Survey Package

Maciej Beręsewicz

Contents

Podstawowe definicje i oznaczenia

Populację o liczebności N oznaczymy jako:

$$\mathcal{U} = 1, 2, \ldots, N$$

Natomiast $\mathcal{S}$ oznaczymy próbę o liczebności n pobraną z populacji $\mathcal{U}$. Przez $P(\mathcal{S})$ oznaczymy prawdopodobieństwo wylosowania danej jednostki. Na podstawie tego prawdopodobieństwa możemy wyliczyć:

$$\pi_i = P(i \in \mathcal{S})$$

Losowanie proste ze zwracaniem

Losujemy jednostki do próby w sposób prosty ze zwracaniem. W takim przypadku estymator wartości średniej opisany jest następującym wzorem:

$$\bar{y}_{\mathcal{S}} = \frac{1}{n} \sum_{i \in \mathcal{S}} y_i.$$

Natomiast wariancja tego estymatora określona jest następująco

$$V(\bar{y}) = \frac{S^2}{n}(1 - \frac{n}{N}),$$

gdzie $(1 - \frac{n}{N})$ jest korektą dla populacji skończonej (ang. *finite population correction*). Natomiast S^2 jest wariancją w populacji, której estymatorem jest $s^2 = \frac{1}{n-1}\sum_{i\in\mathcal{S}}(y_i - \bar{y})^2$. W związku z tym estymatorem $V(\bar{y})$ jest

$$\hat{V}(\bar{y}) = (1 - \frac{n}{N})\frac{s^2}{n},$$

a błąd standardowy określony jest

$$SE(\bar{y}) = \sqrt{(1 - \frac{n}{N})\frac{s^2}{n}}.$$

Miarą, którą wykorzystujemy do oceny precyzji jest współczynnik zmienności CV, który jest określony następująco

$$CV(\bar{y}) = \frac{\sqrt{V(\bar{y})}}{E(\bar{y})},$$

a jego estymatorem jest

$$\hat{CV}(\bar{y}) = \frac{SE(\bar{y})}{\bar{y}}.$$

Podobne zależności możemy obserwować dla wartości globalnej oraz innych miar.

Wagi wynikające ze schematu losowania określamy jako w_i, które są równe odwrotnosci prawdopodobieństwa inkluzji do próby:

$$w_i = \frac{1}{\pi_i}.$$

W przypadku losowania prostego ze zwracaniem $\sum_{i\in\mathcal{S}} w_i = \sum_{i\in\mathcal{S}} \frac{N}{n} = N$, $\bar{y} = \frac{\sum_{i\in\mathcal{S}} w_i y_i}{\sum_{i\in\mathcal{S}} w_i}$.

Przedział ufności dla średniej w przypadku losowania prostego określony jest następujacym wzorem.

$$(\bar{y} - z_{\alpha/2}SE(\bar{y}), \bar{y} + z_{\alpha/2}SE(\bar{y})).$$

Podsumowanie:

- Wartość globalna - estymator $\hat{y} = \sum_{i\in\mathcal{S}} w_i y_i = N\bar{y}$, błąd standardowy estymatora $N\sqrt{(1 - \frac{n}{N})\frac{s^2}{n}}$
- Warość średnia - estymator $\bar{y} = \frac{\hat{y}}{n}$, błąd standardowy estymatora $\sqrt{(1 - \frac{n}{N})\frac{s^2}{n}}$
- Proporcja - estymator $\hat{y}$, błąd standardowy estymatora $\sqrt{(1 - \frac{n}{N})\frac{\hat{p}(1-\hat{p})}{n-1}}$

Losowanie stratyfikacyjne (warstwowe) ze zwracaniem

Populacja $\mathcal{U}$ o wielkości N podzielona jest na h warstw, które spełniają następującą zależność:

$$N_1 + N_2 + \ldots + N_H = N.$$

Następie dokonujemy losowania prostego (lub innego) z każdej warstwy h i próbę tę oznaczamy $\mathcal{S}_h$ o wielkości n_h. Wielkość całej próby spełnia następujące zależność $n = n_1 + n_2 + \ldots + n_H$.

Niech:

- y_{hj} - oznacza wartość j tej jednostki w warstwie h
- $t_h = \sum_{j=1}^{N_h} y_{hj}$ jest wartością globalną w warstwie h
- $t = \sum_{h=1} Ht_h$ - jest wartością globalną dla populacji
- $\bar{y}_{hU} = \frac{\sum_{j=1}^{N_h} y_{hj}}{N_h}$ - jest wartością średnią w warstwie h
- $\bar{y}_U = \frac{t}{N} = \frac{\sum_{h=1}^{H}\sum_{j=1}^{N_h} y_{hj}}{N}$ - jest średnią dla całej populacji
- $S_h^2 = \sum_{j=1}^{N_h} \frac{(y_{hj} - \bar{y}_h U)^2}{N_h - 1}$ jest wariancją w warstwie h

a estymatory to odpowiednio

- $\bar{y}_h = \frac{1}{n_h}\sum_{j\in\mathcal{S}_h} y_{hj}$
- $\hat{y} = \frac{N_h}{n_h}\sum_{j\in\mathcal{S}_h} y_{hj} = N_h\bar{y}_h$
- $s_h^2 = \sum_{j\in\mathcal{S}_h} \frac{(y_{hj} - \bar{y}_h)^2}{n_h - 1}$

Wartość globalna dla populacji

$$\hat{t}_{str} = \sum_{h=1}^{N} \hat{y}_h = \sum_{h=1}^{N}\sum_{j\in\mathcal{S}_h} w_{hj} y_{hj}$$

wariancja jest równa

$$\hat{V}(\hat{y}_{str}) = \sum_{h=1}^{N} \hat{V}(\hat{t}_h) = \sum_{h=1}^{N} (1 - \frac{n_h}{N_h}) N_h^2 \frac{s_h^2}{n_h}.$$

Natomiast wartość średnia równa będzie $\bar{y}_U = t/N$, a wariancja $\hat{V}(\bar{y}_{str}) = \hat{V}(\hat{t}_{str})/N^2$.

Losowanie dwustopniowe

Jednostka losowania pierwszego stopnia (ang. *Primary Sampling Unit*) określa jednostki losowanie w pierwszej kolejności.

Estymatorem Horvitza-Thompsona dla losowania dwustopniowego jest następująca wartość

$$\hat{y}_{HT} = \sum_{i\in\mathcal{S}} \frac{\hat{y}_i}{\pi_i} = \sum_{i=1}^{N} Z_i \frac{\hat{t}_i}{\pi_i}$$

gdzie $Z_i = 1$ jeżeli jednostka i została wylosowana do badania. Natomiast wariancja estymatora HT jest określona wzorem:

$$\hat{V}_{HT}(\hat{y}_{HT}) = \sum_{i\in\mathcal{S}} (1 - \pi_i) \frac{\hat{t}_i^2}{\pi_i^2} + \sum_{i\in\mathcal{S}} \sum_{k\in\mathcal{S}, k\neq i} \frac{\pi_{ik} - \pi_i\pi_k}{\pi_{ik}} \frac{\hat{t}_i}{\pi_i} \frac{\hat{t}_k}{\pi_k} + \sum_{i\in\mathcal{S}} \frac{\hat{V}(\hat{t}_i)}{\pi_i}$$

Metody szacowania wariancji estymatorów

Balanced Repeated Replication (BRR)

Idea - dzielimy w każdym przekroju na dwie części i nadajemy im odpowiednią wagę. Zacznijmy od następujących oznaczeń.

Niech r oznacza próbę podzieloną na dwie części, $\alpha_r = (\alpha_{r1}, \alpha_{r2}, \ldots, \alpha_{rH})$. Następnie niech

$$y_h(\alpha_r) = \begin{cases} y_{h1} & \text{jeżeli } \alpha_{rh} = 1 \\ y_{h2} & \text{jeżeli } \alpha_{rh} = -1 \end{cases}$$

Co równoważnie oznacza

$$y_h(\alpha_r) = \frac{\alpha_{rh} + 1}{2} y_{h1} - \frac{\alpha_{rh} - 1}{2} y_{h2}$$

Próba jest tak zbalansowana żeby spełnione było:

$$\sum_{r=1}^{R} \alpha_{rh}\alpha_{rl} = 0, \text{ dla każdego } l \neq h$$

Następnie dla każdej replikacji r, obliczamy $\hat{\theta}(\alpha_r)$ ale tylko na podstawie połowy próby wyznaczonej przez α_r. W związku z tym estymator dla wariancji jest opisany następująco:

$$\hat{V}_{BRR}(\hat{\theta}) = \frac{1}{R} \sum_{r=1}^{R} (\theta(\hat{\alpha}_r) - \hat{\theta})^2$$

W przypadku losowania z wykorzystaniem procedury wielostopniowej dodatkowo tworzone są wagi zgodnie z

$$w_i(\alpha_r) = \begin{cases} 2w_i & \text{jeżeli jednostka } i \text{ jest w pół-próbie } r \\ 0 & \text{w przeciwnym wypadku} \end{cases}$$

Natomiast w podejściu Fay'a proponowana jest poprawka na te wagi oznaczana jako ρ, gdzie zwykle przyjmuje się wartości

$$w_i(\alpha_r) = \begin{cases} (2-\rho)w_i & \text{jeżeli jednostka } i \text{ jest w pół-próbie } r \\ \rho & \text{w przeciwnym wypadku} \end{cases}$$

Natomiast estymator dla wariancji (dowolnego) estymatora zostaje taki sam.

Jackknife

Niech $\hat{\theta}_{(j)}$ będzie estymatorem o takiej samej formie jak $\hat{\theta}$ ale bez obserwacji j. Więc, jeżeli $\hat{\theta} = \bar{y}$, to $\hat{\theta}_{(j)} = \bar{y}_{(j)} = \sum_{i \neq j} y_i/(n-1)$. W związku z tym wariancja szacowana metodą JK ma następującą postać:

$$\hat{V}_{JK}(\hat{\theta}) = \frac{n-1}{n} \sum_{j=1}^{n} (\hat{\theta}_{(j)} - \hat{\theta})^2$$

Gdy stosujemy metodę losowania dla prób mających wielostopniowych estymator wariancji określony jest następująco

$$\hat{V}_{JK}(\hat{\theta}) = \sum_{h=1}^{H} \frac{n_h - 1}{n_h} \sum_{j=1}^{n_h} (\hat{\theta}_{(hj)} - \hat{\theta})^2$$

gdzie wagi określone są następujaco:

- w_i - jeżeli jednostka i nie jest w warstwie h
- 0 - jeżeli jednostka i jest w PSU j w warstwie h
- $\frac{n_h}{n_h - 1} w_i$ - jeżeli jednostka i jestw warstwie h ale nie w PSU j

Bootstrap

Metoda polega na losowaniu B razy prób $\mathcal{S}^*$ o wielkości n ze zwracaniem z próby $\mathcal{S}$ o wielkości n. Przebieg jest następujący:

1. Dla każdej replikacji $b = 1, 2, \ldots, B$ losujemy ze zwracaniem $n_h - 1$ jednostek z n_h próby PSU w warstwie h. Robimy to niezależnie dla każdej warstwy. Niech $m_{hj}(B)$ określa liczbę wylosowań jednostki PSU j z warstwy h w próbie b.
2. Obliczana jest następnie waga według wzoru:

$$w_i(b) = w_i \times \frac{n_h}{n_h - 1} m_{hj}(b)$$

dla obserwacji i w PSU j w warstwie h.

3. Następnie szacujemy wariancję dla każdej iteracji b. Niech $\hat{\theta}_b*$ będzie estymatorem θ w każdej replikacji b przy użyciu wag $w_i(b)$. Wtedy otrzymujemy estymator wariancji metodą bootstrap, który ma następującą postać

$$\hat{V}_B(\hat{\theta}) = \frac{1}{B - 1} \sum_{b=1}^{B} (\hat{\theta}_b * - \hat{\theta})^2$$

Podsumowanie zalet i wad

- BRR
  - Zalety: nieskomplikowane obliczeniowo w porównaniu do jackknife i bootstrap, może być wykorzystany do szacowania wariancji estymatorów dla kwantyli, jest zbieżny asymptotycznie do podejścia linearyzacji Taylora
  - Wady: może być wykorzystany gdy w ramach warstw mamy parzystą liczbę PSU, może przeszacowywać wariancję dla podejścia bez zwracania.
- Jackknife
  - Zalety: można szacować wariancję dla dowolnej statystyki, nie ma problemu związanego z parzystą liczbą PUS w schemacie wielostopniowym ze stratyfikacją, można wykorzystywać do szacowania wpływu imputacji na wariancję estymatorów.
  - Wady: dla niektórych schematów losowania może być skomplikowany obliczeniowo, może niedoszacowywać wariancji estymatorów, które nie są oparte na wartościach globalnych (np. estymatory dla kwantyli).
- Bootstrap
  - Zalety: działa bez problemu dla statystyk, które oparte są na wartości globalnej lub średniej, można bezpośrednio szacować przedziały ufności o dowolnej szerokości,

– Wady: złożoność obliczeniowa większa niż w BRR czy JK, oszacowana wariancja jest różna w przypadku doboru różnych prób bootstrapowych

Metody ważenia danych

Kalibracja

Kalibracja jest wykorzystywana w badaniach częściowych, gdy znane są wagi wynikające ze schematu losowania próby. Polega na korygowaniu wag wejciowych przy wykorzystaniu informacji dodatkowych dzięki czemu możliwe będzie polepszenie oszacowań parametrów w~przypadku występowania braków danych. Istnieją inne metody wykorzystujące dodatkowe zmienne, na przykład ważenie oparte na regresji logistycznej czy uogólnionych estymatorach regresyjnych (ang. *Generalized Regression Estimators – GREG*), jednak kalibracja różni się od innych tym, że wykorzystuje odpowiednie równanie kalibracyjne.

Kalibrację można zdefiniować jako *metodę polegającą na skorygowaniu wyjciowych wag wynikających ze schematu losowana próby, celem redukcji obciążeń wynikających z istnienia braków odpowiedzi, tak aby spełnione było odpowiednie równanie kalibracyjne. Wagi te obliczane są w oparciu o wykorzystanie informacji dodatkowych z lub spoza próby.*

Formalnie, problem kalibracji polega na poszukiwaniu tak zwanych wag kalibracyjnych, który można przedstawiść następująco:

Warunek 1. Minimalizacja funkcji odległości

$$D(\mathbf{d}, \mathbf{w}) = \sum_{i=1}^{n} d_i G(\frac{d_i}{w_i}) \rightarrow min,$$

Warunek 2. Równania kalibracyjne:

$$\sum_{i=1}^{n} d_i x_{ij} = \mathbf{X}_j, \qquad j = 1, \ldots, k,$$

Warunek 3. Warunki ograniczające:

$$L \leq \frac{d_i}{w_i} \leq U, \qquad L < 1 \ \ i \ \ U > 1, i = 1, \ldots, n.$$

gdzie:

$\mathbf{w} = (w_1, \ldots, w_n)^T$ – jest wektorem wag wynikających ze schematu losowania próby,

$\mathbf{d} = (d_1, \ldots, d_n)^T$ – jest poszukiwanym wektorem wag kalibracyjnych,

n – liczebność próby,

N – liczebność populacji,

x_k – oznacza zmienne pomocnicze,

$\mathbf{X}_j = \sum_{i=1}^{N} x_{ij}$ – oznacza wartość globalną zmiennej $x_j, j = 1, \ldots, k$

x_{ij} – oznacza wartość j–tej zmiennej pomocniczej dla i–tej jednostki badania,

k – liczba zmiennych pomocniczych,

i – oznacza jednostkę.

G – jest dowolną funkcją spełniającą następujące warunki:

- $G(\bullet)$ jest ścile wypukła i~dwukrotnie różniczkowalna,
- $G(\bullet) \geq 0$,
- $G(1) = 0$,
- $G'(1) = 0$,
- $G''(1) = 1$.

Pierwszy warunek oznacza, że odległość między wagami wynikającymi ze schematu losowania próby i szukanymi wagami kalibracyjnymi powinna być jak najmniejsza. Drugi informuje o tym, że wagi kalibracyjne powinny być tak ustalone aby po ich zastosowaniu do wszystkich zmiennych pomocniczych otrzymaść ich wartości globalne. Trzeci warunek jest warunkiem ograniczającym, zapobiegającym osiągnięciu przez wagi kalibracyjne wartości ujemnych bądź ekstremalnych.

Poniżej znajdują się przykłady funkcji $G(\bullet)$, które są najczęciej rozważane w~literaturze:

$$G_1(x) = \frac{1}{2}(x-1)^2,$$

$$G_2(x) = \frac{(x-1)^2}{x},$$

$$G_3(x) = x(\log x - 1) + 1,$$

$$G_4(x) = 2x - 4\sqrt{x} + 2,$$

$$G_5(x) = \frac{1}{2\alpha}\int_1^x sinh[\alpha(t - \frac{1}{t})]dt.,$$

Jedynie pierwsza funkcja pozwala na jawną postać estymatora kalibracyjnego w~postaci macierzowej tj. można ją oprogramowaść bezpośrednio w języku macierzowym. Pozostałe funkcje odległoci wymagają numerycznych obliczeń przez co są bardziej skomplikowane. W~dalszej częci pracy zakładać będziemy że dokonujemy estymacji wartości globalnej zmiennej Y.

Wartością globalną zmiennej Y jest:

$$Y = \sum_{i=1}^{N} y_i$$

gdzie:

y_i – oznacza wartość zmiennej y dla i–tej jednostki

Oszacowania wartości globalnej dokonywać będziemy w~oparciu o n–elementową próbą wylosowaną zgodnie z~okrelonym schematem losowania z N–elementowej populacji. Klasycznym estymatorem wartości globalnej jest estymator Horvitza–Thompsona, który wyraża się wzorem:

$$\hat{Y} = \sum_{i=1}^{n} d_i y_i$$

gdzie: d_i – jest wagą przypisaną i–tej jednostce wylosowanej do próby w wyniku kalibracji.

Ponieważ nie wszystkie jednostki z n–elementowej próby udzielają odpowiedzi, ważona suma będzie zazwyczaj niedoszacowywała prawdziwą wartość globalną.

Zachodzi zatem potrzeba odpowiedniej korekty wag w_i tak, aby zniwelować ujemny wpływ braków odpowiedzi w~odniesieniu do zmiennej Y. Sposobów wyznaczania nowych wag (tzw. wag kalibracyjnych) opisany zostanie w~następnym podrozdziale.

Estymator kalibracyjny wartości globalnej ze znanym wektorem X

W~podejciu kalibracyjnym ze znanym wektorem $\mathbf{X}$ zakładamy, że znamy wartości zmiennych pomocniczych dla każdego respondenta (na przykład wypełniona metryczka) oraz że znany jest wektor wartości globalnych wszystkich zmiennych pomocniczych.

Estymator kalibracyjny wartości globalnej ze znanym wektorem $\mathbf{X}$ ma następującą postać:

$$\hat{Y}_{\mathbf{X}} = \sum_{i=1}^{m} d_i y_i$$

gdzie:

wektor $\mathbf{d} = (d_1, d_2, \ldots, d_m)^T$ jest rozwiązaniem zadania optymalizacyjnego:

$$\mathbf{d} = argmin_v D(\mathbf{v}, \mathbf{w})$$

przy warunku:

$$\mathbf{X} = \widetilde{\mathbf{X}}$$

gdzie:

$$\widetilde{X} = (\sum_{i=1}^{m} d_i x_{i1}, \sum_{i=1}^{m} d_i x_{i2}, \ldots, \sum_{i=1}^{m} d_i x_{ik})^T$$

a wektor wartości globalnych $\mathbf{X}$ ma postaść:

$$\mathbf{X} = (\sum_{i=1}^{N} x_{i1}, \sum_{i=1}^{N} x_{i2}, \ldots, \sum_{i=1}^{N} x_{ik})^T$$

gdzie:

- N to liczba respondentów,
- k to liczba zmiennych pomocniczych,
- m oznacza liczebność zbioru respondentów.

Rozwiązaniem zadania optymalizacyjnego przy warunku jest wektor wag kalibracyjnych postaci $d = (d_1, \ldots, d_m)^T$, którego składowe spełniają równanie:

$$d_i = w_i + w_i(\mathbf{X} - \hat{\mathbf{X}})^T (\sum_{i=1}^{m} w_i \mathbf{x}_i \mathbf{x}_i^T)^{-1} \mathbf{x}_i,$$

przy czym

$$\hat{\mathbf{X}} = (\sum_{i=1}^{m} w_i x_{i1}, \sum_{i=1}^{m} w_i x_{i2}, \ldots, \sum_{i=1}^{m} w_i x_{ik})^T,$$

a

$$\mathbf{x}_i = (x_{i1}, \ldots, x_{ik})^T$$

jest wektorem złożonym z wartości wszystkich k zmiennych pomocniczych dla i–tego respondenta, $i = 1, \ldots, m$.

Estymatory kalibracyjne w badaniach pełnych

W~badaniach marketingowych nie zawsze mamy do czynienia z budowaniem próby na podstawie schematu losowania. Zwykle takie badania generują duże koszty, które uniemożliwiają ich przeprowadzanie na dużą skalę. W związku z tym pojawia się problem braku wag wynikających ze schematu losowania, które są niezbędne z punktu widzenia wykorzystania estymatorów kalibracyjnych w wypadku istnienia braków odpowiedzi.

Jak wskazuje literatura możliwe jest jednak wykorzystanie estymatorów kalibracyjnych w badaniach pełnych, w których nie losujemy jednostek. Problem ten dotyczy głównie prób kwotowych oraz celowych, gdzie w takim przypadku należy utworzyć sztuczne wagi, które będą następnie korygowane przy użyciu podejcia kalibracyjnego. Wagi tworzone są według następującego schematu.

$$w_i = \begin{cases} 1, \text{ gdy nie występowały braki odpowiedzi,} \\ 0, \text{ w przeciwnym wypadku,} \end{cases}$$

gdzie:

w_i oznacza sztuczną wagę przypisaną i–temu respondentowi.

Wagi mogą być tworzone dla tabel jednowymiarowych, które mają na celu przedstawienie rozkładu odpowiedzi na dane pytanie, jak również dla tabel dwuwymiarowych (tabel kontyngencji), które mają na celu zestawienie dwóch pytań. Istnieje również możliwość wykorzystania podejcia kalibracyjnego gdy zachodzi potrzeba konstrukcji p–wymiarowych tabel kontyngencji, $p \geq 2$. W~pierwszym wypadku, wagi będą tworzone uwzględniając braki odpowiedzi tylko w~jednym pytaniu, natomiast w drugim, wagi będą konstruowane zgodnie z~poniższym schematem.

$$w_i = \begin{cases} 1, \text{ gdy dla i–tego respondenta mamy odpowiedz na obydwa pytania,} \\ 0, \text{ w przeciwnym wypadku.} \end{cases}$$

Pakiet survey

Podstawowe informacje o pakiecie

Pakiet został przygotowany przez Thomas'a Lumley'a z Uniwersytetu w Auckland, Nowa Zalandia. Szczegółowy opis można znaleźć na stronie http://r-survey.r-forge.r-project.org/survey/. Obecnie pakiet jest w wersji 3.30-3. Pakiet wykorzystywany jest do estymacji z wykorzystaniem prób nieprostych (*complex surveys*). Umożliwia korzystanie ze zbiorów wczytanych do R, jak i połączenia z bazami danych. Jest również drugi (póki co eksperymentalny) pakiet *sqlsurvey*, który umożliwia wykorzystanie różnych typów baz danych.

Deklaracja schematu losowania

Aby rozpocząć pracę z pakietem *survey* należy w pierwszym kroku zadeklarować sposób doboru próby (schemat losowania). Jest to istotny element ponieważ wpływa na poprawność estymacji wariancji (błędów szacunku) poszczególnych estymatorów. Zadeklarowany obiekt będzie również wykorzystywany w przypadku korzystania z metod replikacyjnych (np. bootstrap) w którym należy odtworzyć sposób losowania jednostek do badania.

Deklarujemy schemat losowania funkcją *svydesign*, która ma następujące parametry:

- ids - formuła lub ramka danych w której określamy losowane zespoły jednostek (od pierwszego do kolejnych etapów). Jeżeli w badaniu nie były losowane jednostki wpisujemy ~0 lub ~1.
- probs - formuła lub ramka danych określająca prawdopodobieństwa inkluzji.
- strata - formuła lub wektor określający warstwy, w przypadku ich braku określamy NULL.
- variables - formuła lub ramka danych określająca zmienne wybrane do analizy. Jeżeli nie zostaną określone wybrany zostanie cały zbiór danych.
- fpc - poprawka na szkończoną populację (ang. *Finite population correction*).
- weights - formuła lub wektor określający wagi wynikające z losowania (alternatywa dla prob).
- data - ramka danych, która zawiera zmienne określone we wcześniejszych formułach. Można również wykorzystać połączenie do bazy danych czy obiekt imputationList.
- nest - jeżeli TRUE wskazuje, że należy stworzyć nowe ID na podstawie zadeklarowanych zmiennych w argumencie ids.
- check.strata - jeżeli TRUE wskazuje, że należy sprawdzić czy zespoły jednostek są zawarte w warstwach określonych w argumencie strata.
- pps - metody określające proste losowanie bez zwracania - “brewer” lub “overton”.
- dbtype - wskazujemy z jakiej bazy danych korzystamy.
- dbname - wskazujemy nazwę bazy danych.
- variance - określamy w przypadku losowania prostego bez zwracania (variance=“YG”). W takim przypadku zostanie zastosowany estymatoFor Yatesa-Grundiego zamiast estymatora Horvitza-Thompsona.

Argumenty wymagane: ids (oraz data). Zmienne, które posłużą do określenia argumentów nie mogą zawierać braków danych.