Polskie Forum Psychologiczne, 2018, tom 23, numer 2, s. 291-310
DOI: 10.14656/PFP20180205

WYBRANE STATYSTYCZNE METODY RADZENIA SOBIE Z BRAKAMI DANYCH*

Artur Pokropek

Instytut Filozofii i Socjologii Polskiej Akademii Nauk w Warszawie
Institute of Philosophy and Sociology of the Polish Academy of Sciences in Warsaw

SELECTED STATISTICAL METHODS FOR HANDLING MISSING DATA

Summary. This article presents selected modern statistical way of understanding of missing data. It focuses on explaining missing data mechanisms introduced by Rubin: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). Classical and modern statistical tools for coping with missing data are presented end evaluated including: missing data deletion, mean imputation, regression imputation, stochastic regression imputation, hot deck imputation, multiple imputation (MI) and maximum likelihood estimation with missing data (ML). The article finishes with practical guidance for missing data handling.
Key words: missing data, maximum likelihood, multiple imputation, hot deck imputation

Wprowadzenie

Występowanie braków danych w badaniach społecznych jest sytuacją naturalną i częstą. Respondenci niekiedy odmawiają udziału w badaniu, z różnych przyczyn nie chcą lub nie potrafią odpowiedzieć na zadawane im pytania, czasami przypadkowo opuszczają niektóre pytania. Występowanie braków danych w badaniu rodzi wiele niebagatelnych problemów, takich jak zniekształcenie rozkładów analizowanych zmiennych czy wzrost obciążenia i wariancji wykorzystywanych estymatorów. Ignorowanie problemu braków danych lub niewłaściwe traktowanie tego problemu może zniekształcić wyniki i prowadzić badacza do błędnych wniosków. Ewolucja

* Publikacja ta została przygotowana w ramach projektu *From school to work: indywidualne i instytucjonalne wyznaczniki kształtowania się ścieżek karier edukacyjno-zawodowych młodych Polaków (FS2W)*, który jest finansowany przez Narodowe Centrum Nauki, w ramach konkursu grantowego Maestro 3, umowa nr UMO-2012/06/A/HS6/00323.

Adres do korespondencji: Artur Pokropek, e-mail: artur.pokropek@gmail.com

teorii i metod statystycznych przyniosła wiele rozwiązań mogących stanowić pomoc w rozstrzyganiu problemów wynikających z braków danych. Część z nich przedstawiono w tym artykule.

Artykuł ten bazuje na podejściu Rubina i jego statystycznym rozumieniu braków danych. Zaprezentowane zostaną tutaj klasyczne metody radzenia sobie z brakami danych oraz metody nawiązujące bezpośrednio do koncepcji Rubina. Nie jest to zatem kompletny przegląd metod rozwiązywania problemów braków danych, a wybór potencjalnie najbardziej użytecznych dla badacza zjawisk psychologicznych i społecznych.

Mechanizmy powstawania braków danych i ignorowalność

W 1976 roku Rubin opublikował dziewięciostronicowy artykuł *Interference and missing data* (Rubin, 1976). Na tych kilku stronach sformułował podstawy, które stały się fundamentem dla nowoczesnych technik radzenia sobie z brakami danych. W artykule po raz pierwszy przedstawiony jest przejrzysty, matematycznie rygorystyczny, system klasyfikacji braków danych oraz założenia, które za nim stoją. W tym krótkim artykule Rubin stawia podwaliny estymacji za pomocą metody największej wiarygodności (MNW; ang. *maximum likelihood*, ML) oraz techniki wielokrotnych imputacji (ang. *multiple imputation*, MI) w sytuacji występowania braków danych.

W podejściu wywodzącym się od Rubina brak danych jest uważany za fenomen probabilistyczny i podległy „procesowi powodującemu braki danych" (*process that causes missing data*) (Rubin, 1976, s. 582). Mechanizm, który skutkuje tym, iż niektóre dane są rejestrowane, a inne nie, opisuje powstawanie braków danych oraz relacje między brakami danych a kompletnym zbiorem danych. Nazywany on będzie mechanizmem generującym braki danych.

Rubin, a za nim większość badaczy zajmujących się problematyką braków danych, wyróżnia trzy typy mechanizmów generujących braki danych:

(1) Mechanizm całkowicie losowy (MCAR: *Missing completely at random*);
(2) Mechanizm losowy (MAR: *Missing at random*);
(3) Mechanizm nielosowy (MNAR Missing not at random).

Pełny statystyczny opis mechanizmów generujących braki danych i wynikające z nich konsekwencje dla statystycznego modelowania zostały umieszczone w aneksie. Dla badaczy zainteresowanych jedynie aplikacyjnymi problemami modelowania statystycznego problematykę można uprościć. Z mechanizmem MCAR mamy do czynienia, gdy proces powstawania braków danych jest całkowicie losowy. W sytuacji MAR występowanie zjawiska braków danych zależy wyłącznie od zmiennych obserwowalnych. Mechanizm MNAR opisuje sytuację, w której braki danych związane są ze zmiennymi nieobserwowalnymi. Na przykład mechanizm MCAR występuje w badaniach, w których respondentom losowo przydziela się różne wersje kwestionariuszy, częściowo różniące się zestawami pytań (Pokropek, 2011). Z mechanizmem MAR możemy mieć do czynienia, gdy prawdopodobieństwo odmowy odpowiedzi

związane jest jedynie z obserwowalną, zarejestrowaną podczas wywiadu cechą respondenta, np. płcią. Mechanizm MNAR może uwidaczniać się w sytuacji, gdy niechęć respondentów do odpowiadania na dane pytanie związana jest z nieobserwowalną w danym badaniu cechą, np. zgeneralizowanym zaufaniem, statusem społecznym czy inteligencją.

Dlaczego rozróżnienie typów mechanizmów powodujących braki danych jest takie ważne? Ponieważ w zależności od tego z jakim mechanizmem mamy do czynienia różne metody radzenia sobie z brakami danych będą w różnym stopniu efektywne. Rubin stworzył i formalnie omówił koncepcję kryterium ignorowalności (Rubin, 1976). Kryterium te mówi o tym dla jakiego mechanizmu braków danych i metod radzenia sobie z brakami danych, problem braków danych może być ignorowany, a dla jakiej sytuacji nie można go ignorować bez poważnych konsekwencji dla trafności uzyskanych wyników.

Jeżeli mamy do czynienia z mechanizmem MCAR, nawet proste metody jak usunięcie obserwacji parami nie będą prowadziły do znacznych błędów w estymacji (kosztem jaki tutaj poniesiemy będzie utrata mocy testów statystycznych wraz ze wzrostem udziału braków danych w próbie). Braki danych w tej sytuacji (mechanizm MCAR i sposób radzenia sobie z brakami danych: usunięcie parami) są ignorowane. Jeżeli mamy do czynienia z sytuacją MAR proste metody nie wystarczą do spełnienia kryterium ignorowalności. Rubin (1976) udowodnił, że aby mówić o ignorowalności w sytuacji MAR trzeba odwołać się do takich metod jak estymacja metodą największej wiarygodności, która bierze pod uwagę braki danych lub metody oparte na bayesowskich wielokrotnych imputacjach, czyli MNW i WI. Jeżeli chodzi o sytuację MNAR to ignorowalność uzyskamy jedynie wtedy, kiedy znamy dokładną wartość parametrów stojących za mechanizmem powodującym powstawanie braków danych. W praktyce oznacza to, że ignorowalność dla sytuacji MNAR jest praktycznie niemożliwa.

„Klasyczne" metody radzenia sobie z brakami danych

W tej części przyjrzymy się klasycznym metodom radzenia sobie z brakami danych. Przegląd ten pokaże jakie konsekwencje wiążą się z wykorzystaniem klasycznych metod radzenia sobie z brakami danych. Zademonstrowane zostanie jak decyzje związane z radzeniem sobie z brakami danych wpływają na końcowe oszacowania. Pokazane zostanie w jakich sytuacjach klasyczne metody dają zadowalające wyniki, a w jakich sytuacjach należy ich unikać.

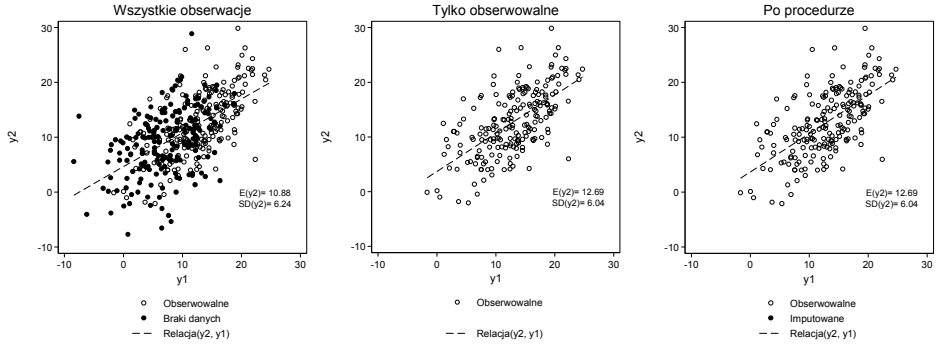
Każdemu opisowi metody radzenia sobie z brakami danych towarzyszył będzie rysunek pokazujący istotę przeprowadzanych zabiegów. Każdy z rysunków odnosi się do jednego zbioru danych o liczebności 400. W przedstawionych przykładach zmienne wylosowane zostały z wielowymiarowego rozkładu normalnego. Dla y_1 średnia wynosiła 10, a dla y_2 11. Odchylenie standardowe dla obydwu zmiennych wynosiło 6, a korelacja między nimi 0,6. W zbiorze tym mamy dwie zmienne y_1 i y_2. Dodatkowo za pomocą mechanizmu MAR wygenerowano braki danych dla zmien-

nej y_2. Innymi słowy, prawdopodobieństwo pojawienia się braków danych skorelowane jest ze zmienną y_1 (pozostającą w pełni obserwowalną), a nie ze zmienną y_2. Procent braków danych dla zmiennej y_2 wynosi 50%.

Dla każdej metody radzenia sobie z brakami danych graficzna reprezentacja procedury ma taki sam schemat. Rysunek przedstawiający procedurę radzenia sobie z brakami danych składa się z trzech wykresów. Na pierwszym wykresie (patrząc od lewej strony) znajduje się wykres rozrzutu dwóch zmiennych y_2 i y_1. Jest to kompletny zbiór danych, w którym obserwujemy wszystkie wartości zmiennych niezależnie od tego czy są one obserwowane (puste kropki) czy mamy do czynienia z brakami danych (pełne kropki). Na wykresie przedstawione są tylko obserwowalne wartości zmiennych. Na trzecim wykresie odnaleźć można zbiór po zastosowaniu procedury mającej poradzić sobie z brakami danych. Na każdym z wykresów zamieszczono przerywaną linię reprezentującą relację między y_1 i y_2 oraz podano średnią i odchylenie standardowe dla zmiennej y_2, czyli tej dla której obserwujemy braki danych.

Usuwanie wszystkich jednostek obserwacji z analiz

Pierwsza metoda polega na usunięciu wszystkich obserwacji posiadających braki danych przynajmniej dla jednej zmiennej. W tym wypadku drugi i trzeci wykres rozrzutu przedstawione na rysunku 1 są tożsame.



Rysunek 1. Ilustracja podejścia do analiz z brakami danych: usuwanie wszystkich obserwacji mających braki danych, mechanizm generowania braków danych: MAR

Źródło: opracowanie własne.

Jak widać po usunięciu obserwacji z brakami danych, relacja między y_1 i y_2 nie zmienia się – nachylenie linii pozostaje niezmienione. Zmienia się jednak średnia y_2, zwiększa się wyraźnie z 10,88 dla całego zbioru na 12,69 dla zbioru jedynie z obserwowalnymi danymi. To zrozumiałe, skoro mechanizm powstawania braków danych

jest MAR i prawdopodobieństwo pojawienia się braków danych jest związane z wartością y_1 (w prezentowanym przykładzie im niższa wartość y_1 tym wyższe prawdopodobieństwo pojawienia się braków danych), natomiast gdy y_1 i y_2 są skorelowane to średnio rzecz biorąc prawdopodobieństwo braków danych dla y_2 dla niższych wartości jest większe niż dla wyższych wartości (pośrednio przez wpływ y_1).

Prezentowana metoda ma swoje zalety. Gdy mechanizm generowania braków danych ma charakter MCAR braki danych są ignorowalne. W przypadku MAR właściwości tej metody są słabsze, błędnie estymowana jest średnia zmiennej z brakami danych lecz relacje, które nie zależą od wartości średniej (np. współczynnik korelacji) pozostają nieobciążone.

Usuwanie jednostek obserwacji z analiz parami

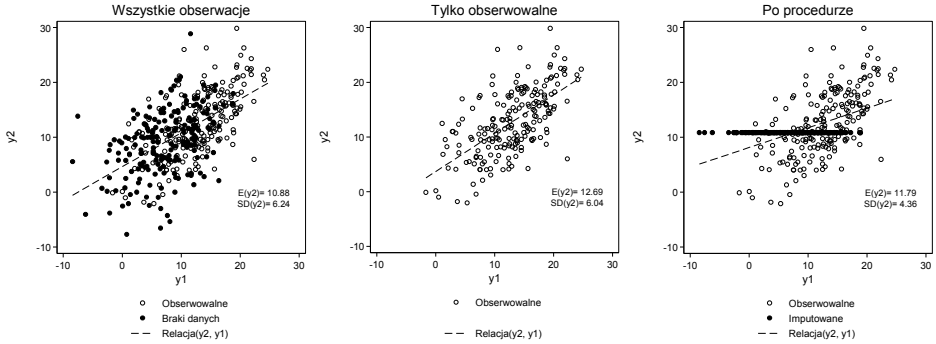
Usuwanie wszystkich jednostek obserwacji jest problematyczne w sytuacji, kiedy mamy do czynienia ze zbiorem, w którym figuruje wiele zmiennych. Dla przykładu, jeżeli zbiór danych zawiera 20 zmiennych, a każda ze zmiennych charakteryzuje się 2% braków danych. Jeżeli braki danych są czysto losowe to średnio chcąc usuwać wszystkie obserwacje mające przynajmniej jeden brak danych musielibyśmy usunąć około 67% zbioru danych (McKnight, 2007; Enders, 2010).

Innym rozwiązaniem jest usuwanie obserwacji parami. Dla przykładu, jeżeli szacujemy macierz korelacji między 20 zmiennymi, to dla każdej pary zmiennych potrzebnej do oszacowania współczynnika korelacji wykorzystuje się wszystkie dostępne informacje. Skutkuje to tym, że każda para zmiennych może mieć inną liczebność. Jest to problematyczne w analizach wielowymiarowych, utrudnia szacowanie liczby stopni swobody, a co za tym idzie problematyzuje stosowanie testów statystycznych. Jest to jednak cena za zwiększenie mocy analiz statystycznych, do analiz wielowymiarowych wykorzystywana jest bowiem znacznie większa liczba obserwacji.

Pozostałe właściwości metody usuwania jednostek obserwacji parami pozostają takie same jak dla usuwania wszystkich jednostek z brakami danych. W szczególnym wypadku, gdy mamy do czynienia z dwoma zmiennymi, tak jak w prezentowanym przykładzie, sytuacja usuwania jednostek obserwacji analiz parami redukuje się do usuwania wszystkich obserwacji. Przedstawione to zostało na rysunku 1 i nie wymaga oddzielnej prezentacji graficznej.

Zastępowanie braków danych średnią

Alternatywą do usuwania braków danych jest zastępowanie braków danych określoną wartością liczbową. Procedurę taką nazywa się imputacją. Najprostszym rodzajem imputacji jest zastępowanie nieobserwowalnych wartości wartością średnią zmiennej w próbie (średnią z próby obserwowalnych wartości zmiennych). Przedstawiono to na rysunku 2.



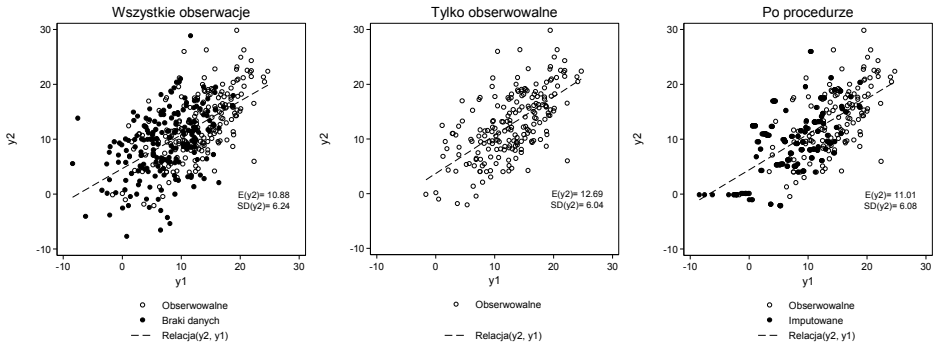
Rysunek 2. Ilustracja podejścia do analiz z brakami danych: zastępowanie braków danych średnią, mechanizm generowania braków danych: MAR

Źródło: opracowanie własne.

Jak widać procedura taka zaniża znacząco odchylenie standardowe zmiennej y_2. Musiało się tak stać, skoro 50% wartości zmiennej przypisano jedną wartość (równą wartości średniej spośród wartości obserwowalnych). Znaczne obniżenie wariancji pociąga za sobą również zmianę nachylenia krzywej reprezentującej relację między y_2 i y_1. Jak widać zastępowanie średnią działa mniej efektywnie niż wykluczenie z analizy obserwacji z brakami danych. Zastępowania braków danych średnią w sytuacji MAR wiąże się z obciążonym szacowaniem średniej y_2, wariancji y_2, relacji między y_2 i y_1 oraz błędów standardowych i opartych na nich testów statystycznych. Metoda, która na pozór zachowuje więcej informacji (nie odrzuca jednostek z brakami danych), sprawia więcej kłopotów niż korzyści.

Imputacja nieparametryczna (*hot-deck imputation*)

Bardziej złożoną procedurą zastępowania braków danych jest imputacja nieparametryczna. Polega ona na znajdowaniu jednostek „podobnych". W przypadku dwóch zmiennych, z których jedna obarczona jest brakami danych. Procedura konceptualnie jest bardzo prosta. Wybieramy parę zmiennych $y_{1(a)}$ i $y_{2(a)}$, dla której $y_{2(a)}$ nie ma obserwowalnej wartości. Następnie w grupie par zmiennych $y_{1(b)}$ i $y_{2(b)}$, gdzie nie zdarzają się braki danych szukamy tzw. „dawcy". Dawca to para zmiennych dla której $y_{1(a)} = y_{1(b)}$ lub przynajmniej wartości te są bliskie sobie. Jeżeli dawca zostanie odnaleziony, wtedy brakująca $y_{2(b)}$ zastąpiona zostanie wartością $y_{2(a)}$. W przypadku większej liczby zmiennych algorytm poszukuje dawcy jak najbardziej podobnego ze względu na cały zbiór dostępnych zmiennych.



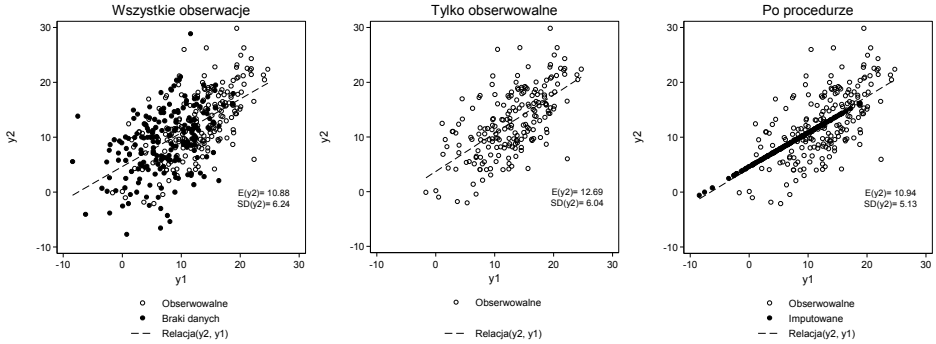
Rysunek 3. Ilustracja podejścia do analiz z brakami danych: nieparametryczne zastępowanie braków danych, mechanizm generowania braków danych: MAR

Źródło: opracowanie własne.

Na rysunku 3 przedstawiono wyniki imputacji nieparametrycznej. Jak widać rezultaty uzyskane za pomocą tej procedury zdają się być zadowalające. Zarówno średnia, odchylenie standardowe dla y_2 pozostają bliskie wartościom z całej próby. Relacja między y_1 i y_2 również nie odbiega od tej z pełnego zbioru danych. Charakterystyczne dla tego typu imputacji jest ułożenie nienaturalnej liczby punktów dla niektórych grup wartości w jednej horyzontalnej linii. Spowodowane jest to tym, że podobnym brakom danych przypisywani są podobni dawcy. Dla dużej frakcji braków danych może mieć to duże znaczenie i powodować dużą redukcję wariancji zmiennej z brakami danych, a w konsekwencji prowadzi do błędnych oszacowań większości modeli statystycznych. W sytuacji, gdy tylko jedna wartość y_2 będzie obserwowalna, wynik imputacji nieparametrycznej będzie tożsamy z wynikiem imputacji za pomocą średniej (w tym wypadku specyficznej średniej z jednej wartości).

Imputacja regresyjna

Kolejną metodą wykorzystywaną do tego, by pokonywać kłopoty związane z brakami danych jest imputacja regresyjna. Przeprowadzenie jej wymaga wykonania dwóch kroków. W pierwszym kroku należy oszacować parametry modelu regresji $y_2 = \beta_0 + \beta_1 y_1 + e$ dla zbioru obserwowalnych danych. Następnie na podstawie parametrów oszacowanych w pierwszym kroków oraz wartości y_1 należy wyznaczyć przewidywane wartości i zastąpić nimi braki danych.



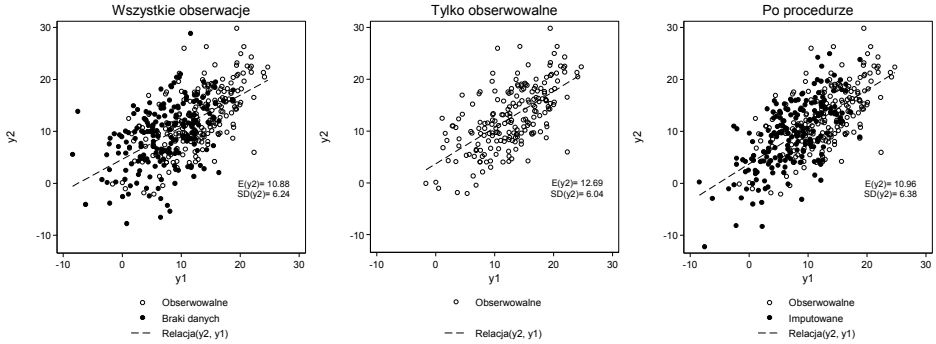
Rysunek 4. Ilustracja podejścia do analiz z brakami danych: imputacja regresyjna, mechanizm generowania braków danych: MAR

Źródło: opracowanie własne.

Na rysunku 4 przedstawiono w sposób graficzny wynik imputacji regresyjnej. Imputacja regresyjna daje charakterystyczny obraz imputowanych wartości. Wartości imputowane układają się wzdłuż linii wyznaczającej relacje między y_2 i y_1. Dzieje się tak oczywiście dlatego, że linia wyznaczająca tę relację jest zarazem linią przewidywania regresji. Imputacja regresyjna daje dobre oszacowanie średniej y_2 choć jej deterministyczny charakter sprawia, iż odchylenie standardowe tej zmiennej jest zaniżane. W konsekwencji zarówno imputacja regresyjna, jak i wcześniej opisywane metody imputacyjne spowodują, iż zdecydowana większość parametrów estymowanych modeli statystycznych będzie błędnie szacowana. Trudno zatem rekomendować opisywane wcześniej metody imputacyjne.

Stochastyczna imputacja regresyjna

W odpowiedzi na problemy związane z redukcją wariancji w imputowanych danych zaproponowana została tzw. stochastyczna imputacja regresyjna. Jest to rozwinięcie imputacji regresyjnej o jeden dodatkowy krok. Pierwsze dwa kroki znane z imputacji regresyjnej są takie same. Trzeci krok polega na dodaniu losowej wartości do przewidywanej wartości y_2. Wartość ta losowana jest na podstawie z rozkładu o średniej 0 i wariancji oszacowanej za pomocą modelu regresyjnego szacowanego w kroku pierwszym procedury, a mianowicie: *var*(e). Zakłada to oczywiście spełnienie warunku homoscedastyczności. Po spełnieniu tego klasycznego wymagania regresja stochastyczna okazuje się jedną z najlepszych klasycznych metod radzenia sobie z brakami danych.



Rysunek 5. Ilustracja podejścia do analiz z brakami danych: imputacja metodą regresji stochastycznej, mechanizm generowania braków danych: MAR

Źródło: opracowanie własne.

Na rysunku 5 w sposób graficzny przedstawiono wynik zastosowania regresji stochastycznej do zbioru danych znanego ze wcześniejszych przykładów. Widać, że za jej pomocą bardzo dokładnie oddawana jest zarówno średnia, jak i odchylenie standardowe zmiennej z brakami danych (y_2). Relacja między zmiennymi również odtwarzana jest bardzo dokładnie. Rzut oka na skrajny lewy i skrajny prawy wykres z rysunku 5 wskazuje na to iż dwuwymiarowy rozkład y_1 i y_2 danych imputowanych bardzo dobrze odzwierciedla rozkład, w którym wszystkie zmienne są w pełni obserwowalne. Uważa się, że metoda ta jest najbardziej efektywna spośród klasycznych metod radzenia sobie z brakami danych (Graham, Cumsille, Elek-Fisk, 2003; Enders 2010). Metoda ta nie jest jednak wolna od wad. Główny problem z imputacją stochastyczną polega na szacowaniu błędów standardowych i testowaniu hipotez statystycznych. Problem polega na tym, iż do obliczenia większości testów statystycznych potrzebna jest informacja o liczbie obserwacji. W przypadku danych imputowanych odpowiedź na to, ile obserwacji wykorzystanych zostało do oszacowania danego parametru nie jest trywialna. Nie jest to liczba wszystkich dostępnych obserwacji, ponieważ część posiada braki danych. Nie jest to również liczba wszystkich obserwacji bez braków danych, gdyż proces imputacyjny do pewnego stopnia rekompensuje utratę informacji spowodowaną brakami danych.

„Nowoczesne" metody radzenia sobie z brakami danych

W tej części opisane zostaną nowoczesne metody radzenia sobie z brakami danych. Precyzyjniej – metoda największej wiarygodności uwzględniająca braki danych i metoda wielokrotnych imputacji. Metody te charakteryzują się najlepszymi właściwościami statystycznymi i gwarantują uzyskanie najprecyzyjniejszych wyników zakładając mechanizmy MAR lub MCAR, gwarantując nie tylko najlepsze osza-

cowania punktowe estymatorów, lecz również poprawne oszacowania błędów standardowych (Graham, Cumsille, Elek-Fisk, 2003). Problem estymacji liczby obserwacji dla tych metod nie istnieje, w konsekwencji poprawne szacowanie błędów standardowych i testów statystycznych przez obydwie metody jest wielkim krokiem naprzód w stosunku do klasycznych metod.

Obydwie z prezentowanych „nowoczesnych" metod są asymptotycznie ekwiwalentne, tj. przy dużych próbach dają porównywalne wyniki (Graham, Cumsille, Elek-Fisk, 2003). Główna różnica polega na sposobie implementacji. Metoda największej wiarygodności uwzględniająca braki danych stosowana jest przede wszystkim w programach statystycznych szacujących modele cech ukrytych, takich jak Mplus czy LISREL. Implementacja dla użytkownika jest prosta i polega na wybraniu odpowiedniej opcji w programie, ewentualnie wyspecyfikowaniu zmiennych pomocniczych (*auxiliary variables*), czyli związanych z mechanizmem generowania braków danych, lecz nie używanych bezpośrednio w modelu.

Wielokrotne imputacje mogą być szacowane za pomocą specjalistycznych programów do estymacji bayesowskiej, takich jak STAN lub WinBUGS, JAGS oraz programów dedykowanych do generowania wielokrotnych imputacji, takich jak NORM. Stosunkowa złożoność tych programów do pewnego stopnia ograniczała dostęp do wykorzystania wielokrotnych imputacji. Sytuacja ta jednak zmieniła się, gdyż w najnowszych wersjach najpopularniejszych wielozadaniowych programów statystycznych, takich jak SPSS, Stata, R czy SAS, użytkownik uzyskał możliwość używania prostych i funkcjonalnych procedur generowania wielokrotnych imputacji. Podejście wykorzystujące wielokrotne imputacje jest bardziej wymagające, lecz również bardziej elastyczne. Raz oszacowane wielokrotne imputacje mogą być wykorzystywane w dowolnym programie statystycznym dla dowolnych modeli. Dlatego w tym artykule poświęcamy im trochę więcej uwagi.

Metoda największej wiarygodności uwzględniająca braki danych

Korzenie estymacji metodą największej wiarygodności sięgają początku lat 20. XX w. i są związane z pracą jednego z twórców nowoczesnej statystyki Sir Ronalda Fishera (Andrich, 1999). Chociaż statystyczne podstawy tej metody znane są od wielu lat, dopiero niedawno prędkość obliczeniowa nowoczesnych komputerów umożliwiła wykorzystanie wszelkich zalet tego bardzo elastycznego podejścia. Jedną z głównych zalet MNW w kontekście braków danych jest to, że metoda ta pozwala na uwzględnienie obserwacji z brakami danych w procesie estymacji wraz z obserwacjami z kompletnymi danymi w naturalny dla tego podejścia sposób. W klasycznym podejściu takie rozwiązanie nie jest możliwe – obserwacje z brakującymi danymi należy wykluczyć lub braki danych muszą zostać zastąpione oszacowanymi wartościami. MNW umożliwia wykorzystanie obserwacji z częściowo brakującymi danymi i częściowych informacji z nich płynących do oszacowań parametrów modelu statystycznego.

Stosowanie MNW do estymacji uwzględniającej braki danych konceptualnie jest bardzo proste. W przypadku większości analiz statystycznych funkcja wiarygodności jest szacowana na podstawie danych z każdej obserwacji. W przypadku obecności braków danych funkcja wiarygodności musi zostać podzielona na dwie części – dla danych zawierających braki i danych niezawierających braków. Znalezienie właściwych szacunków dla estymowanych parametrów staje się tylko problemem technicznym. Problem ten w większości wypadków rozwiązywany jest poprzez algorytm (*Expectation-Maximization*, EM; Dempster, Laird, Rubin, 1977) implementowany w większości programów zdolnych do szacowania modeli za pomocą MNW uwzględniającej braki danych.

Wielokrotne imputacje: bayesowska imputacja braków danych

Alternatywą dla MNW jest metoda wielokrotnych imputacji. Wymaga ona dwóch etapów. W pierwszym kroku generowane są zestawy imputacji, podobnie jak dla stochastycznej imputacji regresyjnej z tym, że w tym wypadku jednemu brakowi danych przypisanych jest od kilku do kilkudziesięciu wartości imputowanych. W drugim kroku imputowane wartości wykorzystywane są do szacowania interesujących badacza parametrów statystycznych.

Generowanie wielokrotnych imputacji odbywa się w paradygmacie bayesowskim i składa się z dwóch faz tzw. „kroku I" (imputacyjnego) i „kroku P" (*posterior*). W procesie tym w kolejnych krokach bayesowski model statystyczny generuje parametry opisujący relacje między zmiennymi (krok P) oraz zestawy imputowanych dla braków danych wartości (krok I). Pierwsze kroki takiego algorytmu są zwykle niestabilne, z każdym kolejnym krokiem algorytm stabilizuje się i łańcuch wyników staje się stacjonarny. Różnice wartości parametrów w kolejnych krokach odzwierciedlają losowe wahania. Teoretycznie zatem mamy możliwość wygenerowania tylu wartości imputacji, ile odbyło się iteracji. Zazwyczaj jednak liczba wielokrotnych imputacji nie przekracza kilkunastu. Zwiększanie jej ponad tę liczbę nie poprawia znacząco oszacowań (szczególnie chodzi tu o błędy standardowe) ani nie zwiększa mocy statystycznej obliczeń. Graham i współpracownicy (2007) w swoich symulacjach pokazali, że górną praktyczną granicą liczby imputacji jest 20. Dwadzieścia imputacji wystarcza dla zdecydowanej większości sytuacji. Inne analizy sugerują również, iż nawet mniejsza liczba imputacji (od 5) daje zadowalające wyniki (Enders, 2010). Algorytmy i programy komputerowe generują zatem mniejszą liczbę imputacji i robą to zazwyczaj, wykorzystując wylosowane wartości z co *n*-tego kroku I. Na przykład, chcąc wylosować 10 wartości imputacyjnych z przytaczanego przykładu, gdzie dokonano 1000 iteracji wartości, wylosowane byłyby z 100, 200, 300, 400 etc. kroku I. Przy czym zachowuje się dużą liczbę iteracji, tak aby uzyskać pewność iż łańcuch uzyskał stacjonarność oraz, że kolejne losowania i wartości imputacyjne są od siebie niezależne.

Wielokrotne imputacje mogą być stosowane do zmiennych o różnych rozkładach (nie ma tutaj założenia o normalności lub wielowymiarowej normalności). Wielokrotne imputacje można zastosować do każdego typu danych. Jedyną ich wadą, jaką warto wyraźnie podkreślić jest wrażliwość na błędną lub niekompletną specyfikację modelu. Błędy mogą powstać jeżeli używana zmienna w analizach nie została użyta również do generowania wielokrotnych imputacji. Mianowicie relacja między zmienną wykorzystywaną w analizach lecz nie używaną do generowania wielokrotnych imputacji a inną zmienną będzie niedoszacowana. Dodatkowo nie tylko zmienne, ale i interakcje między nimi, jeżeli takie będą stosowane, muszą zostać uwzględnione w modelu imputacyjnym. Stosowanie wielokrotnych imputacji wymusza zatem przewidywanie modelu statystycznego przed procesem imputacji, a w praktyce oznacza iż model imputacyjny powinien zawierać tak dużo zmiennych, jak to tylko możliwe (Rubin, 1996).

Szacowanie statystyk na podstawie wielokrotnych imputacji

Po wygenerowaniu wielokrotnych imputacji dane organizowane są zwykle w osobne zbiory danych. W każdym zbiorze danych znajdują się dane obserwowalne oraz wygenerowane imputacje. W każdym ze zbiorów danych przeprowadza się osobna analizę, dowolnym narzędziem statystycznym wykorzystywanym w analizach na pełnych zbiorach danych. Po przeprowadzeniu tych analiz badacz zostaje z *m* liczbą wyestymowanych parametrów i *m* liczbą błędów standardowych. Wyniki te należy zagregować zgodnie z tzw. regułą Rubina, która w przypadku punktowych wartości parametrów jest prostą średnią:

$$\bar{\theta} = \frac{1}{m}\sum_{t=1}^{m} \hat{\theta}_t$$

gdzie $\hat{\theta}_t$ jest parametrem estymowanym ze zbioru danych *t*, a *m* to całkowita liczba imputacji. $\bar{\theta}$ można interpretować zarówno z perspektywy klasycznej statystyki częstościowej, jako estymację stałej wartości populacyjnej, jak również z perspektywy bayesowskiej, jako przybliżenie rozkładu *a posteriori* dla badanego parametru (Little, Rubin, 1987).

Zagregowane błędy standardowe są średnią błędów standardowych oszacowanych dla każdego zbioru danych powiększone o wariancje między oszacowaniami błędów standardowych (Little, Rubin, 1987):

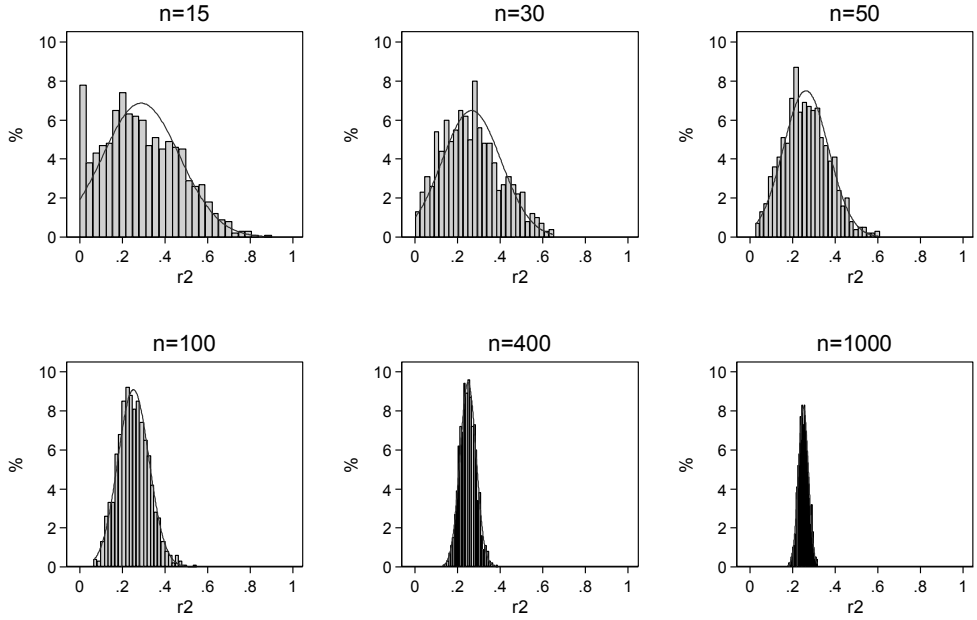
$$SE(\overline{\theta}) = \sqrt{V_w + V_B + {}^{V_B}/_{m}}$$

gdzie V_w jest wariancją wewnątrzimputacyjną (*within-imputation variance*), a V_b jest wariancją międzyimputacyjną (*between-imputation variance*) kwantyfikującą zróżnicowanie między oszacowaniami dla różnych imputacji. Bardziej formalnie:

$$V_w = \frac{1}{m}\sum_{t=1}^{m} SE_t^2 \qquad V_B = \frac{1}{m-1}\sum_{t=1}^{m}(\hat{\theta}_t - \bar{\theta})^2$$

gdzie SE^2 to kwadrat odchylenia standardowego parametru dla danych t.

Agregowanie parametrów zakłada, że są one asymptotycznie normalne. Jest to problematyczne szczególnie w przypadku małych próbek. Dla przykładu rozkład współczynnika korelacji Pearsona z próby jest normalny, gdy wartość populacyjna korelacji wynosi zero, skośność rozkładu tej statystyki z próby rośnie, gdy wartość bezwzględna korelacji rośnie w populacji. Skośne rozkłady statystyk z próby mają też parametry oparte na wariancji: R^2, odchylenie standardowe, korelacje wewnątrzgrupowe etc. Skośność tych statystyk maleje jednak wraz z liczebnością próby. W zasadzie, gdy próba liczy powyżej 400 jednostek, skośność znika. Przykład dla statystyki $R^2 = 0{,}25$ i próbek liczebności 15, 30, 50, 100, 400, 1000 przedstawiono na rysunku. Należy jednak dodać, iż w wielu warunkach wielokrotna imputacja jest dość odporna na złamanie założenia o normalności rozkładu (Leite, Beretvas, 2010), tj. potencjalne błędy wynikające ze złamania założeń odnoszących się do normalności rozkładu nie obciążają znacznie wyników przeprowadzanych analiz.



Rysunek 6. Symulacyjne rozkłady statystyk z próby dla $R^2 = 0{,}25$ i próbek liczebności 15, 30, 50, 100, 400, 1000

Źródło: opracowanie własne.

Ilość informacji a braki danych

Razem z rozwojem metod imputacji powstały dwa popularne wskaźniki opisujące ilość informacji utraconych za pomocą braków danych zaproponowane przez Rubina (1987). Pierwszy z nic to FMI (*Fraction of Missing Information*). Można go zapisać następującym wzorem:

$$\text{FMI}=\frac{V_B+V_B/m}{V_T}$$

W mianowniku tego równania znajduje się wariancja z próby ($V_t = (SE(\overline{\theta})^2$), a licznik wyraża dodatkową wariancje wynikającą z braków danych. Jest to zatem stosunek całej wariancji z próby estymowanego parametru do wariancji wynikającej z braków danych i mówi nam jaki procent wariancji estymatora wynika z braków danych.

Podobnym wskaźnikiem jest RIV (*Relative Increase in Variance*), który można wyrazić wzorem:

$$\text{RIV}=\frac{V_B+V_B/m}{V_W}=\frac{\text{FMI}}{1-\text{FMI}}$$

Relatywny wzrost wariancji wskazuje na proporcjonalny wzrost w wariancji estymatora, który jest spowodowany brakami danych. Innymi słowy, mówi nam, o ile zmienia się wariancja estymatora przez to, że mamy do czynienia z brakami danych w stosunku do sytuacji bez braków danych. Więcej informacji o tych wskaźnikach można znaleźć w książkach Rubina (1987) i Endersa (2009).

Zamiast podsumowania. Praktyczne uwagi dotyczące braków danych

Braki danych stanowią element niemalże każdego zbioru danych z zakresu badań społecznych. Czasem problem braków danych jest marginalny, w innych sytuacjach liczba braków danych może być znacząca. Nigdy jednak braków danych nie należy ignorować. Inspekcja braków danych powinna być jednym z pierwszych kroków każdej analizy statystycznej. Pozostawienie problemu braków danych na łaskę domyślnych procedur stosowanych w popularnych programach statystycznych nie będzie prowadzić do rozwiązań optymalnych i prawie zawsze będą one błędne (dla sytuacji MAR i NMAR, gdyż większość programów usuwa braki danych z analizy).

Niektórzy autorzy (Newman, 2014) sugerują, iż „nowoczesne" metody radzenia sobie z brakami danych stosowane powinny być w sytuacji, gdy braki danych przekraczają 10%. Argumentują to tym, że gdy proporcja braków danych jest niska, „nowoczesne" metody są działaniem zbędnym i nieekonomicznym. Trudno jednak zaakceptować prostą wartość brzegową w postaci 10%. W różnych sytuacjach, dla

różnych mechanizmów generowania braków danych, 10% może mieć diametralnie różne znaczenie. Dodatkowo w dobie szybkich komputerów i powszechnego dostępu do nowoczesnych metod radzenia sobie z brakami danych postulat ekonomiczności przestaje mieć duże znaczenie. Problem braków danych każdorazowo powinien zostać przemyślany, a sposób rozwiązania tej kwestii – bezpośrednio opisany i uargumentowany. Jeżeli badacz stwierdza, iż braki danych mogą mieć istotne znaczenie dla analizy, rekomendowane są MNW lub wielokrotne imputacje. W sytuacji MCAR przy małej liczbie braków danych wykluczanie jednostek obserwacji z analizy nie doprowadzi do znacznych obciążeń estymacji parametrów i może być stosowane. Klasycznych metod imputacyjnych, z wyjątkiem imputacji stochastycznej, powinno się unikać.

Jeżeli odsetek danych jest znaczący i/lub prawdopodobne jest, iż mechanizm generowania braków danych nie spełnia kryteriów MAR, należy przeprowadzić dodatkowe analizy mające na celu oszacowanie prawdopodobnych błędów. Warto też dokonać analizy wrażliwości (*sensitivity analysis*) modeli dla różnych scenariuszy i mechanizmów generowania braków danych (Newman, Sin, 2009; Newman, 2010). Newman (2014) sugeruje, że dodatkowe analizy powinny zostać przeprowadzone każdorazowo, gdy poziom braków danych przekracza 30% dla interesujących badacza zmiennych.

Warto zaznaczyć, że gdy mówimy o brakach danych, mamy do czynienia z kilkoma ich rodzajami. Podstawowy podział to jednostkowe i pozycyjne braki danych. Z pierwszym przypadkiem (ang. *unit nonresponse*) mamy do czynienia, gdy respondent nie udziela odpowiedzi na wszystkie pytania w badaniu (na przykład na skutek odmowy). Drugi przypadek (ang. *item nonresponse*) napotykamy, gdy respondent nie udziela odpowiedzi na niektóre pytania (na przykład na skutek ich drażliwości). Zarówno MNW, jak i wielokrotne imputacje mogą być stosowane w obydwu tych sytuacjach.

Inne rozróżnienie braków danych wyszczególnia braki danych na poziomie skal (ang. *scale/construct-level*) oraz braki danych na poziomie pytań (ang. *item-level*). Rozróżnienie to staje się istotne, gdy respondent odpowiada tylko na część pytań skali (np. skali poczucia własnej wartości, skali lęku, testu na inteligencję etc.), inne pozostawiając bez odpowiedzi. W takiej sytuacji rekomendowanymi metodami radzenia sobie z barkami danych pozostają MNW oraz wielokrotne imputacje. Należy jednak zdecydować czy koncentrować się na poszczególnych pytaniach, czy raczej uznać, iż w przypadku respondenta, który odpowiedział tylko na część pytań, wartość dla całej skali uznać jako brak danych. Powszechnie przyjmowaną rekomendacją jest przyjęcie pierwszej strategii (Newman, 2014). Jeśli respondent odpowiedział choćby na jedno pytanie skali, rozwiązania dotyczące braków danych powinny odnosić się do poszczególnych pytań, nie zaś do całych skal.

Literatura cytowana

Allison, P.D. (2001). *Missing data*. Thousand Oaks: Sage.

Enders, C.K. (2010). *Applied missing data analysis*. New York-London: Guilford Press.

Graham, J.W., Cumsille, P.E., Elek-Fisk, E. (2003). Methods for Handling Missing Data. W: J.A. Schinka, W.F. Velicer (red.), *Handbook of Psychology* (t. 2, s. 87-114). Hoboken, NJ: John Wiley & Sons, Inc.

Graham, J.W., Olchowski, A.E., Gilreath, T.D. (2007). How Many Imputations are Really Needed? Some Practical Clarifications of Multiple Imputation Theory. *Prevention Science*, 8 (3), 206-213, doi: 10.1007/s11121-007-0070-9

Heckman, J. (2013). Sample selection bias as a specification error. *Applied Econometrics*, 31 (3), 129-137.

Leite, W., Beretvas, S. (2010). The Performance of Multiple Imputation for Likert--type Items with Missing Data. *Journal of Modern Applied Statistical Methods*, 9 (1), 64-74.

Little, R.J.A., Rubin, D.B. (1987). *Statistical Analysis with Missing Data*. New York: John Wiley & Sons, Inc.

McKnight, P.E., McKnight, K.M., Sidani, S., Figueredo, A.J. (2007). *Missing Data: A Gentle Introduction*. New York: Guilford Press.

Molenberghs, G., Kenward, M.G. (2007). *Missing Data in Clinical Studies*. Chichester: John Wiley & Sons, Inc.

Newman, D.A. (2010). Missing Data Techniques and Low Response Rates. W: C.E. Lance, R.J. Vandenberg (red.), *Statistical and Methodological Myths and Urban Legends: Doctrine, Verity and Fable in Organizational and Social Sciences*. New York: Routledge.

Pokropek, A. (2011). Missing by design: Planned missing-data designs in social science. *Ask: Research & Methods*, 20 (1), 81-105.

Rubin, D.B. (1976). Inference and Missing Data. *Biometrika*, 63 (3), 581-592, doi: 10.2307/2335739

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons, Inc.

Rubin, D.B. (1996). Multiple Imputation after 18+ Years. *Journal of the American Statistical Association*, 91 (434), 473-489, doi: 10.1080/01621459.1996.10476908

Schafer, J.L., Graham, J.W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7 (2), 147-177, doi: 10.1037/1082-989X.7.2.147

Aneks – Statystyczny opis mechanizmów powstawania braków danych

Podstawowe pojęcia i notacja

Dla każdej jednostki $i = 1, \ldots, N$ zbierany jest zestaw zmiennych $j = 1, ..., n_j$, gdzie n_j to liczba zmiennych. Y_{ij} określa wartość zmiennej dla i-tego respondenta i j-tej zmiennej. Pełny zbiór zmiennych oznaczymy jako $\boldsymbol{Y}$ i nazwijmy go zbiorem odpo-

wiedzi (mając na myśli, iż w praktycznych zastosowaniach realizacja tej macierzy oparta będzie najczęściej na odpowiedziach respondentów na zadawane im pytania). Dodatkowo dla każdej zmiennej *j* obserwacji *i* zdefiniowany zostanie wskaźnik R_{ij}. Nazwiemy go wskaźnikiem braków danych, mówiącym o tym, czy obserwowalna jest wartość zmiennej dla konkretnej obserwacji czy też nie, tak że:

$$R_{ij} = \begin{cases} 1, \text{ jeżeli } Y_{ij} \text{ jest obserwowalne,} \\ 0, \text{ jeżeli } Y_{ij} \text{ nie jest obserwowalne.} \end{cases} \quad (1)$$

Podobnie jak w przypadku $\boldsymbol{Y}$ pełny zbiór wskaźników braków odnosić będziemy do $\boldsymbol{R}$. Dla przedstawienia mechanizmów generujących braki danych i konsekwencji dla estymacji z nich płynących wygodnie będzie rozdzielić $\boldsymbol{Y}$ na dwie części, tak że $\boldsymbol{Y} = (\boldsymbol{Y}_0, \boldsymbol{Y}_m)$, gdzie:

$$\mathbf{Y} = \begin{cases} \mathbf{Y_0} \text{ zawiera } Y_{ij}, \text{ dla których } R_{ij} = 1, \\ \mathbf{Y_m} \text{ zawiera } Y_{ij}, \text{ dla których } R_{ij} = 0. \end{cases} \quad (2)$$

Macierz **R** określającą strukturę braków danych. Proces generujący macierz nazywać będziemy procesem powstawania braków danych (*missing data process*). Przez pełny zbiór danych należy rozumieć zbiór zawierający kompletną informację o wszystkich wartościach zmiennych oraz macierz określającą strukturę braków danych ($\mathbf{Y_i}$, $\mathbf{R_i}$).

Modele statystyczne i braki danych

W sytuacji, gdzie nie mamy do czynienia z brakami danych, informacja o zmiennych losowych $\mathbf{R_i}$ może zostać pominięta. Model statystyczny, dzięki któremu możemy dokonywać wnioskowania na temat parametrów stojących za analizowanymi danymi (oznaczonych macierzą $\boldsymbol{\theta}$) nazywamy modelem pełnych odpowiedzi (*full response model*). W sposób ogólny można zapisać go następująco:

$$p(\mathbf{y} \mid \boldsymbol{\theta}) \quad (3)$$

Jest to model probabilistyczny indeksowany parametrem $\boldsymbol{\theta}$, gdzie $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, a $\boldsymbol{\Theta}$ jest przestrzenią parametrów[1]. W przypadku, gdy problem braków danych istnieje, wnioskowanie dotyczy wybranej przez analityka funkcji parametru $\boldsymbol{\theta}$, np. $\mu(\boldsymbol{\theta}) = E(\mathbf{Y} \mid \boldsymbol{\theta})$

[1] Sformułowanie tego modelu może być bardziej złożone i wprowadzać podział na zmienne zależne i niezależne (Molenberghs and Kenward 2007) tak, że wzór (3) rozwinie się do postaci $p(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\theta})$. Dla celów tego rozdziału takie rozwinięcie nie jest konieczne i pozostaniemy tu przy prostej formie.

albo w przypadku bardziej złożonych modeli $\boldsymbol{\beta}(\boldsymbol{\theta}) = E(\mathbf{Y} \mid \mathbf{X}, \boldsymbol{\beta})$ dla modelu regresji liniowej. Inaczej mówiąc badacza interesują statystyki, czyli estymatory prawdziwych parametrów.

Gdy problem braków danych pojawia się, model (3) staje się niewystarczający (w większości wypadków) i należy się odnieść do bardziej ogólnego, tak zwanego pełnego modelu (*full data model*):

$$p(\mathbf{y}, \mathbf{r} \mid \boldsymbol{\omega}) \tag{4}$$

Model (4) opisuje łączne prawdopodobieństwo $\mathbf{Y}$ i $\mathbf{R}$ oraz jest indeksowany przez parametry $\boldsymbol{\omega}$, gdzie $\boldsymbol{\omega} \in \Omega$, a Ω jest przestrzenią parametrów. Gdy podejmujemy problem braków danych, zakładamy, że $\boldsymbol{\theta}$ jest funkcją parametru $\boldsymbol{\omega}$ indeksującego bardziej ogólny model $\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{\omega})$. W sytuacji braków danych musimy bowiem przeprowadzić wnioskowanie, które z modelu (4) (z brakami danych) prowadziło będzie do wnioskowania analogicznego do modelu (3) (gdzie braki danych nie istnieją). Inaczej mówiąc, szukając nieobciążonych estymatorów $\boldsymbol{\theta}$, musimy brać również pod uwagę parametry modelu, które definiują nam mechanizmy tworzące braki danych.

Przedstawiony w równaniu model (4) można rozłożyć na kilka sposobów (Molenberghs i Kenward 2007). Najpopularniejszym i w większości przypadków najwygodniejszym sposobem faktoryzacji funkcji (4) jest sposób rozkładu, który każe myśleć o modelowanych danych w kategoriach modelu selekcji:

$$p(\mathbf{y}, \mathbf{r} \mid \boldsymbol{\omega}) = p(\mathbf{y} \mid \boldsymbol{\theta}\,(\boldsymbol{\omega}))\; p(\mathbf{r} \mid \mathbf{y}, \boldsymbol{\psi}(\boldsymbol{\omega})) \tag{5}$$

Pierwszy człon sfaktoryzowanej funkcji jest modelem pełnych odpowiedzi (równanie 3). Drugi człon to model opisujący rozkład braków danych warunkowy ze względu na wartości odpowiedzi ($\mathbf{y}$) i indeksowany przez parametr $\boldsymbol{\psi}$ będący funkcją parametru pełnego modelu $\boldsymbol{\psi} = \boldsymbol{\psi}(\boldsymbol{\omega})$, gdzie $\boldsymbol{\psi} \in \boldsymbol{\Psi}$, a $\boldsymbol{\Psi}$ jest przestrzenią parametrów. Nazwa modelu selekcji bierze się z popularnego w ekonomii modelu sformułowanego przez Heckmana (1976), służącego do modelowania mechanizmu powodującego braki danych tak, by uzyskać nieobciążone parametry populacji (Heckman, 1979, s. 153-161). W tym tekście część modelu (5) odnoszącego się do braków danych częściej będziemy jednak nazywać modelem braków danych, aby odróżnić nasze konkretne zainteresowanie od szerokiej klasy modeli selekcji.

Mechanizm powodujący powstawanie braków danych

Mechanizm generujący braki danych w podejściu wywodzącym się od Rubina rozumiany jest jako proces stochastyczny przejawiający się uwidacznianiem braków danych w $\mathbf{Y}$. Analizę mechanizmu generowania braków danych można rozpocząć od części równania (5) odnoszącego się do modelu selekcji. Wygodnie jest rozdzielić wektor zmiennych Y na dwie części (tak jak było to opisane we wzorze 2) i zapisać mechanizm powstawania braków danych jako:

$$p(\mathbf{r} \mid \mathbf{y}, \mathbf{x}, \boldsymbol{\psi}) = p(\mathbf{r} \mid \mathbf{y}_0, \mathbf{y}_m, \mathbf{x}, \boldsymbol{\psi}) \tag{6}$$

Odwołując się do tak sformułowanego modelu, przedstawimy trzy podstawowe założenia formułujące trzy typy mechanizmów stojących za brakami danych: mechanizm całkowicie losowy (MCAR: *Missing completely at random*), mechanizm losowy (MAR: *Missing at random*), mechanizm nielosowy (MNAR: *Missing not at random*).

Mechanizm całkowicie losowy (MCAR: Missing completely at random)

Mechanizm powstawania braków danych jest całkowicie losowy (MCAR), gdy braki danych nie są związane ani z obserwowalną częścią $\mathbf{Y}$, ani z nieobserwowalną częścią, tak że:

$$p(\mathbf{r} \mid \mathbf{y}, \psi) = p(\mathbf{r} \mid \psi) \tag{7}$$

Innymi słowy, pojawianie się braków danych nie jest związane z niczym oprócz losowego procesu. Takie założenie ma poważne konsekwencje dla modelu pełnego (5). Zakładając, że równanie (7) jest prawdzie oraz że parametry θ i ψ są niezależne, model (5) możemy zapisać następująco:

$$p(\mathbf{y}, \mathbf{r} \mid \omega) = p(\mathbf{y} \mid \theta)\, p(\mathbf{r} \mid \psi) \tag{8}$$

Implikacja tego zapisu jest niezwykle ważna. Jeżeli będziemy potrafili opisać model pełny tylko za pomocą danych obserwowalnych, możliwym stanie się nieobciążona estymacja parametrów populacyjnych.

Przyjmując założenie MCAR można pokazać, że model pełny daje się uprościć całkując po części odnoszącej się do braków danych $\boldsymbol{y}_{\mathbf{m}}$ tak, że:

$$\int p(\mathbf{y}, \mathbf{r} \mid \omega) d\mathbf{y}_{\mathbf{m}} = \int p(\mathbf{y}_0, \mathbf{y}_{\mathbf{m}} \mid \theta)\, p(\mathbf{r} \mid \psi) d\mathbf{y}_{\mathbf{m}} = p(\mathbf{y}_0 \mid \theta)\, p(\mathbf{r} \mid \psi) \tag{9}$$

Jako że przy założeniu MCAR $p(\mathbf{r} \mid \psi)$ jest stałą, nie ma ona praktycznego znaczenia podczas modelowania większości parametrów. Widać zatem, iż jeżeli mamy do czynienia z mechanizmem MCAR modelowanie z brakami danych nie powinno nastręczać większych trudności.

Mechanizm losowy (MAR: Missing at random)

Mechanizm losowy zakłada, iż prawdopodobieństwo powstania braków danych związane jest z obserwowalną częścią zmiennych wykorzystywanych w analizie, nie jest jednak związane z częścią nieobserwowalną, tak że:

$$p(\mathbf{r} \mid \mathbf{y}, \psi) = p(\mathbf{r} \mid \mathbf{y}_0, \psi) \tag{10}$$

Tym samym model pełny można zapisać następująco:

$$p(\mathbf{y}, \mathbf{r} \mid \omega) = p(\mathbf{y} \mid \theta)\, p(\mathbf{r} \mid \mathbf{y}_0, \psi) \tag{11}$$

Podobnie jak w przypadku silniejszego założenia (MCAR) przy założeniu MAR model pełny da się uprościć całkując po części odnoszącej się do braków danych $\boldsymbol{y}_{\mathbf{m}}$ tak, że:

$$\int p(\mathbf{y}, \mathbf{r} \mid \boldsymbol{\omega}) d\mathbf{y}_{\mathbf{m}} = \int p(\mathbf{y}_0, \mathbf{y}_{\mathbf{m}} \mid \boldsymbol{\theta})\, p(\mathbf{r} \mid \mathbf{y}_0, \boldsymbol{\psi}) d\mathbf{y}_{\mathbf{m}} = p(\mathbf{y}_0 \mid \boldsymbol{\theta})\, p(\mathbf{r} \mid \mathbf{y}_0, \boldsymbol{\psi}) \tag{12}$$

Przy założeniu MAR możliwe jest zatem opisanie modelu pełnego jedynie za pomocą danych obserwowalnych. Co prawda mechanizm powodujący powstawanie braków danych nie jest tutaj tak bezproblemowy jak w MCAR, ale daje się modelować za pomocą danych obserwowalnych $p(\mathbf{r} \mid \mathbf{y}_0, \boldsymbol{\psi})$. Otwiera to możliwości do estymacji nieobciążonych parametrów modelu odpowiedzi jedynie za pomocą danych obserwowalnych, przy kontroli mechanizmu powstawania braków danych.

Mechanizm nielosowy (MNAR: Missing not at random)

Nielosowy mechanizm powstawania braków danych oznacza, że prawdopodobieństwo pojawienia się braku danych w $\mathbf{Y}$ zależy zarówno od danych obserwowalnych ($\boldsymbol{Y}_0$), jak i nieobserwowalnych ($\boldsymbol{Y}_{\mathbf{m}}$). Przyjęcie założeń MNAR uniemożliwia jakiekolwiek sensowne uproszczenie modelu pełnego tak, by umożliwiało to szacowanie parametrów modelu bez informacji o brakach danych (tak jak w równaniach 9 i 12):

$$\int p(\mathbf{y}, \mathbf{r} \mid \boldsymbol{\omega}) d\mathbf{y}_{\mathbf{m}} = \int p(\mathbf{y}_0, \mathbf{y}_{\mathbf{m}} \mid \boldsymbol{\theta})\, p(\mathbf{r} \mid \mathbf{y}_0, \mathbf{y}_{\mathbf{m}}, \boldsymbol{\psi}) d\mathbf{y}_{\mathbf{m}} = p(\mathbf{y}_0 \mid \boldsymbol{\theta})\, p(\mathbf{r} \mid \mathbf{y}_0, \mathbf{y}_{\mathbf{m}}, \boldsymbol{\psi}) \tag{13}$$

Innymi słowy, gdy przyjmiemy założenie MNAR do poprawnego wnioskowania o parametrach interesujących badacza niezbędne są dodatkowe informacje o części nieobserwowalnej zbioru oraz o parametrach modelu w którym prawdopodobieństwa braków danych są modelowane ($\boldsymbol{\psi}$).

Streszczenie. W artykule przedstawiono wybrane współczesne statystyczne metody radzenia sobie z brakami danych. Artykuł opiera się na podejściu Rubina, który zaproponował trzy typy mechanizmów generujących braki danych: mechanizm całkowicie losowy (MCAR: *Missing completely at random*), mechanizm losowy (MAR: *Missing at random*), mechanizm nielosowy (MNAR: *Missing not at random*). Przedstawione i ocenione zostały zarówno klasyczne, jak i „nowoczesne" metody radzenia sobie z brakami danych, takie jak: usuwanie braków danych, zastępowanie średnią, imputacja regresyjna, stochastyczna imputacja regresyjna, nieparametryczna imputacja *hot deck*, metoda największej wiarygodności uwzgledniająca braki danych i wielokrotne imputacje. Artykuł kończy się praktycznymi wskazówkami dotyczacymi radzenia sobie z brakami danych.
Słowa kluczowe: braki danych, metoda największej wiarygodności, wielokrotne imputacje, imputacje nieparametryczne

Data wpłynięcia: 20.10.2017
Data wpłynięcia po poprawkach: 5.03.2018
Data zatwierdzenia tekstu do druku: 31.03.2018