW-based weights *wgt* aim to attain the multivariate version of balanced distribution of all covariates X.
- Ebal has been extensively used as a raking method for causal inference in Economics.

Kang and Schafer [2007] (KS) simulation study design as a check point to evalute methods

Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data

JDY Kang, JL Schafer - Statistical science, 2007 - projecteuclid.org

When outcomes are missing for reasons beyond an investigator's control, there are two different ways to adjust a parameter estimate for covariates that may be related both to the ...

☆ Save 99 Cite Cited by 1322 Related articles All 18 versions »

- KS designed the following simulation to critically evaluate methodologies for causal inference and missing data analysis.

$$y = 210 + 27.4z_1 + 13.7z_2 + 13.7z_3 + 13.7z_4 + \epsilon,$$
$$\pi = 1 - (1 + exp(-z_1 + 0.5z_2 - 0.25z_3 - 0.1z_4))^{-1},$$

$x_1 = exp(z_1/2)$, $x_2 = z_2/(1 + exp(z_1)) + 10$, $x_3 = (z_1 z_3/25 + 0.6)^3$, $x_4 = (z_2 + z_4 + 20)^2$, and ϵ and z's were from standard normal distributions.

- z's are not available but x's are available to analysts.
- x's are reasonable predictors for y and r in lieu of z.

KS simulation study: continued

- KS design

ID	R	X	Y	$Y1$
1	1	X_1	Y_1	Y_1
2	1	X_2	Y_1	Y_1
...	...	...	...	...
m	1	X_m	Y_m	Y_m
$m+1$	0	X_{m+1}	Y_m	??
...	...	...	...	...
n	0	X_n	Y_n	??

n: sample size, m: $\Sigma_{i=1}^{n}(R_i = 1)$,
??: missing Y

- $Y1$ is a potential outcome that could have been observed if a unit had responded ($R = 1$).
- $E(Y1) = 210$ is the value to be inferred with X's and Y within the subgroup of $R = 1$.

KS study results for 1000 samples

Weights	MBias	VBias	RMSE	MAE	RNG
wgt_ebal	1.881	2.310	2.418	1.894	-1.005:4.966
wgt_gbm1	2.997	1.806	3.285	2.973	0.36:5.666
wgt_gbm2	3.394	1.745	3.642	3.383	0.789:5.972
wgt_gbm1ebal	1.271	1.879	1.869	1.365	-1.342:4.024
wgt_gbm2ebal	0.880	1.754	1.590	1.098	-1.726:3.514
ols	0.725	2.263	1.669	1.159	-2.213:3.693

$bias = 210 - est$

$MBias = \frac{1}{10^3} \sum bias$

$VBias = \frac{1}{10^3 - 1} \sum (bias - MBias)^2$

$RMSE = \sqrt{\frac{1}{10^3} \sum bias^2}$

$MAE = median \text{ of } abs(bias)$

$RNG = 2.5\% - 97.5\%$

- 'ols' indicates ordinal least square regression estimates in fitting a regression for units with $R = 1$ and use its estimated coefficients to predict the entire sample.

Ebal results

Weights	MBias	VBias	RMSE	MAE	RNG
wgt_ebal	1.881	2.310	2.418	1.894	-1.005:4.966
wgt_gbm1ebal	1.271	1.879	1.869	1.365	-1.342:4.024
wgt_gbm2ebal	0.880	1.754	1.590	1.098	-1.726:3.514
ols	0.725	2.263	1.669	1.159	-2.213:3.693

bias = 210 - est; MBias=$\frac{1}{10^3}\sum bias$; VBias = $\frac{1}{10^3-1}\sum(bias - MBias)^2$; RMSE = $\sqrt{\frac{1}{10^3}\sum bias^2}$; MAE = median of abs(bias); RNG = 2.5% - 97.5%

- The simple ols method was reported to be the most unbiased method by Kang and Schafer [2007].
- Ebal's MBias was improved with GBM.
- wgt_gbm2ebal outperformed the ols in RMSE, MAE, and RNG!
- *ebal alone is inferior* to ebal-raked GBM IPW.

GBM results

Weights	MBias	VBias	RMSE	MAE	RNG
wgt_gbm1	2.997	1.806	3.285	2.973	0.36:5.666
wgt_gbm2	3.394	1.745	3.642	3.383	0.789:5.972
wgt_gbm1ebal	1.271	1.879	1.869	1.365	-1.342:4.024
wgt_gbm2ebal	0.880	1.754	1.590	1.098	-1.726:3.514
ols	0.725	2.263	1.669	1.159	-2.213:3.693

bias = 210 - est; MBias=$\frac{1}{10^3}\sum bias$; VBias = $\frac{1}{10^3-1}\sum(bias - MBias)^2$; RMSE = $\sqrt{\frac{1}{10^3}\sum bias^2}$; MAE = median of abs(bias); RNG = 2.5% - 97.5%

- GBM, if with ebal, worked better than itself.
- GBM worked worse when interaction.depth increased from 1 to 2.
- GBM, with ebal, worked better when interaction.depth increased from 1 to 2.

1000 simulated samples for balancing results I

Table: Averages of variables

method	$\bar{X}_1$	$\bar{X}_2$	$\bar{X}_3$	$\bar{X}_4$	$E(Y(1))$
all	1.134	10.000	0.219	401.985	210.023
wgt_ebal	1.134	10.000	0.219	401.985	208.119
wgt_gbm1	1.083	10.020	0.217	403.406	207.003
wgt_gbm2	1.068	10.025	0.218	403.224	206.606
wgt_gbm1ebal	1.134	10.000	0.219	401.985	208.729
wgt_gbm2ebal	1.134	10.000	0.219	401.985	209.120

- GBM methods alone (wgt_gbm1 and wgt_gbm2) did not completely balance covariate distributions.

1000 simulated samples for balancing results II

Table: Averages of standardized difference of variables

method	x1	x2	x3	x4
wgt_ebal	0.394	0.530	0.050	0.443
wgt_gbm1	2.877	3.236	1.331	1.348
wgt_gbm2	4.165	3.894	1.051	1.189
wgt_gbm1ebal	0.440	0.438	0.265	0.426
wgt_gbm2ebal	**0.031**	**0.032**	**0.032**	**0.045**

- A standardized difference:

$$d = \frac{\bar{\Delta}_B}{\sqrt{\sum_{b=1}^{B}(\Delta_b - \bar{\Delta}_B)^2/(B-1)}}, \text{ B=1000}$$

where $\bar{\Delta}_B = \frac{1}{B}\sum_{b=1}^{B}\Delta_b$, Δ_b is $(\bar{x}_{1b} - \bar{x}_b)$, $\bar{x}_{1b}$ indicates the sample mean of covariate among $R = 1$ and $\bar{x}_b$ is for all units in the b^{th} bootstrap (simulation) sample.

- $d < 0.1$ implies that univariate balancing conditions were met by negligible differences in the mean of a covariate between the subgroup with $R = 1$ and all [Austin, 2011].

Summary

- Calibrated machine learning-based IPW weights produced as reasonable estimates as the ols method in inferring the average of an outcome with missing data via the KS simulation study.
- The GBM machine-learning method was used for IPW to meet the multivariate balancing condition.
- The ebal method was used to meet the univariate balancing conditions.
- Either GBM alone or ebal alone is inferior to the combined method.
- The combined method used 1) GBM to meet the multivariate condition, which is hard to check, and 2) ebal as 'insurance' to complete the univariate conditions.

References

Jens Hainmueller. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, 20(1):25–46, 2012.

Reducing bias on benchmarks. `https://www.pewresearch.org/methods/2018/01/26/reducing-bias-on-benchmarks/`. Accessed: 2022-08-30.

J. M. Robins, M. A. Hernán, and B. Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560, Sep 2000.

James M. Robins Miguel A. Hernán. *Causal Inference: What If*. Chapman & Hall/CRC, Boca Raton, 2020.

P. C. Austin. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behav Res*, 46(3):399–424, May 2011.

Matthew Cefalu, Greg Ridgeway, Dan McCaffrey, Andrew Morral, Beth Ann Griffin, and Lane Burgette. *twang: Toolkit for Weighting and Analysis of Nonequivalent Groups*, 2021. URL `https://CRAN.R-project.org/package=twang`. R package