

Szkolenie z estymacji na podstawie prób nielosowych z wykorzystaniem pakietu *nonprobsvy* w języku R

dr Maciej Beręsewicz, prof. UEP

Katedra Statystyki, Uniwersytet Ekonomiczny w Poznaniu

Ośrodek Statystyki Małych Obszarów, Urząd Statystyczny w Poznaniu

lic. Łukasz Chrostowski

Uniwersytet im. Adama Mickiewicza

Spis treści

1 Wprowadzenie

- O szkoleniu
- O nas
- O grancie
- O pakiecie

2 Próby losowe i nielosowe

- Jaka jest różnica między próbą losową a nielosową?
- Literatura

3 Metody estymacji dla prób nielosowych

- Metody quasi-randomizacyjne
- Metody oparte na modelu
 - Ogólna idea
 - Podwójnie odporne estymatory

4 Podsumowanie

Spis treści

1 Wprowadzenie

- O szkoleniu
- O nas
- O grancie
- O pakiecie

2 Próby losowe i nielosowe

3 Metody estymacji dla prób nielosowych

4 Podsumowanie

O szkoleniu

O szkoleniu

- Część 1: 11:00-11:50 (teoria)
- Część 2: 12:00-12:50 (praktyka)
- Pytania: proszę zadawać na czacie

O nas

- **Maciej Beręsewicz** – Katedra Statystyki, Uniwersytet Ekonomiczny w Poznaniu; Ośrodek Statystyki Małych Obszarów, Urząd Statystyczny w Poznaniu.
- **Łukasz Chrostowski** - student, Wydział Matematyki i Informatyki, Uniwersytet im. Adama Mickiewicza.

O grancie

O grancie

- Prace nad pakietem i metodyką finansowane są z grantu NCN: *Statystyka cudzoziemców bez spisu powszechnego - jakość, integracja danych i estymacja finansowanego* (2020/39/B/HS4/00941).
- **Głównym celem projektu** jest opracowanie metod estymacji wielkości i charakterystyk populacji cudzoziemców w Polsce w oparciu o dostępne źródła danych.
- Opis: https://projekty.ncn.gov.pl/index.php?projekt_id=501831

O grancie

O grancie

The screenshot shows the GitHub organization profile for 'ncn-foreigners'. The header includes the GitHub icon, the organization name 'ncn-foreigners', a search bar, and navigation links for Overview, Repositories (34), Projects (2), Packages, Teams, People (6), and Settings. The main content area displays the organization's logo, name, description ('Project "Towards census-like statistics for foreign-born populations"'), follower count (4), and location (Poland). Below this, four repositories are listed under 'Pinned': 'outputs' (Public, 1 star, 1 fork), 'singleRcapture' (Public, 3 stars, 1 fork), 'nonprobsvy' (Public, 10 stars, 3 forks), and 'software-tutorials' (Public, 3 stars, 1 fork). Each repository card includes a brief description and its status as Public.

ncn-foreigners

Project "Towards census-like statistics for foreign-born populations"

4 followers Poland

Pinned

Customize pins

outputs Public

Repository with the list of project's outputs

1 star 1 fork

singleRcapture Public

Repository for single source capture-recapture models

1 R 3 stars 1 fork

nonprobsvy Public

An R package for modern methods for non-probability surveys

1 R 10 stars 3 forks

software-tutorials Public

Repo with tutorials about the software that are developed in this project

HTML

O grancie – single-source capture-recapture models

singleRcapture 0.2.1.1 Reference Changelog

Search for

Overview

Capture-recapture type experiments are used to estimate the total population size in situations when observing only a part of such population is feasible. In recent years these types of experiments have seen more interest.

Single source models are distinct from other capture-recapture models because we cannot estimate the population size based on how many units were observed in two or three sources which is the standard approach.

Instead in single source models we utilize count data regression models on positive distributions (i.e. on counts greater than 0) where the dependent variable is the number of times a particular unit was observed in source data.

This package aims to implement already existing and introduce new methods of estimating population size from single source to simplify the research process.

Links

[View on CRAN](#)

[Browse source code](#)

[Report a bug](#)

License

[Full license](#)

[MIT + file LICENSE](#)

Citation

[Citing singleRcapture](#)

Developers

Piotr Chlebicki

Author, maintainer

O grancie - kalibracja dla kwantyli

jointCalib 0.1.2 Reference Articles ▾ Changelog

Search for

Overview

Details

A small package for joint calibration of totals and quantiles (see [Beręsewicz and Szymkowiak \(2023\)](#) working paper for details). The package combines the following approaches:

- Deville, J. C., and Särndal, C. E. (1992). [Calibration estimators in survey sampling](#). Journal of the American statistical Association, 87(418), 376-382.
- Harms, T. and Duchesne, P. (2006). [On calibration estimation for quantiles](#). Survey Methodology, 32(1), 37.
- Wu, C. (2005) [Algorithms and R codes for the pseudo empirical likelihood method in survey sampling](#), Survey Methodology, 31(2), 239.
- Zhang, S., Han, P., and Wu, C. (2023) [Calibration Techniques Encompassing Survey Sampling, Missing Data Analysis and Causal Inference](#), International Statistical Review 91, 165–192.

Links

[View on CRAN](#)

[Browse source code](#)

[Report a bug](#)

License

[GPL-3](#)

Citation

[Citing.jointCalib](#)

Developers

Maciej Beręsewicz

Author, maintainer 

Dev status

9 / 60

O grancie - blokowanie na podstawie ANN

blocking 0.1.0 Reference Articles ▾ Changelog

Search for

Overview

Description

An R package that aims to block records for data deduplication and record linkage (a.k.a. entity resolution) based on [approximate nearest neighbours algorithms \(ANN\)](#) and graphs (via the `igraph` package).

Currently supports the following R packages that binds to specific ANN algorithms:

- [rnnedcent](#) (default, very powerful, supports sparse matrices),
- [RcppHNSW](#) (powerful but does not support sparse matrices),
- [RcppAnnoy](#),
- [mlpack](#) (see `mlpack::lsh` and `mlpack::knn`).

The package also supports integration with the [recln2](#) package via `blocking::pair_ann` function.

Links

[Browse source code](#)

[Report a bug](#)

License

[GPL-3](#)

Citation

[Citing blocking](#)

Developers

Maciej Beręsewicz

Author, maintainer 

Dev status

 R-CMD-check passing

Po co kolejny pakiet?

- Pakiety do prób nielosowych: NonProbEst, WeightIt lub GJRM.
- Ograniczona integracja z pakietem *survey*.
- Brak implementacji aktualnych rozwiązań przedstawionych w literaturze.
- Stworzenie narzędzia, które pozwala w spójny sposób zastosować różne estymatory.
- Pakiet jest rozwijany więc zachcemy do testowania i komentowania oraz śledzenia na github.

Spis treści

1 Wprowadzenie

2 Próby losowe i nielosowe

- Jaka jest różnica między próbą losową a nielosową?
- Literatura

3 Metody estymacji dla prób nielosowych

4 Podsumowanie

Jaka jest różnica między próbą losową a nielosową?

Co GPT + Dall-e o tym myśli?



"A diagram of a probability sample survey"

 Image Creator from Designer

Powered by DALL·E 3



"A diagram of a non probability sample survey"

Obsługiwane przez DALL·E 3

Jaka jest różnica między próbą losową a nielosową?

Co GPT + Dall-e o tym myśli?



"A non probability sample of people"

Obsługiwane przez DALL-E 3



"A probability sample of people"

Obsługiwane przez DALL-E 3

Jaka jest różnica między próbą losową a nielosową?

Definicje

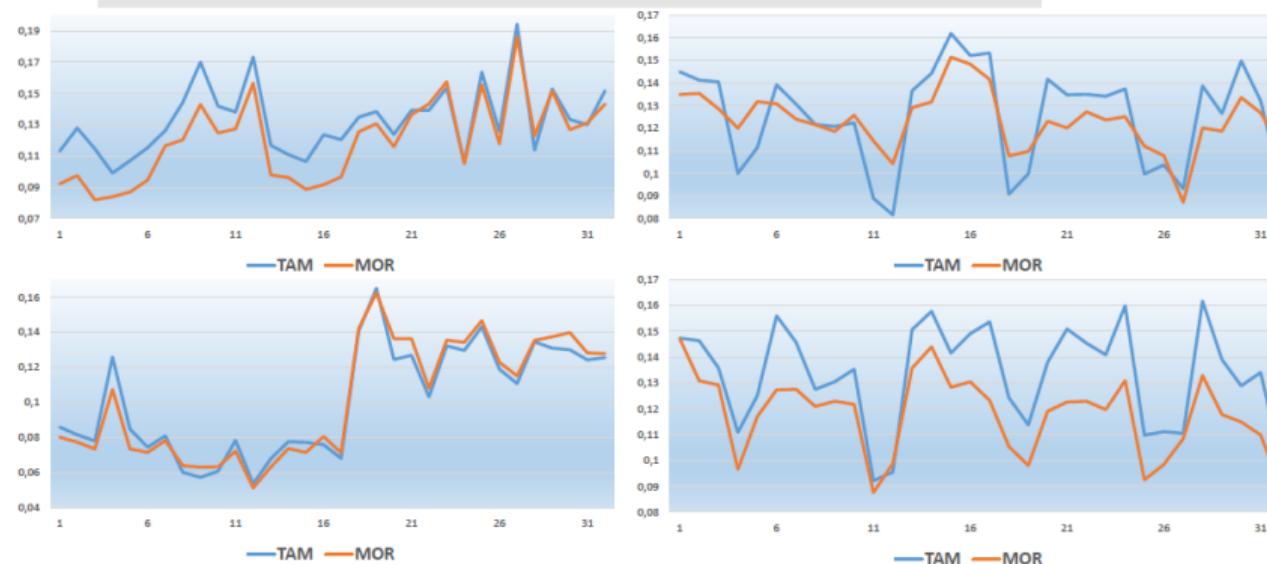
W skrócie:

- **Próba losowa:** sami ustalamy schemat losowania i znamy prawdopodobieństwa inkluzji.
- **Próba nielosowa:** nie znamy procesu generowania danych i nie znamy prawdopodobieństw inkluzji.

Jaka jest różnica między próbą losową a nielosową?

Przykład – pomiar oglądalności

Przykładowe udziały – TAM (|) vs MOR (||)



Rysunek 1: Porównanie badania NAM i Netii. Źródło: Wirtualne Media

Jaka jest różnica między próbą losową a nielosową?

Przykład - CBOP



Język: PL



Kontrast:



Czcionka:



Wsparcie:

[Oferty pracy, staże i praktyki](#)[Kalendarz targów, giełd i szkoleń](#)[Wyszukiwanie pracowników](#)[Zaloguj się](#)[Zarejestruj się](#)Jesteś tutaj: [CBOP](#) > Oferty pracy, staże i praktykiLiczba propozycji: **22 239**, w tym w urzędach pracy**18 796** | Ofert pracy**52 582** | Wolnych miejsc pracy

Wpisz nazwę stanowiska

Wpisz nazwę lokalizacji lub kod pocztowy

+ 0 km

Szukaj

Wyszukiwanie zaawansowane

Sortowanie

Data dodania



Poziom

szczegółowości

Niski



Pozycji na stronie

10



Strona

1

z 2224 następna

WYBIERZ KRYTERIA

STANOWISKO

MIEJSCE PRACY

RODZAJ UMOWY

PRACODAWCA

DOSTĘPNA OD

[SPRZEDAWCA](#)Bytom,
śląskie

Umowa o pracę

kontakt przez PUP

dzisiaj

WYBRANE KRYTERIUM

Jaka jest różnica między próbą losową a nielosową?

Przykład - Banki

Bank	Liczba klientów indywidualnych		
	III kw. 2020	II kw. 2020	III kw. 2019
PKO BP i Inteligo	10 508 000	10 465 700	10 401 000
Bank Pekao	5 434 134	5 388 766	5 349 673
Santander Bank Polska	4 743 041	4 698 385	4 610 781
Alior Bank i TMUB	4 278 399	4 211 010	4 075 953
ING Bank Śląski	4 215 000	4 133 000	4 288 000
mBank	4 099 820	4 086 000	3 979 263
Bank Millennium	3 859 084	3 810 561	2 693 843

Jaka jest różnica między próbą losową a nielosową?

Porównanie

Tabela 1: Próby losowe, a nielosowe

Czynnik	Próba losowa	Próba nielosowa
Dobór	Schemat losowania	Auto-selekcja
Pokrycie	Zwykle dobre	Pewne grupy są wykluczone
Obciążenie	Zwykle mniejsze	duże, lub bardzo duże
Wariancja	Zwykle większa	Mała, lub bardzo mała
Koszt	Duży lub bardzo duży	Zwykle nieduży

Próby losowe i kalibracja

Marcin Szymkowiak

Podejście kalibracyjne w badaniach społeczno-ekonomicznych

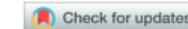
Statistical Science

Statistical Science
2017, Vol. 32, No. 2, 249–264
DOI: 10.1214/16-STS598
© Institute of Mathematical Statistics, 2017

Inference for Nonprobability Samples

Michael R. Elliott and Richard Valliant

Abstract. Although selecting a probability sample has been the standard for decades when making inferences from a sample to a finite population, incentives are increasing to use nonprobability samples. In a world of “big data”, large amounts of data are available that are faster and easier to collect than are probability samples. Design-based inference, in which the distribution for inference is generated by the random mechanism used by the sampler, cannot be used for nonprobability samples. One alternative is quasi-randomization in which pseudo-inclusion probabilities are estimated based on covariates available for samples and nonsample units. Another is superpopulation modeling for the analytic variables collected on the sample units in which the model is used to predict values for the nonsample units. We discuss the pros and cons of each approach.



Doubly Robust Inference With Nonprobability Survey Samples

Yilin Chen, Pengfei Li, and Changbao Wu

Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON, Canada

ABSTRACT

We establish a general framework for statistical inferences with nonprobability survey samples when relevant auxiliary information is available from a probability survey sample. We develop a rigorous procedure for estimating the propensity scores for units in the nonprobability sample, and construct doubly robust estimators for the finite population mean. Variance estimation is discussed under the proposed framework. Results from simulation studies show the robustness and the efficiency of our proposed estimators as compared to existing methods. The proposed method is used to analyze a nonprobability survey sample collected by the Pew Research Center with auxiliary information from the Behavioral Risk Factor Surveillance System and the Current Population Survey. Our results illustrate a general approach to inference with nonprobability samples and highlight the importance and usefulness of auxiliary information from probability survey samples. Supplementary materials for this article are available online.

ARTICLE HISTORY

Received May 2018
Accepted September 2019

KEYWORDS

Design-based inference;
Inclusion probability; Missing
at random; Propensity score;
Regression modeling;
Variance estimation



Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2020)

Doubly robust inference when combining probability and non-probability samples with high dimensional data

Shu Yang,

North Carolina State University, Raleigh, USA

Jae Kwang Kim,

Iowa State University, Ames, USA

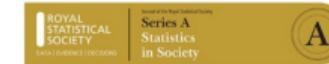
and Rui Song

North Carolina State University, Raleigh, USA

Received: 8 January 2020

Accepted: 20 March 2021

DOI: 10.1111/rssc.12696

ORIGINAL ARTICLE

Combining non-probability and probability survey samples through mass imputation

Jae Kwang Kim¹ | Seho Park² | Yilin Chen³ | Changbao Wu³

¹Department of Statistics, Iowa State University, Ames, IA 50011, USA

²Department of Biostatistics, Indiana University School of Medicine,

Abstract

Analysis of non-probability survey samples requires auxiliary information at the population level. Such information

Survey Methodology

Survey Methodology, December 2022
Vol. 48, No. 2, pp. 283-311
Statistics Canada, Catalogue No. 12-001-X

283

Statistical inference with non-probability survey samples

Changbao Wu¹

Abstract

We provide a critical review and some extended discussions on theoretical and practical issues with analysis of non-probability survey samples. We attempt to present rigorous inferential frameworks and valid statistical procedures under commonly used assumptions, and address issues on the justification and verification of assumptions in practical applications. Some current methodological developments are showcased, and problems which require further investigation are mentioned. While the focus of the paper is on non-probability samples, the essential role of probability survey samples with rich and relevant information on auxiliary variables is highlighted.

Key Words: Auxiliary information; Bootstrap variance estimator; Calibration method; Doubly robust estimator; Estimating equations; Inverse probability weighting; Model-based prediction; Poststratification; Pseudo likelihood; Propensity score; Quota survey; Sensitivity analysis; Variance estimation.

Literatura (wybrana)

- Baker, R, J Michael Brick, Nancy A Bates, Mike Battaglia, Mick P Couper, Jill A Dever, Krista J Gile, and Roger Tourangeau (2013). Summary Report of the AAPOR task force on non-probability sampling. *Journal of Survey Statistics and Methodology* 1, pp. 90–143.
- Bethlehem, J. (2010). Selection Bias in Web Surveys. *International Statistical Review*, 78(2), 161–188.
doi:10.1111/j.1751-5823.2010.00112.x.
- Bethlehem, J., and Biggignandi, S. (2012). *Handbook of Web Surveys*, John Wiley and Sons, Inc. doi:10.1086/318641.
- Callegaro M., Baker R., Bethlehem J., Göritz A.S., Krosnick J.A., Lavrakas P. J. (2014) *Online Panel Research A Data Quality Perspective*, Wiley.
- Lee, S. (2006). Propensity score adjustment as a weighting scheme for volunteer panel web surveys. *Journal of Official Statistics*, 22(2), 329–349.
- Kim, J. K., and Tam, S. M. (2020). Data integration by combining big data and survey sample data for finite population inference. *International Statistical Review*.
- S. Yang, J.K. Kim, and R. Song (2020). "Doubly Robust Inference when Combining Probability and Non-probability Samples with High-dimensional Data", *Journal of the Royal Statistical Society: Series B*, 82, 445-465.
- J.K. Kim and Z. Wang (2019). "Sampling techniques for big data analysis in finite population inference", *International Statistical Review*, 87, S177-S191.

Spis treści

1 Wprowadzenie

2 Próby losowe i nielosowe

3 Metody estymacji dla prób nielosowych

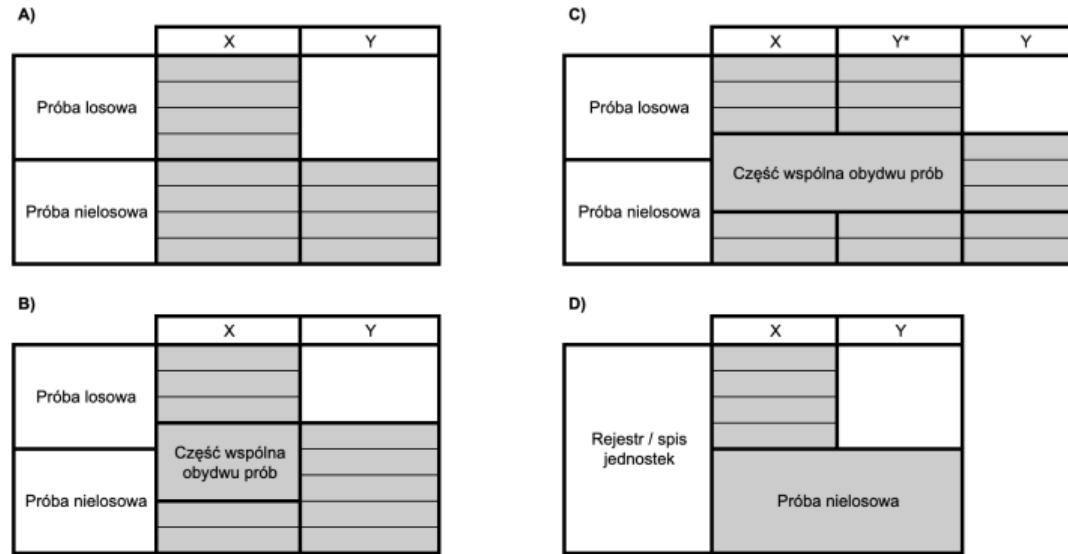
- Metody quasi-randomizacyjne
- Metody oparte na modelu

4 Podsumowanie

Podstawowy układ

Sample		Auxiliary variables X	Target variable Y	Design (d) or calibrated (w) weights
S_A (non-probability)	1	✓	✓	?
	...	✓	✓	?
	n_A	✓	✓	?
S_B (probability)	$n_A + 1$	✓	?	✓
	...	✓	?	✓
	$n_A + n_B$	✓	?	✓

Metody estymacji – rozważmy następujące przypadki



Rysunek 2: Cztery przykładowe przypadki źródeł danych, gdzie celem jest oszacowanie wybranej charakterystyki cechy Y . Cechy X są wspólne, cecha Y^* to tzw. zmienna proxy.

Elliott i Valliant (2017) wyróżniają dwa podejścia:

- **quasi-randomizacyjne** – w której konstruujemy *pseudo-wagi* z wykorzystaniem próby losowej lub znanych (albo estymowanych) wartości globalnych.

	X	W	Y	W*
Próba losowa				
Próba nielosowa				Ostatecznie, do wnioskowania, korzystamy tylko z próby nielosowej

- **oparte na modelu** – w którym zakładamy pewien model.

	X	W	Y	
Próba losowa			$f(Y X) = Y_{pred}$	Ostatecznie, do wnioskowania, korzystamy tylko z próby losowej
Próba nielosowa				

Metody quasi-randomizacyjne

W przypadku metod quasi-randomizacyjnych możemy rozważyć następujące metody:

- **Post-stratyfikację** (ang. post-stratification) – wymaga znajomości wartości globalnych \mathbf{X} (Holt & Smith, 1979)
- **Kalibrację** (ang. calibration) – wymaga znajomości wartości globalnych / średnich \mathbf{X} (Deville & Särndal, 1992)
- **Ważenie przez odwrotność prawdopodobieństwa inkluzji** (ang. inverse probability weighting/propensity score weighting) – wymaga dostępu do danych jednostkowych lub wartości globalnych cech \mathbf{X} (por. Lee, 2006)

Literatura:

- Holt, D., & Smith, T. F. (1979). Post stratification. *Journal of the Royal Statistical Society: Series A (General)*, 142(1), 33-46.
- Deville, J. C., & Särndal, C. E. (1992). Calibration estimators in survey sampling. *Journal of the American statistical Association*, 87(418), 376-382.
- Lee, S. (2006). Propensity score adjustment as a weighting scheme for volunteer panel web surveys. *Journal of official statistics*, 22(2), 329.

Inverse Probability Weighting (IPW) – idea

	X1 (płeć)	X1 (wiek)	Y (słuchanie podcastów)	w (waga)	R	P(R=1 X1, X2)
Próba losowa	M	34	x	4	0	0,65
	M	32	x	2	0	0,85
	K	50	x	5	0	0,35
	K	40	x	10	0	0,23
Próba nielosowa	M	32	Tak	?	1	0,85
	M	34	Nie	?	1	0,65
	K	40	Tak	?	1	0,23
	K	50	Tak	?	1	0,35

w* = 1 / P(waga finalna)
x
x
x
x
1,1765
1,5385
4,3478
2,8571

Ten zbiór wykorzystamy do estymacji

- Mamy dwa zbiory danych (próba losowa i nielosowa).
- Mamy dwie wspólne zmienne (X_1, X_2) oraz jedną obserwowaną tylko w próbie nielosowej (Y).
- W próbie losowej mamy wagi (d).
- Szacujemy prawdopodobieństwo $P(R = 1)$ uwzględniając wspólne zmienne (X_1, X_2).
- Dla jednostek z próby nielosowej przypisujemy wagi $w = 1/P(R = 1|X_1, X_2)$.

Propensity score – założenia (cz. 1)

- Różne nazwy na to samo: propensity score adjustment (PSA), propensity score weighting (PSW), inverse probability weighting (IPW).
- W metodzie *propensity score* zakładamy, że
 - ① potrafimy rozróżnić jednostki, które są i nie są w próbie nielosowej,
 - ② dysponujemy źródłem, które zawiera jednostki nie występujące w źródle nielosowym.
- Przykład:
 - Źródło big data oraz dane dotyczące całej populacji
 - Źródło big data oraz próba losowa, w której znajdują się również jednostki obserwowane w big data.

IPW – oznaczenia

Podstawowe oznaczenia

- Niech $R_i = \{0, 1\}$ oznacza zmienną określającą przynależność do próby nielosowej ($R_i = 1$).
- Niech zmienne \mathbf{x}_i oznacza wektor zmiennych pomocniczych, obserwowanych w obydwu źródłach danych.
- Niech $\pi_i = P(R_i = 1 | \mathbf{x}_i)$ prawdopodobieństwo przynależności do danego źródła.
- π_i nie jest znane jest estymowane na podstawie danych.

IPW – założenia (cz. 2)

Formalnie, założenia metody *propensity score* są następujące:

- Zmienna selekcji / inkluzji R_i oraz badana przez nas cecha y_i są warunkowo niezależne gdy pod uwagę weźmiemy cechy \mathbf{x}_i . Innymi słowy $\pi_i = P(R_i = 1|y, \mathbf{x}_i) = P(R_i = 1|\mathbf{x}_i)$ – **Inaczej:** dobór jest nieinformatywny (w literaturze przedmiotu: missing at random, ignorable).
- Wszystkie jednostki w populacji mają niezerowe prawdopodobieństwo inkluzji do próby nielosowej ($\pi_i > 0$) – **Inaczej:** brak błędów pokrycia.
- Zmienne R_i oraz R_j są niezależne gdy uwzględnimy \mathbf{x}_i dla $i \neq j$ – **Inaczej:** obserwacje są niezależne (m.in. brak duplikatów lub jakichś zmiennych \mathbf{z}_i , których nie obserwujemy).

IPW – estymator

Po oszacowaniu π , wykorzystujemy następujący estymator, który oznaczamy jako IPW/PSW/PSA

$$\hat{\theta}_{IPW/PSW/PSA} = \frac{\sum_{i \in S_A} y_i \hat{\pi}_i^{-1}}{\sum_{i \in S_A} \hat{\pi}_i^{-1}}, \quad (1)$$

gdzie S_A oznacza próbę nielosową, π_i^{-1} to odwrotność sklonności. Uwaga: nie robimy tutaj sumy ponieważ suma π_i^{-1} nie daje nam populacji.

Estymacja wariacji tego estymatora nie jest taka prosta i zostanie pokrótce omówiona na kolejnym slajdzie.

IPW – teoria

Ostatnie, najbardziej aktualne artykuły poświęcone wykorzystaniu metody propensity score weighting dla prób nielosowych:

- Kim, J. K., & Wang, Z. (2019). **Sampling techniques for big data analysis.** *International Statistical Review*, 87, S177-S191
 - estymacja tylko na podstawie próby losowej,
 - propozycja estymatora IPW/PSW/PSA oraz jego wariancji (dla prostych schematów losowania próby losowej).
- Chen, Y., Li, P., & Wu, C. (2020). **Doubly robust inference with nonprobability survey samples.** *Journal of the American Statistical Association*, 115(532), 2011-2021.
 - estymacja IPW/PSW/PSA dla wszystkich obserwacji (oba źródła jednocześnie).
 - propozycja estymatora IPW/PSW/PSA oraz jego wariancji (ogólna postać dla wszelkiego rodzaju schematów losowania próby losowej).

IPW – wady i zalety

Wśród zalet możemy wymienić:

- prosta idea i implementacja,
- jeden zestaw wag dla całego zbioru danych,
- w bardziej (powiedzmy) zaawansowanej wersji wymaga wyłącznie znajomości wartości globalnych / średnich dla cech x_i .

Wśród wad możemy wymienić:

- metoda działa **tylko wtedy gdy poprawnie określmy model** dla π (postać, zmienne),
- estymator PS nie jest efektywny,
- w podstawowej wersji wymaga danych jednostkowych (próba losowa i nielosowa),
- musimy zidentyfikować jednostki między źródłami,
- wagi ($w_i = 1/\hat{\pi}_i$) nie odtwarzają wartości globalnych z populacji (np. liczby kobiet, mężczyzn; charakterystyk podmiotów gospodarczych),
- jeden zestaw wag dla całego zbioru danych.

Podejście oparte na modelu

- W podejściu opartym na modelu zakładamy, że interesuje nas $E(Y|X)$,
- Zakładamy, że model $E(Y|X, R = 1) = E(Y|X) = \mu(y_i|\mathbf{x}_i)$,
- Budujemy model na próbie nielosowej (np. regresja liniowa, logistyczna) i aplikujemy na całą populację (ewentualnie próbę)
- Estymator, w przypadku gdy znamy całą populację na postać:

$$\hat{\theta}_M = \sum_{i \in S_A} y_i + \sum_{i \in U \setminus S_A} \hat{y}_i, \quad (2)$$

gdzie S_A to próba nielosowa, a $U \setminus S_A$ to pozostała część populacji.

Metody oparte na modelu

Podejście oparte na modelu – klasyfikacja metod

- Gdy znamy wszystkie jednostki z populacji,
- Gdy znamy tylko jednostki z próby nielosowej – masowa imputacja (metoda Riversa, metody opracowane przez Jae-Kwang Kim'a i współpracowników).

Podejście oparte na modelu – założenia

- Model z próby losowej możemy przenieść na resztę jednostek (ang. missing at random) – tzw. model dla super-populacji.
- Dysponujemy zmiennymi \mathbf{X} , które są obserwowane w próbie/próbach i populacji.
- Możemy zidentyfikować jednostki między próbą nielosową i populacją.
- Zakładamy, że zmienna Y oraz \mathbf{X} są obserwowane bez błędów.
- Zakładamy brak korelacji między obserwacjami, brak błędu nadreprezentacji itp.

Metody oparte na modelu

Masowa imputacja – dwa podejścia

	X1 (płeć)	X1 (wiek)	Y (słucha podcastów)	w (waga)	R	Y* (przepisane z nielosowej)	
Próba losowa	M	34	?	4	0	Nie	Ten zbiór wykorzystamy do estymacji
	M	32	?	2	0	Tak	
	K	50	?	5	0	Tak	
	K	40	?	10	0	Tak	
Próba nielosowa	M	32	Tak	?	1		Ten zbiór wykorzystamy do estymacji
	M	34	Nie	?	1		
	K	40	Tak	?	1		
	K	50	Tak	?	1		

Rysunek 3: Podejście I: Przepisanie wartości z próby nielosowej

	X1 (płeć)	X1 (wiek)	Y (słuchanie podcastów)	w (waga)	R	\hat{y} (przewidywana)	
Próba losowa	M	34	?	4	0	Nie	Ten zbiór wykorzystamy do estymacji
	M	32	?	2	0	Tak	
	K	50	?	5	0	Tak	
	K	40	?	10	0	Nie	
Próba nielosowa	M	32	Tak	?	1		Ten zbiór wykorzystamy do estymacji
	M	34	Nie	?	1		
	K	40	Tak	?	1		
	K	50	Tak	?	1		

Rysunek 4: Podejście II: wartości przewidywane z modelu zbudowanego na próbie nielosowej

Masowa imputacja – podejście I

W pierwszym podejściu, zaproponowanym przez Rivers (2007), dokonujemy masowej imputacji przez tzw. sample matching, który polega na następujących krokach:

- ① Dla próby nielosowej \mathcal{S}_A oraz losowej \mathcal{S}_B określamy zestaw wspólnych cech \mathbf{X} .
- ② Następnie, dla każdej jednostki $i \in \mathcal{S}_B$ szukamy najbliższej jednostki ze zbioru $k \in \mathcal{S}_A$, tak żeby

$$d(\mathbf{x}_k, \mathbf{x}_i) = \|\mathbf{x}_k - \mathbf{x}_i\| \quad (3)$$

była jak najmniejsza. Możemy w tym celu wykorzystać np. odległość euklidesową. Jednostce i przypisujemy wartość cechy y_k ze zbioru \mathcal{S}_A .

- ③ Po znalezieniu odpowiednich sąsiadów, wyznaczamy estymator. Przykładowo, estymator wartości średniej $\theta = N^{-1} \sum_{i \in U} y_i$ dany będzie:

$$\hat{\theta}_{M1} = \sum_{i \in \mathcal{S}_B} w_i y_i / \sum_{i \in \mathcal{S}_B} w_i. \quad (4)$$

Masowa imputacja – podejście I – UWAGA

Na podstawie pracy: Abadie & Imbens (2006) **Large sample properties of matching estimators for average treatment effects** (Econometrica, 74(1), 235-267) można wykazać, że obciążenie estymatora $\hat{\theta}_{M1}$ rośnie wraz z liczbą zmiennych użytych do obliczenia $d(\mathbf{x}_k, \mathbf{x}_i)$. Dokładnie, ta zależność wyrażona jest następującym wzorem:

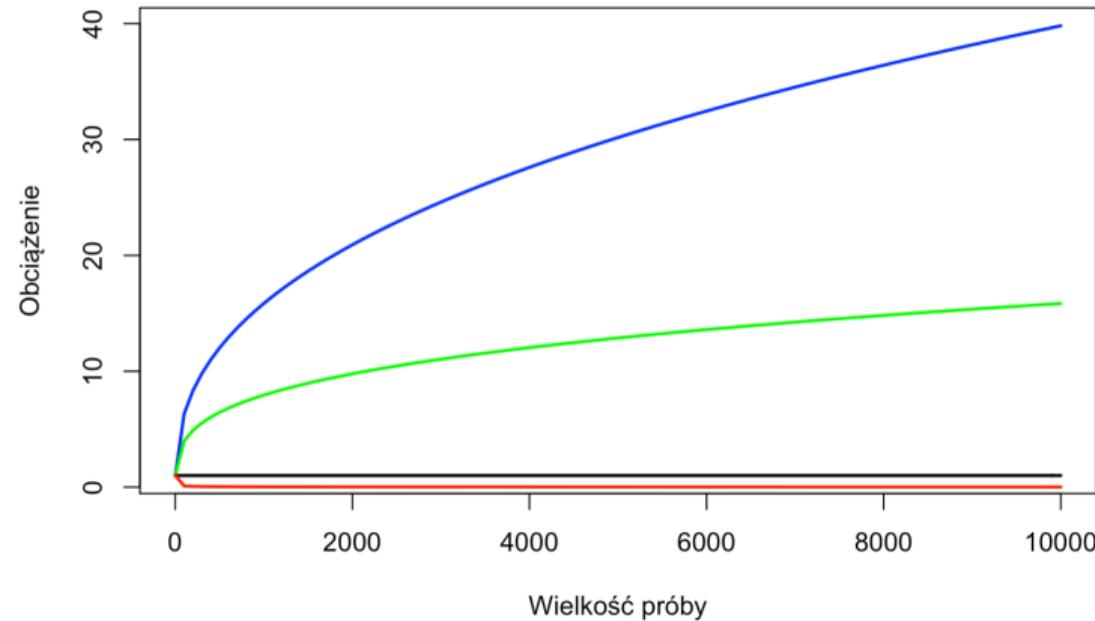
$$\text{Bias}(\hat{\theta}_{M1}) = O_p\left(n^{1/2 - 1/p}\right) \quad (5)$$

gdzie p oznacza liczbę zmiennych (wymiar wektora \mathbf{x}_k), a O_p oznacza notację wielkie-O i określa asymptotyczne tempo wzrostu (od liczby próby losowej n_B).

Uwaga 1: Oznacza to, że estymator $\hat{\theta}_{M1}$ będzie nieobciążony wyłącznie gdy $p = 1$, $O_p(n^{-1/p}) = O_p(n^{-1/1}) = O_p(n^{-1})$ czyli obciążenie będzie mało wraz ze wzrostem próby losowej B .

Uwaga 2: ta zależność dotyczy zarówno prób nielosowych, jak i imputacji czy ekonometrycznego badania wpływu.

Masowa imputacja – podejście I – obciążenie



Rysunek 5: Wizualizacja obciążenia $O_p(n^{1/2-1/p})$. Kolor czerwony: $p=1$; czarny: $p=2$, zielony: $p=5$ i niebieski: $p=10$.

Masowa imputacja – podejście I – Praktyka

Mając na uwadze fakt, że przypisania wartości y_i z próby nielosowej dla jednostek k z próby losowej należy wykorzystać na podstawie wyłącznie jednej zmiennej, stosuje się następujące podejście:

- ① Budujemy model $m(\mathbf{x}_i; \boldsymbol{\beta})$ na próbie nielosowej \mathcal{S}_A otrzymując $\hat{y}_i = m(\mathbf{x}_i; \hat{\boldsymbol{\beta}})$.
- ② Stosujemy model $m(\mathbf{x}; \hat{\boldsymbol{\beta}})$ na próbie losowej \mathcal{S}_B otrzymując $\hat{y}_k = m(\mathbf{x}_k; \hat{\boldsymbol{\beta}})$.
- ③ Dla każdej jednostki $k \in \mathcal{S}_B$ znajdujemy najbliższą jednostkę na podstawie $d(\hat{y}_k, \hat{y}_i)$ i przepisujemy wartość y_i .
- ④ Następnie wyznaczamy estymator

$$\hat{\theta}_{M1} = \sum_{i \in \mathcal{S}_B} w_i y_i / \sum_{i \in \mathcal{S}_B} w_i. \quad (6)$$

To podejście w literaturze nazywa się *predictive mean matching*.

Masowa imputacja – podejście II

- W pracy Kim, Park, Chen i Wu (2021) **Combining Non-probability and Probability Survey Samples Through Mass Imputation**, (Journal of the Royal Statistical Society: Series A) zaproponowano trochę inne podejście ale oparte na solidnych, teoretycznych podstawach.
- Zamiast dokonywać poszukiwania najbliższego sąsiada wykorzystuje się wyłącznie 1 oraz 2 krok z poprzedniego slajdu.
- Estymator wartości średniej dany jest wtedy

$$\hat{\theta}_{M2} = \sum_{i \in S_B} w_i \hat{y}_i / \sum_{i \in S_B} w_i. \quad (7)$$

- Ten sposób nazywamy na potrzeby zajęć podejściem II do masowej imputacji.
- W wyżej wymienionej pracy zaproponowano również estymator wariancji w postaci zlinearyzowanej (konkretny wzór), jak i na podstawie metody bootstrap.

Podwójnie odporne estymatory

- Propensity score weighting działa **wyłącznie wtedy, gdy** model $P(R_i = 1|\mathbf{x}_i)$ jest poprawnie **Wyspecyfikowany** – zakładany model jest poprawny dla całej populacji.
- Dlatego w literaturze zaproponowano nową klasę estymatorów pod nazwą: podwójnie odporne estymatory (ang. *doubly robust estimators*).
- Dlaczego podwójnie odporne? Estymator ten składa się z dwóch części

$$\hat{\theta}_{\text{DR}} = \underbrace{\frac{1}{N} \sum_{i=1}^N \frac{R_i \left\{ y_i - m(\hat{\beta}; \mathbf{x}_i) \right\}}{P(\hat{\lambda}; \mathbf{x}_i)}}_{\text{Średnia ważona reszt z modelu}} + \underbrace{\frac{1}{N} \sum_{i=1}^N m(\hat{\beta}; \mathbf{x}_i)}_{\text{Średnia z predykcji dla całej populacji}}, \quad (8)$$

gdzie R_i to zmienna 0-1, gdzie 1 gdy próba nielosowa, $P(\hat{\lambda}; \mathbf{x}_i)$ to prawdopodobieństwo przynależności do próby nielosowej, a $m(\hat{\beta}; \mathbf{x}_i)$ to pewien model parametryczny ($E(y|\mathbf{x}) = m(\hat{\beta}; \mathbf{x}_i)$).

Podwójnie odporne estymatory – w telegraficznym skrócie

Do wartości przewidywanych dla jednostek spoza próby nielosowej dodajemy ważone reszty z modelu dla próby nielosowej.

Podwójnie odporne estymatory

Własności:

- Estymator ten jest nieobciążony gdy model dla $P(\hat{\lambda}; \mathbf{x}_i)$ jest źle wyspecyfikowany, ale $m(\hat{\beta}; \mathbf{x}_i)$ jest dobrze wyspecyfikowany,
- Estymator ten jest nieobciążony gdy $m(\hat{\beta}; \mathbf{x}_i)$ jest źle wyspecyfikowany ale $P(\hat{\lambda}; \mathbf{x}_i)$ jest dobrze wyspecyfikowany.

Literatura:

- Chen, Y., Li, P., & Wu, C. (2020). Doubly Robust Inference With Nonprobability Survey Samples. *Journal of the American Statistical Association*, 115(532), 2011–2021.
- Kim, J. K., & Wang, Z. (2019). Sampling Techniques for Big Data Analysis. *International Statistical Review*, 87(S1), S177–S191.

Podwójnie odporne estymatory

Powyższe wzory miały zastosowanie gdy znamy wszystkie jednostki z populacji. Jednak gdy dysponujemy wyłącznie dwiema próbami (losową i nielosową) estymator ten może mieć postać:

$$\hat{\theta}_{\text{DR1}} = \underbrace{\frac{1}{N} \sum_{i \in \mathcal{S}_A} w_i \left\{ y_i - m(\mathbf{x}_i, \hat{\beta}) \right\}}_{\text{Próba nielosowa}} + \underbrace{\frac{1}{N} \sum_{i \in \mathcal{S}_B} d_i m(\mathbf{x}_i, \hat{\beta})}_{\text{Próba losowa}}, \quad (9)$$

gdzie N to znana wielkość populacji, \mathcal{S}_A to próba nielosowa, \mathcal{S}_B to próba losowa, $w_i = 1/P(\hat{\lambda}; \mathbf{x}_i)$, d_i to waga wynikająca z losowania.

Podwójnie odporne estymatory

W przypadku gdy wielkość populacji nie jest znana możemy zastosować następujący estymator

$$\hat{\theta}_{DR2} = \underbrace{\frac{1}{\hat{N}^A} \sum_{i \in S_A} w_i \left\{ y_i - m(x_i, \hat{\beta}) \right\}}_{\text{Próba nielosowa}} + \underbrace{\frac{1}{\hat{N}^B} \sum_{i \in S_B} d_i m(x_i, \hat{\beta})}_{\text{Próba losowa}}, \quad (10)$$

gdzie $\hat{N}^A = \sum_{i \in S_A} w_i$, $\hat{N}^B = \sum_{i \in S_B} d_i$.

Podwójnie odporne estymatory – estymacja wariancji

- Kim & Wang (2019) pokazali, że jeżeli próba losowa stanowi niewielki ułamek próby nielosowej ($n_B/N_A = o(1)$) to wariancję estymatora $\hat{\theta}_{DR1}$ lub $\hat{\theta}_{DR2}$ można wyznaczyć wyłącznie na podstawie próby losowej (zgodnie z jej schematem losowania).
- Chen, Li & Wu (2020) wyznaczyli estymatory wariancji bez takiego założenia oraz zaproponowali podejście oparte na metodzie bootstrap, którą można scharakteryzować następującymi niezależnymi krokami:
 - ❶ dla próby nielosowej S_A losujemy ze zwracaniem prostą próbę S_A^* o liczebności n_A ,
 - ❷ dla próby losowej S_B losujemy z prawdopodobieństwem $1/d_i^B$ ze zwracaniem próbę S_B o liczebności n_B

Następnie wyznaczamy $\hat{\theta}_{DR1}^*$ lub $\hat{\theta}_{DR2}^*$ z każdej próby bootstrapowej.

Podwójnie odporne estymatory – rozszerzenie

- Praca: Yang, S., Kim, J. K., & Song, R. (2020). *Doubly robust inference when combining probability and non-probability samples with high dimensional data*. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(2), 445-465 przedstawia rozwiązanie w tym zakresie oparte na doborze zbliżonym do regresji LASSO (dokładnie Smoothly Clipped Absolute Deviation; SCAD). Jest również pakiet w R ale o dość ograniczonych możliwościach.

Spis treści

1 Wprowadzenie

2 Próby losowe i nielosowe

3 Metody estymacji dla prób nielosowych

4 Podsumowanie

Co jest zaimplementowane?

- Inverse probability weighting - 3 estymatory.
- Mass imputation – 3 estymatory.
- Doubly robust estimators – 2 estymatory.
- Dobór zmiennych: SCAD, LASSO, MCP.
- Wariancja: analityczna i bootstrap.
- ... i wiele innych funkcjonalności.
- Więcej w materiałach: <https://ncn-foreigners.github.io/nonprobsvy-book/>

Prawo wielkich populacji (Meng, 2018)

W przypadku *Big data* i występowaniu autoselekcji mierzonej następuje zmiana paradygmatu w ocenie błędów estymacji, tj.

przechodzimy z *prawa wielkich liczb* i *centralnego twierdzenia granicznego* według, którego

$$\text{error} \propto \frac{\sigma}{\sqrt{n}}, \quad (11)$$

do relatywnego systematycznego błędu (*prawa wielkich populacji*) według, którego

$$\text{error} \propto \hat{\pi}_i \sqrt{N}. \quad (12)$$

Kiedy zaproponowane metody działają?

- Redukcja obciążenia występuje tylko wtedy kiedy \mathbf{X} i Y są ze sobą silnie skorelowane ($|Corr(Y, \mathbf{X})| > 0$)
- Redukcja obciążenia występuje tylko wtedy kiedy \mathbf{X} i R są ze sobą silnie skorelowane ($|Corr(R, \mathbf{X})| > 0$)
- Korelacja między Y i R jest bliska零u gdy ($|Corr(Y, R|\mathbf{X})| \approx 0$)
- Konieczne jest posiadanie informacji o X , a najlepiej gdybyśmy dysponowali zmienną X_k , która jest tzw. *proxy variable* – zbliżona ale nie ta sama definicja (np. cena ofertowa vs cena transakcyjna).

Nie ma jednej metody estymacji, którą można zastosować!

