Uniwersytet Gdański

Dokładność estymacji złożonej na podstawie prób nielosowych w badaniu społecznym – symulacja na bazie danych z badania EU-SILC

Arkadiusz Kozłowski

IV Kongres Statystyki Polskiej
2-4 lipca 2024 r. w Warszawie

CEL I PLAN PREZENTACJI

Celem badania jest ocena dokładności hipotetycznej strategii badania społecznego:

- podstawą wnioskowania jest próba nielosowa,
- estymacja jest wspomagana informacjami dodatkowymi z niezależnej, losowej próby referencyjnej i/lub całej populacji,
- przedmiotem badania są parametry rozkładu dochodów gospodarstw domowych.

Plan prezentacji:

1) Dlaczego próby nielosowe?
2) Obciążenie wyboru
3) Metody estymacji z prób nielosowych
4) Warunki eksperymentu
5) Wyniki eksperymentu

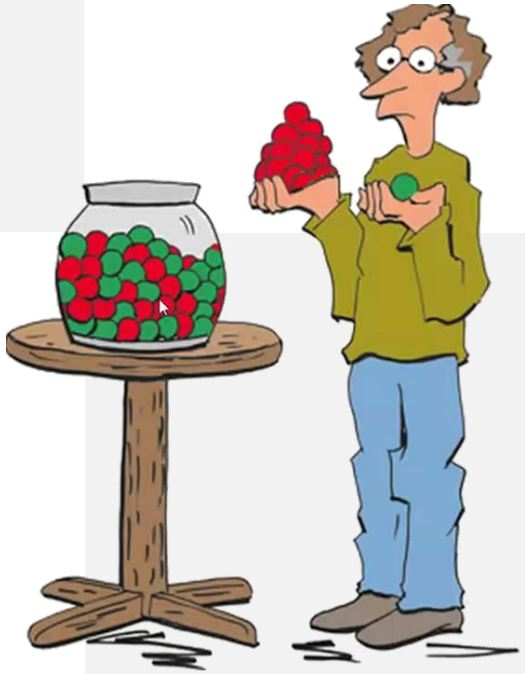
DLACZEGO PRÓBY NIELOSOWE?

- Zbieranie danych jest względnie tanie i łatwe
- Zapotrzebowanie na aktualne statystyki
- Większe możliwości zachowania anonimowości respondenta
- W badaniach społecznych próby *stricte* losowe w zasadzie nie występują
- Dużo źródeł prób nielosowych (w tym *big data*)

SELECTION BIAS

Dwie najważniejsze słabości próbkowania nieprobabilistycznego to:

- brak możliwości określenia błędu wnioskowania (porównywalnego z błędem wnioskowania z próby losowej)
- **obciążenie/błąd wyboru** (ang. *selection bias*) - wyniki uzyskane z próby nielosowej średnio rzecz biorąc różnią się od rzeczywistych charakterystyk populacji



SELECTION BIAS

Indeks **wadliwości danych** (miara błędy wyboru; Meng, 2018):

$$D_I \equiv E_R\left[\rho_{R,Y}{}^2\right]$$

gdzie:

$\rho_{R,Y}$ - współczynnik korelacji (populacyjny) między **wskaźnikiem udziału w badaniu** $\boldsymbol{R}$ ($R_k = 1$ jeżeli k-ta jednostka uczestniczy w badaniu, $R_k = 0$ w pozostałych przypadkach) a badana cechą Y.

ESTYMACJA Z PRÓBY NIELOSOWEJ

Jednym z założeń warunkujących słuszność estymacji z prób nielosowych jest to, że próba nielosowa i jednostki spoza próby są „**wymienialne**” pod warunkiem zmiennych zakłócających.

Kluczem do sukcesu estymacji z prób nielosowych jest posiadanie informacji dodatkowej (spoza próby) o silnych **zmiennych pomocniczych**, wyjaśniających skłonność do udziału w badaniu i/lub wartości badanych zmiennych.

Celem eksperymentu jest sprawdzenie w jakim stopniu wykorzystanie zmiennych społeczno-demograficznych, zwykle wykorzystywanych do kalibracji wag w badaniach reprezentacyjnych, jest w stanie zmniejszyć obciążenie próby nielosowej w przypadku szacowania parametrów rozkładu dochodów.

WARUNKI EKSPERYMENTU

- Dane jednostkowe z Europejskiego Badania Dochodów i Warunków Życia Ludności (**EU-SILC**) zrealizowanego w Polsce w 2022 r.
- Dane dla gospodarstwa domowego lub osoby wypełniającej kwestionariusz
- Szacowane parametry:
  - Parametry rozkładu **dochodów ekwiwalentnych do dyspozycji**:
    - średnia, mediana,
    - wskaźnik zagrożenia ubóstwem (*at-risk-of-poverty rate after social transfers*), P_U
    - głębokość ubóstwa (*relative median at-risk-of-poverty gap*), D_U
    - nierówność rozkładu dochodów S80/S20
    - współczynnik Giniego
  - Parametry rozkładu **wzrostu i wagi ciała** (w tym BMI):
    - średnia, mediana, współczynnik korelacji

WARUNKI EKSPERYMENTU

Zmienne pomocnicze:

- Region (7 kategorii)
- Stopień urbanizacji (3 kategorie)
- Typ gospodarstwa domowego (10 kategorii)
- Status własności mieszkania (2 kategorie)
- Liczba pomieszczeń w mieszkaniu (6 kategorii)
- Płeć (2 kategorie)
- Stan cywilny (4 kategorie)
- Poziom wykształcenia (5 kategorii)
- Status zawodowy (3 kategorie)
- Wiek (7 kategorii)

40 lub 16 zmiennych zerojedynkowych

WARUNKI EKSPERYMENTU

Teoretyczna populacja:

- Oryginalna próba liczyła 19 757 gospodarstw domowych, do których należały 48 283 osoby
- Korzystając z wag dla jednostek odtworzono prawdopodobną populację składającą się z 12 705 064 gospodarstw domowych
- Dla celów symulacji wykorzystano losowo wybrane **N=20000** rekordów i one pełniły rolę **populacji badania**

WARUNKI EKSPERYMENTU

Plan losowania próby nielosowej:

1) Prawdopodobieństwa inkluzji zależne od dochodów ekwiwalentnych:

$$\pi_{NP,k} = \frac{1}{1 + e^{-(b_0 + b_1 \cdot DOCH_k)}}$$

gdzie: $DOCH_k$ - dochód ekwiwalentny do dyspozycji k-tego gospodarstwa domowego

b_0 - wyraz wolny dobrany w taki sposób, aby $\sum_k \pi_{NP,k} = n_{NP}$

b_1 - parametr decydujący o obciążeniu wyboru

- Dwa warianty obciążenia:
  - **silne**: $b_1 = -1$ => $E_R[\rho_{R,DOCH}] = -0{,}11\ (-0{,}16)$
  - **słabe**: $b_1 = -0{,}2$ => $E_R[\rho_{R,DOCH}] = -0{,}03\ (-0{,}05)$

WARUNKI EKSPERYMENTU

Plan losowania próby nielosowej:

2) Prawdopodobieństwa inkluzji zależne od posiadania komputera i dostępu do Internetu:

Komputer	Internet	
	nie	tak
nie	11%	7%
tak	2%	80%

$$\pi_{NP,k} = \begin{cases} \frac{n_{NP}}{N^*} & jeżeli \text{ Komputer} = tak \text{ \& Internet} = tak \\ 0 & w\ pozostałych\ przypadkach \end{cases}$$

gdzie: N^* - liczba jednostek posiadających komputer i dostęp do Internetu

$$E_R[\rho_{R,DOCH}] = 0{,}017\ (0{,}025)$$

WARUNKI EKSPERYMENTU

Schemat losowania próby nielosowej:

- Losowanie systematyczne z różnymi prawdopodobieństwami wyboru po wcześniejszym przemieszaniu populacji (ang. *random systematic sampling*)
- W symulacji losowane są **wszystkie możliwe próby** (liczba wszystkich możliwych prób = N)

WARUNKI EKSPERYMENTU

Schemat losowania próby referencyjnej (losowej):

- Losowanie warstwowe z proporcjonalną alokacją próby
  - Warstwy według przekroju *regionu* i *stopnia urbanizacji* (21 warstw)

WARUNKI EKSPERYMENTU

Liczebność próby:

- Próby nielosowe: $n_{NP} = 1000$ lub $n_{NP} = 2000$
- Próby losowe: $n_P = 1000$

Zakres informacji dodatkowej:

- Zmienne warstwujące dla próby losowej znane dla całej populacji
- Pozostałe zmienne pomocnicze znane dla próby referencyjnej lub całej populacji

WARUNKI EKSPERYMENTU

Testowane metody estymacji

Kryteria wyboru:

- Bez konieczności tworzenia osobnego modelu dla każdej zmiennej celu
- Pozwala na stworzenie jednego zestawu wag dla szacowania różnych parametrów i różnych zmiennych celu

WARUNKI EKSPERYMENTU

Testowane metody estymacji

- Podejście quasi-randomizacyjne (**QR**):

$$w_{NP,i}^{(QR)} = \frac{1}{\hat{\pi}_{NP,i}}$$

gdzie: $\hat{\pi}_{NP,i}$ jest szacowane według procedury największej pseudo-wiarygodności z wykorzystaniem losowej próby odniesienia (Chen i in., 2020, s. 2013)

- Podejście modelowe (**SM**):
  - przyjmując model liniowy ze stałą wariancją

$$w_{NP,i}^{(SM)} = \hat{\boldsymbol{X}}^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{x}_i$$

gdzie: $\hat{\boldsymbol{X}}$ - wektor sum populacyjnych cech pomocniczych (rzeczywisty lub oszacowany na podstawie losowej próby odniesienia); $\boldsymbol{X}, \boldsymbol{x}_i$ – macierz i wektor wartości cech pomocniczych z próby nielosowej

WARUNKI EKSPERYMENTU

Testowane metody estymacji

- Estymacja podwójnie odporna (**DR**):
  - przyjmując model liniowy ze stałą wariancją

$$w_{NP,i}^{(DR)} = w_{NP,i}^{(QR)}\left[1 + \left(\widehat{\boldsymbol{X}} - \widehat{\boldsymbol{x}}_{QR}\right)^{T}\left(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{W}\boldsymbol{X}\right)^{-1}\boldsymbol{x}_i\right]$$

gdzie: $\boldsymbol{W} = diag\left\{w_{NP,i}^{(QR)}\right\}$; $\widehat{\boldsymbol{x}}_{QR}$ - wektor sum populacyjnych cech pomocniczych oszacowany metodą QR

WYNIKI EKSPERYMENTU

Obciążenie względne (%) szacunków bezpośrednich dla parametrów rozkładu dochodów

Wadliwość próby	n_{NP}	Me	$\bar{Y}$	P_U	D_U	$\frac{S80}{S20}$	$Gini$
silna	1000	-21,9	-26,6	16,4	20,4	-5,0	-6,6
	2000	-20,6	-25,6	14,0	19,2	-7,3	-7,7
słaba	1000	-5,4	-7,9	2,2	7,1	-5,4	-4,7
	2000	-5,1	-7,5	1,6	6,6	-5,7	-4,5
Komp. & Internet	1000	4,0	3,8	-5,9	0,7	5,8	-2,1
	2000	4,1	3,7	-5,2	0,2	5,5	-2,0

WYNIKI EKSPERYMENTU

Przeciętna zmiana obciążenia względnego (pkt. proc.)
w stosunku do szacunków bezpośrednich

Wadliwość próby	Estymator	Me	$\overline{Y}$	P_U	D_U	$\frac{S80}{S20}$	$Gini$
silna	QR	-4,4	-4,6	-2,1	2,6	-0,5	0,8
	SM	-4,4	-4,5	-2,2	3,8	-0,5	0,9
	DR	-4,5	-4,8	-2,1	2,9	-0,6	0,7
słaba	QR	-1,8	-1,7	-0,4	-0,4	-0,8	-0,1
	SM	-1,7	-1,7	-0,3	-0,4	-0,8	-0,1
	DR	-1,7	-1,7	-0,3	-0,4	-0,8	-0,2
Komp. & Internet	QR	-2,3	-2,8	0,9	-0,1	-3,7	1,6
	SM	-3,0	-3,3	0,0	0,3	-3,6	1,8
	DR	-3,0	-3,4	0,3	-0,1	-4,0	1,7

WYNIKI EKSPERYMENTU

Przeciętna zmiana względnego odchylenie standardowego estymatora (pkt. proc.) w stosunku do szacunków bezpośrednich

Wadliwość próby	Estymator	Me	$\overline{Y}$	P_U	D_U	$\frac{S80}{S20}$	$Gini$
silna	QR	0,0	0,2	0,6	2,7	0,5	0,2
	SM	-0,1	0,1	0,6	2,9	0,5	0,2
	DR	-0,1	0,1	0,6	2,8	0,5	0,2
słaba	QR	0,2	0,1	-0,3	0,6	0,3	0,1
	SM	0,1	0,0	-0,3	0,4	0,2	0,1
	DR	0,1	0,0	-0,3	0,5	0,2	0,1
Komp. & Internet	QR	-0,2	-0,1	0,6	1,2	-0,2	-0,1
	SM	-0,3	-0,2	1,0	1,9	-0,2	-0,1
	DR	-0,3	-0,2	0,9	1,6	-0,3	-0,2

WYNIKI EKSPERYMENTU

Przeciętna zmiana pierwiastka ze średniego względnego błędu kwadratowego estymatora (pkt. proc.)

w stosunku do szacunków bezpośrednich

Wadliwość próby	Estymator	Me	$\overline{Y}$	P_U	D_U	$\frac{S80}{S20}$	$Gini$
silna	QR	-4,4	-4,6	-1,7	3,5	-0,2	0,8
	SM	-4,4	-4,5	-1,8	4,7	-0,2	0,9
	DR	-4,5	-4,8	-1,7	3,8	-0,3	0,7
słaba	QR	-1,6	-1,7	-0,4	0,3	-0,4	0,0
	SM	-1,6	-1,7	-0,4	0,1	-0,5	-0,1
	DR	-1,6	-1,7	-0,4	0,2	-0,5	-0,1
Komp. & Internet	QR	-2,1	-2,3	1,1	1,2	-2,7	1,1
	SM	-2,6	-2,6	0,8	1,9	-2,6	1,3
	DR	-2,6	-2,6	0,9	1,6	-2,9	1,2

WYNIKI EKSPERYMENTU

Przeciętna zmiana obciążenia względnego (pkt. proc.)
w stosunku do szacunków bezpośrednich

liczba zmiennych pomocniczych	Me	$\overline{Y}$	P_U	D_U	$\frac{S80}{S20}$	$Gini$
16	-2,6	-2,9	-0,7	0,7	-1,6	0,9
40	-3,4	-3,4	-0,7	1,2	-1,8	0,7

WYNIKI EKSPERYMENTU

Przeciętna zmiana obciążenia względnego (pkt. proc.)
w stosunku do szacunków bezpośrednich

źródło informacji dodatkowej	Me	$\overline{Y}$	P_U	D_U	$\frac{S80}{S20}$	$Gini$
populacja	-3,0	-3,2	-0,7	0,9	-1,7	0,8
próba losowa	-3,0	-3,2	-0,7	0,9	-1,7	0,8

WYNIKI EKSPERYMENTU

Przeciętna zmiana obciążenia względnego (pkt. proc.)
w stosunku do szacunków bezpośrednich

liczebność próby nielosowej	Me	$\overline{Y}$	P_U	D_U	$\frac{S80}{S20}$	$Gini$
1000	-2,8	-3,0	0,0	1,2	-2,0	0,7
2000	-3,2	-3,4	-1,4	0,6	-1,4	0,9

WYNIKI EKSPERYMENTU

Obciążenie względne (%) szacunków bezpośrednich dla parametrów rozkładu wzrostu i wagi ciała

Wadliwość próby	n_{NP}	Waga ciała		Wzrost			BMI	
		Me	$\overline{Y}$	Me	$\overline{Y}$	ρ	Me	$\overline{Y}$
silna	1000	-0,9	-0,3	-0,6	-0,5	-5,9	0,6	0,7
	2000	-0,7	-0,3	-0,6	-0,5	-5,7	0,5	0,7
słaba	1000	-0,5	-0,1	0,0	-0,2	-1,5	0,1	0,2
	2000	-0,5	-0,1	0,0	-0,1	-1,5	0,1	0,2
Komp. & Internet	1000	0,4	0,5	0,6	0,5	1,7	-0,5	-0,5
	2000	0,6	0,5	0,6	0,5	1,7	-0,5	-0,5

WYNIKI EKSPERYMENTU

Przeciętna zmiana obciążenia względnego (pkt. proc.)
w stosunku do szacunków bezpośrednich

Wadliwość próby	Esty-mator	Waga ciała		Wzrost			BMI	
		Me	$\overline{Y}$	Me	$\overline{Y}$	ρ	Me	$\overline{Y}$
silna	QR	0,0	-0,1	-0,5	-0,4	-4,2	-0,5	-0,6
	SM	0,0	-0,1	-0,4	-0,4	-4,6	-0,4	-0,6
	DR	-0,1	-0,1	-0,5	-0,4	-4,4	-0,4	-0,6
słaba	QR	-0,1	0,0	0,2	-0,1	-1,1	0,0	-0,2
	SM	-0,1	0,0	0,2	-0,1	-1,1	0,0	-0,2
	DR	-0,1	0,0	0,2	-0,1	-1,1	0,0	-0,2
Komp. & Internet	QR	-0,2	0,0	-0,2	-0,4	-0,6	-0,4	-0,2
	SM	-0,2	-0,1	-0,3	-0,4	-0,3	-0,4	-0,1
	DR	-0,2	-0,1	-0,3	-0,4	-0,1	-0,4	-0,1

WYNIKI EKSPERYMENTU

Przeciętna zmiana względnego odchylenie standardowego estymatora (pkt. proc.) w stosunku do szacunków bezpośrednich

Wadliwość próby	Esty-mator	Waga ciała		Wzrost			BMI	
		Me	$\overline{Y}$	Me	$\overline{Y}$	ρ	Me	$\overline{Y}$
silna	QR	0,0	0,1	-0,1	0,0	0,2	0,0	0,0
	SM	0,0	0,0	-0,2	0,0	0,2	-0,1	0,0
	DR	0,0	0,0	-0,2	0,0	0,1	-0,1	0,0
słaba	QR	0,0	0,0	0,0	0,0	0,1	0,0	0,0
	SM	0,0	0,0	0,0	0,0	0,1	0,0	0,0
	DR	0,0	0,0	0,0	0,0	0,1	0,0	0,0
Komp. & Internet	QR	0,1	0,1	0,2	0,0	0,4	0,1	0,1
	SM	0,2	0,1	0,2	0,0	0,4	0,1	0,1
	DR	0,2	0,0	0,2	0,0	0,4	0,1	0,1

WYNIKI EKSPERYMENTU

Przeciętna zmiana pierwiastka ze średniego względnego błędu kwadratowego estymatora (pkt. proc.)

w stosunku do szacunków bezpośrednich

Wadliwość próby	Esty-mator	Waga ciała		Wzrost			BMI	
		Me	$\bar{Y}$	Me	$\bar{Y}$	ρ	Me	$\bar{Y}$
silna	QR	0,1	0,1	-0,4	-0,3	-2,4	-0,3	-0,4
	SM	0,0	0,0	-0,4	-0,3	-2,6	-0,3	-0,4
	DR	0,0	0,0	-0,4	-0,3	-2,5	-0,3	-0,4
słaba	QR	0,0	0,0	0,1	0,0	-0,1	0,0	0,0
	SM	0,0	0,0	0,1	-0,1	-0,1	0,0	0,0
	DR	0,0	0,0	0,1	-0,1	-0,1	0,0	0,0
Komp. & Internet	QR	0,1	0,1	-0,1	-0,3	0,3	-0,2	-0,1
	SM	0,1	0,0	-0,2	-0,4	0,3	-0,2	0,0
	DR	0,1	0,0	-0,2	-0,4	0,3	-0,2	0,0

Un

WNIOSKI

Wykorzystanie zmiennych społeczno-demograficznych do ważenia obserwacji z prób nielosowych, gdy prawdopodobieństwo znalezienia się w próbie nielosowej zależy bezpośrednio od dochodów, może zmniejszyć błędy szacunków (obciążenie, MSE) parametrów rozkładu dochodów

- ✓ poprawa zwykle będzie niewystarczająca
- ✓ poprawa dotyczy głównie szacunków prostych parametrów
- ✓ zwiększenie liczebności próby nielosowej lub liczby zmiennych pomocniczych ma pozytywny efekt głownie dla szacowania parametrów opisowych
- ✓ w większości przypadków ważenie zwiększa zmienność szacunków
- ✓ metoda ważenia nie ma większego znaczenia, jeżeli w modelach dla zmiennej celu i dla prawdopodobieństwa dostania się do próby nielosowej wykorzystuje się ten sam zestaw zmiennych

Uniwersytet Gdański

Dziękuję za uwagę

Arkadiusz Kozłowski

IV Kongres Statystyki Polskiej
2-4 lipca 2024 r. w Warszawie