Survey calibration for causal inference: a simple method to balance covariate distributions

Beręsewicz Maciej[1,2]

(1) Department of Statistics, Poznań University of Economics and Business, Poland

(2) Centre for the Methodology of Population Studies, Statistical Office in Poznań, Poland

Scan me

Introduction

- This paper proposes a simple, yet powerful, method for balancing distributions of covariates for causal inference based on observational studies.
- The method makes it possible to balance an arbitrary number of quantiles (e.g., medians, quartiles, or deciles) together with means if necessary.
- The proposed approach is based on the theory of calibration estimators (Deville and Särndal 1992), in particular, calibration estimators for quantiles, proposed by Harms and Duchesne (2006).
- Valid estimates can be obtained by drawing on existing asymptotic theory.
- An illustrative example of the proposed approach is presented for the entropy balancing method (for more see the working paper).
- Results indicate that the method efficiently estimates average treatment effects on the treated (ATT), the average treatment effect (ATE), the quantile treatment effect on the treated (QTT) and the quantile treatment effect (QTE).
- An open source software, implementing proposed methods, is available (cf. the `Weighit` package).

Calibration for quantiles (survey methodology)

General idea of calibration in survey methodology:

- Let $U = \{1, \ldots, N\}$ denote the target population consisting of N labelled units.
- Each unit k has an associated vector of auxiliary variables $\boldsymbol{x}$ and the target variable y, with their corresponding values $\boldsymbol{x}_k$ and y_k, respectively.
- s denotes a probability sample of size n and $d_k = 1/\pi_k$ is a design weight and π_k is the first-order inclusion probability of the k-th element of the population U.
- In most applications the goal is to estimate a finite population mean $\bar{\tau}_y = \tau_y/N$ of the variable of interest y, where U is the population of size N.
- The well-known estimator of a finite population total is the Horvitz-Thompson estimator $\hat{\tau}_{y\pi} = \sum_{k\in s} d_k y_k$.
- Let $\boldsymbol{x}_k^{\circ}$ be a J_1-dimensional vector of auxiliary variables for which the population totals are assumed to be known.
- In most cases in practice the d_k weights do not reproduce known population totals for auxiliary variables $\boldsymbol{x}_k^{\circ}$.
- The main idea of calibration is to look for new calibration weights w_k which are as close as possible to original design weights d_k and reproduce known population totals $\tau_{\boldsymbol{x}}$ exactly.

Calibration for quantiles:

- We assume that

$$\boldsymbol{Q}_{\boldsymbol{x},\alpha} = \left(Q_{x_1,\alpha}, \ldots, Q_{x_{J_2},\alpha}\right)^T \tag{1}$$

is a vector of known population quantiles of order α for a vector of auxiliary variables $\boldsymbol{x}_k^*$, where $\alpha \in (0,1)$ and $\boldsymbol{x}_k^*$ is a J_2-dimensional vector of auxiliary variables.

- A calibration estimator of quantile $Q_{y,\alpha}$ of order α for variable y is defined as

$$\hat{Q}_{y,cal,\alpha} = \hat{F}^{-1}_{y,cal}(\alpha), \tag{2}$$

where a vector $\boldsymbol{w} = (w_1, \ldots, w_n)^T$ is a solution of optimization problem

$$D(\boldsymbol{d}, \boldsymbol{v}) = \sum_{k\in s} d_k G\left(\frac{v_k}{d_k}\right) \to \min \tag{3}$$

subject to the calibration constraints

$$\sum_{k\in s} v_k = N \tag{4}$$

$$\sum_{k\in s} v_k \boldsymbol{a}_k = \boldsymbol{T}_{\boldsymbol{a}}, \tag{5}$$

where as previously $\boldsymbol{T}_{\boldsymbol{a}} = (N, \alpha, \ldots, \alpha)^T$ and the elements of the vector $\boldsymbol{a}_k = \left(1, a_{k1}, \ldots, a_{kJ_2}\right)^T$ are given by

$$a_{kj} = \begin{cases} N^{-1}, & x_{kj} \leq L_{x_j,s}\left(Q_{x_j,\alpha}\right), \\ N^{-1}\beta_{x_j,s}\left(Q_{x_j,\alpha}\right), & x_{kj} = U_{x_j,s}\left(Q_{x_j,\alpha}\right), \\ 0, & x_{kj} > U_{x_j,s}\left(Q_{x_j,\alpha}\right), \end{cases} \tag{6}$$

with $j = 1, \ldots, J_2$ or alternatively

$$a_{kj} = \frac{1}{1 + \exp\left(-2l\left(x_{kj}^* - Q_{x_j^*,\alpha}\right)\right)} \frac{1}{N}, \tag{7}$$

Distributional balancing for causal inference

- This poster limits the theory and results to entropy balancing; for more results, see the paper.
- Hainmueller (2012) proposed entropy balancing (EB) to reweight the control group to the known characteristics of the treatment group to estimate ATT and QTT. This method can be summarised as follows:

$$\begin{aligned} \max_v H(v) &= -\sum_{k\in s_0} v_k \log\left(v_k/d_k\right) \\ \text{s.t.} & \sum_{k\in s_0} v_k x_{kj}^{\circ} = m_j \text{ for } j \in 1, \ldots, J_1, \\ & \sum_{k\in s_0} v_k = 1 \text{ and } v \geq 0 \text{ for all } k \in s_0, \end{aligned} \tag{8}$$

where s_0 and n_0 is the control group and its size, v_k is defined as previously, $d_k > 0$ is the base weight for unit k set to e.g. $d_k = 1/n_0$ and m_j is the mean of the x_j°-th covariate in the treatment group. As in the case of calibration, ω_k are solutions to (8).

- Instead of using known or estimated population totals $\boldsymbol{Q}_{\boldsymbol{x}^*,\alpha}$, we can use treatment group quantiles denoted by $\boldsymbol{q}_{\boldsymbol{x}^*,\alpha} = \left(q_{x_1^*,\alpha}, \ldots, q_{x_{J_2}^*,\alpha}\right)^T$, where the same α is applied for all $\boldsymbol{x}^*$ variables and the definition of the vector $\boldsymbol{a}_k = \left(1, a_{k1}, \ldots, a_{kJ_2}\right)^T$ changes to

$$a_{kj} = \frac{1}{1 + \exp\left(-2l\left(x_{kj}^* - q_{x_j^*,\alpha}\right)\right)} \frac{1}{n_1},$$

where n_1 is the size of the treatment group.

- The proposal, which leads to distributional entropy balancing (hereinafter DEB), consists of extending the original idea by adding additional constraint(s) on the weights on $\boldsymbol{a}_k$, as presented below where the same α is applied for all $\boldsymbol{x}_j^*$ $j = 1, ..., J_2$ variables

$$\begin{aligned} \max_v H(v) &= -\sum_{k\in s_0} v_k \log\left(v_k/d_k\right), \\ \text{s.t.} & \sum_{k\in s_0} v_k x_{kj}^{\circ} = m_j \text{ for } j \in 1, \ldots, J_1, \\ & \sum_{k\in s_0} v_k a_{kj} = \frac{\alpha_j}{n_1} \text{ for } j \in 1, \ldots, J_2, \\ & \sum_{k\in s_0} v_k = 1 \text{ and } v \geq 0 \text{ for all } k \in s_0. \end{aligned} \tag{9}$$

- This approach can be easily extended for vector $\boldsymbol{\alpha}$, say quartiles $(0.25, 0.5, 0.75)$ or deciles $(0.1, \ldots, 0.9)$.
- The population average treatment is given by $\tau_{\text{PATT}} = \mathbb{E}[Y(1) \mid \mathcal{D} = 1] - \mathbb{E}[Y(0) \mid \mathcal{D} = 1]$ where the counterfactual mean may be estimated by $\sum_{k\in s_0} \omega_k y_k / \sum_{k\in s_0} \omega_k$, where ω_k is the solution to (9).

An example based on Austrian SES

- This study uses Synthetic SES survey data from the `laeken` package, and the `WeightIt` and the `cobalt` packages for causal inference.
- Treatment assignment: females vs males.
- Target variable: hourly earnings.
- Control variables: age (cat), location (cat), education (cat), contract (cat), lengthService (cat), weeks (num) and hoursPaid (num).
- This study compares standard entropy balancing (EB) and distributional entropy balancing (DEB; balanced at 0.10, 0.25, 0.50, 0.75, 0.90 quantiles).
- This study verifies whether DEB improves balance using: Adjusted Variance Ratio (V.Ratio.Adj) and Kolmogorov-Smirnov Statistic (KS.Adj), and how it affects effective sample size (ESS)

Table 1: Variable balance measures

Variable	Measure	EB	DEB
weeks	V.Ratio.Adj	1.0604	**0.9741**
	KS.Adj	0.0226	**0.0120**
hoursPaid	V.Ratio.Adj	**0.7526**	0.9764
	KS.Adj	0.1025	**0.0718**
ESS (control)		**1540.07**	1568.35

Table 2: Hourly wage gender gap estimates based on the synthetic SES using entropy and distributional entropy balancing (standard errors based on M-estimation)

Method	Estimate	Std.Error	Statistic	S-value	CI Lower	CI Upper
EB	-2.48	0.18	-13.53	136.08	-2.84	-2.12
DEB	-2.72	0.21	-13.20	129.71	-3.12	-2.31

Notes: for S-values (Shannon transform) see:
`https://marginaleffects.com/bonus/svalues.html`

References (selected)

- Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. Journal of the American statistical Association, 87(418):376–382.
- Hainmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. Political analysis, 20(1):25–46.
- Harms, T. and Duchesne, P. (2006). On calibration estimation for quantiles. Survey Methodology, 32(1):37–52.

Acknowledgements

This work was supported by the National Science Center grant *Towards census-like statistics for foreign-born populations – quality, data integration and estimation* (NCN OPUS 20 2020/39/B/HS4-/00941).

I would like to thank Noah Greifer (Harvard University; Institute for Quantitative Social Science) who, within a week of uploading the first version of this paper to the arXiv platform (arXiv:2310.11969), implemented the proposed approach in the `WeightIt` package for entropy balancing, non-parametric covariate balancing propensity score, optimization-based, and energy balancing methods.

Claude Sonnet 4 was used to proofreading and as a "second" reviewer.