WIADOMOŚCI STATYSTYCZNE

PL ISSN 0043-518X

Rok LXI
9 (664) 2016, 7—54

STUDIA METODOLOGICZNE

Jacek WESOŁOWSKI
Jakub TARCZYŃSKI

Podstawy matematyczne technik imputacyjnych

Streszczenie. *W artykule przedstawiono podstawy metodologii imputacyjnej (w tym metodologii wielokrotnej imputacji), koncentrując się na wyjaśnieniu matematycznej strony zagadnień. Analizowano sytuację, gdy obserwacje tworzące pierwotną próbkę są niezależnymi zmiennymi losowymi o jednakowym rozkładzie, a braki odpowiedzi pojawiają się losowo w sposób niezależny od obserwacji. W szczególności wskazano na problemy pojawiające się, gdy w imputacji wielokrotnej stosowany jest standardowy estymator Rubina wariancji estymatora wielokrotnej imputacji i wskazano na możliwe ulepszenie tego popularnego estymatora. Punktem wyjścia analiz jest sytuacja, gdy za pojawianie się braków odpowiedzi odpowiada mechanizm deterministyczny*.

Słowa kluczowe: imputacja, imputacja wielokrotna, estymator imputacyjny, estymator Rubina, imputacja średnią, imputacja typu hot-deck, imputacja regresyjna.

Słowem imputacja określa się zespół metod uzupełniania braków w próbce w taki sposób, aby zastosowanie estymatorów standardowych, tzn. takich, które byłyby użyte, gdyby w próbce nie było braków obserwacji, prowadziło do efektywnej procedury estymacyjnej.

Literatura dotycząca technik imputacyjnych jest bardzo obszerna. Składają się na nią setki specjalistycznych artykułów w czasopismach statystycznych, artykuły przeglądowe, np. Andridge i Little (2010), Donders, van der Heijden, Stijnen i Moons (2006) czy Norazian Ramli, Yahaya, Ramli i Yusof (2013) oraz kilka większych publikacji monograficznych, takich jak: Rubin (1987), Little i Rubin

(2002), de Waal, Pannekoek i Scholtus (2011) czy Garson (2012) lub van Buuren (2012). Literatura ta najczęściej dotyczy opisu konkretnych metod, ich zastosowań i porównań. Techniki te bywają dostępne w pakietach statystycznych (Horton i Lipsitz, 2001; van Buuren i Groothuis-Oudshoorn, 2011; Misztal, 2012). We wszystkich tych publikacjach stosunkowo mały nacisk kładziony jest na podstawy teoretyczne.

Okazuje się, że w standardowym modelu statystycznym niezależnych obserwacji o jednakowych rozkładach daje się precyzyjnie opisać (w formie twierdzeń matematycznych) podstawy teoretyczne technik imputacyjnych. Przedstawiane opracowanie jest próbą takiego właśnie systematycznego i matematycznego uściślenia uniwersalnych paradygmatów imputacji. W rozdziale I rozważana jest uproszczona sytuacja, gdy zbiór braków obserwacji jest ustalony. Przy tym założeniu wyprowadzono wzory ogólne na estymator imputacyjny średniej i jego parametry oraz zastosowano je w czterech podstawowych schematach imputacyjnych: imputacji średnią, dwóch procedurach typu hot-deck oraz imputacji regresyjnej. W rozdziale II analizowano imputację wielokrotną w modelu z rozdziału I, a następnie opracowany schemat ogólny wykorzystano w sytuacji, gdy w każdym kroku procedury stosowana jest imputacja typu hot-deck. Używany w literaturze termin „imputacja hot-deck", w tym opracowaniu — nieco rozszerzony — odnosi się do sytuacji, gdy elementy imputowane są zgodnie z pewną procedurą losową, która zależna jest od obserwowanej części próbki.

W rozdziałach III i IV rozważany jest bardziej realistyczny losowy model pojawiania się braków odpowiedzi, tzn. taki, w którym każda obserwacja może być obecna w próbce lub nie, przy czym zdarzenie to ma charakter losowy, czyli można mu przypisać pewne prawdopodobieństwo. Poprzez zastosowanie techniki warunkowania, rezultaty otrzymane w rozdziałach III i IV można stosunkowo prosto wywieść z wyników przedstawionych w rozdziałach I i II. W takim podejściu istotną rolę odgrywa przyjęte w artykule założenie o niezależności statystycznej obserwacji i braków odpowiedzi. W literaturze mówi się w takiej sytuacji o brakach odpowiedzi typu MCAR (*missing completely at random*), np. Little i Rubin (2002). To podejście jest ważne, bo choć często nie jest właściwym modelem dla całej populacji, to okazuje się dobrym przybliżeniem sytuacji rzeczywistej w odpowiednio wybranych podpopulacjach (chodzi o klasy elementów podobnych pod względem skłonności do udzielania odpowiedzi).

Należy zwrócić uwagę na, prezentowane w opracowaniu, ulepszenie klasycznego estymatora Rubina służącego do estymacji wariancji estymatora imputacyjnego w przypadku imputacji wielokrotnej. Zasadnicza część artykułu ma matematyczną strukturę twierdzeń i dowodów. Najciekawsze wyniki matematyczne (w rozdziałach II i IV) zilustrowano eksperymentami numerycznymi. Konkluzje przedstawiono w rozdziale V.

I. DETERMINISTYCZNY ZBIÓR BRAKÓW OBSERWACJI — IMPUTACJA JEDNOKROTNA

I.0. Ogólny schemat imputacyjny

Zmienne losowe $X_1, ..., X_n$ stanowią pierwotną próbkę, $\boldsymbol{X} = (X_1, ..., X_n)$. Niech $\hat{\theta} = g(\boldsymbol{X})$ oznacza estymator parametru θ, gdzie g jest odpowiednią funkcją n-argumentową. Rozważać będziemy sytuację, gdy obserwowane są tylko zmienne X_i, $i \in R \subsetneq \{1, ..., n\}$, czyli R jest zbiorem indeksów obserwowanych elementów próbki pierwotnej. W konsekwencji pierwotny wektor obserwacji dzieli się naturalnie na dwa wektory: wektor odpowiedzi obserwowanych $\boldsymbol{X}_R = (X_i, i \in R)$ oraz wektor odpowiedzi nieobserwowanych $\boldsymbol{X}_{R^c} = (X_i, i \in R^c = \{1, ..., n\} \setminus R)$. Taką niepełną próbkę $\boldsymbol{X}_R$ można uzupełnić przyjmując $\widetilde{\boldsymbol{X}} = \left(\widetilde{X}_i, ..., \widetilde{X}_n\right)$, gdzie $\widetilde{X}_i = X_i$ dla $i \in R$ są zmiennymi obserwowanymi oraz $\widetilde{X}_i$, $i \in R^c$ są tzw. zmiennymi imputowanymi, które mogą zależeć (i często zależą) od zmiennych obserwowanych. Często zmienne są imputowane pewnymi funkcjami tych zmiennych, ale zdarzają się też sytuacje, w których nie ma bezpośredniej zależności funkcyjnej, a zmienne obserwowane wpływają jedynie na rozkład prawdopodobieństwa zmiennych imputowanych.

Definicja I.0.1. Imputacyjną wersją estymatora $\hat{\theta} = g(\boldsymbol{X})$ parametru θ nazywamy estymator postaci

$$\widetilde{\theta}_{Imp} = g(\widetilde{\boldsymbol{X}}).$$

W skrócie mówimy, że $\widetilde{\theta}_{Imp}$ jest estymatorem imputacyjnym parametru $\widetilde{\theta}$.

Zakładamy, że zmienne $X_1, ..., X_n$ są niezależne i mają jednakowe rozkłady, takie jak pewna zmienna losowa X o wartości oczekiwanej $\mathrm{E}\, X = \mu$ i wariancji $\mathrm{Var}\, X = \sigma^2$, które są nieznane.

Do estymacji średniej $\theta = \mu$ standardowo stosowany jest estymator

$$\hat{\mu} = g(\boldsymbol{X}) = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Jest to estymator nieobciążony, czyli $\mathrm{E}\, \hat{\mu} = \mu$, a jego wariancja wynosi $\mathrm{Var}\, \hat{\mu} = \dfrac{\sigma^2}{n}$.

Do estymacji parametru σ^2 standardowo stosowany jest estymator

$$S^2 = S^2(\boldsymbol{X}) = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - g(\boldsymbol{X})\right)^2 \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2.$$

Jest to estymator nieobciążony, czyli $\mathrm{E}\, S^2 = \sigma^2$.

Zatem nieobciążony estymator $d^2\left(\hat{\mu}\right)$ wariancji $\mathrm{Var}\,\hat{\mu}$ estymatora $\hat{\mu}$ ma postać $d^2\left(\hat{\mu}\right) = \frac{1}{n}S^2$.

Z centralnego twierdzenia granicznego (i tzw. twierdzenia Słuckiego) wynika, że $\frac{\hat{\mu}-\mu}{S}\sqrt{n}$ ma asymptotycznie rozkład standardowy normalny, co pozwala na wnioskowanie o przybliżonej precyzji estymatorów, rozumianej jako przedział ufności na zadanym poziomie ufności.

Zadanie imputacji pojawia się, gdy obserwowane są tylko zmienne losowe X_i, $i \in R \subsetneq \{1, ..., n\}$. Polega ono na uzupełnieniu obserwowanej części próbki o część nieobserwowaną $\widetilde{X}_i$, $i \in R^c$. Niech r oznacza liczbę obserwowanych elementów próbki, czyli $r = \#(R)$. Dla potrzeb rozdziałów I i II zakładamy, że podzbiór $R \subsetneq \{1, ..., n\}$, a zatem i jego liczebność r jest ustalona. W kolejnych rozdziałach odejdziemy od tego założenia.

Nowa próbka ma postać $\widetilde{\boldsymbol{X}} = \left(\widetilde{X}_1, ..., \widetilde{X}_n\right)$, przy czym zakłada się, że

$$\mathrm{E}\,\widetilde{X}_i = \widetilde{\mu}_i = \widetilde{\mu} \quad \text{oraz} \quad \mathrm{Var}\,\widetilde{X}_i = \widetilde{\sigma}_i^2 = \widetilde{\sigma}^2, \quad \text{jeśli} \quad i \in R^c$$

oraz

$$\mathrm{Corr}\left(\widetilde{X}_i, \widetilde{X}_j\right) = \widetilde{\rho}_{ij} = \widetilde{\rho}, \quad \text{jeśli} \quad i, j \in R^c, i \neq j,$$

$$\mathrm{Corr}\left(X_i, \widetilde{X}_j\right) = \rho_{ij} = \rho, \quad \text{jeśli} \quad i \in R, j \in R^c.$$

Zajmiemy się teraz, zgodnie z Definicją I.0.1, imputacyjnym estymatorem średniej μ.

TWIERDZENIE I.0.1. ***Imputacyjny estymator średniej*** μ ***ma postać***

$$\hat{\mu}_{Imp} = \frac{r\overline{X}_R + (n-r)\overline{\widetilde{X}}}{n}, \tag{I.0.1}$$

gdzie $\overline{X}_R = \frac{1}{r}\sum_{i\in R} X_i$ ***oraz*** $\overline{\widetilde{X}} = \frac{1}{n-r}\sum_{i\in R^c}\widetilde{X}_i$.

Jego wartość oczekiwana wynosi

$$\mathrm{E}\hat{\mu}_{Imp} = \frac{r}{n}\mu + \frac{n-r}{n}\tilde{\mu}, \tag{I.0.2}$$

a jego obciążenie wynosi

$$\mathrm{B}\hat{\mu}_{Imp} = \frac{n-r}{n}\left(\tilde{\mu} - \mu\right),$$

więc estymator $\hat{\mu}_{Imp}$ ***jest nieobciążony tylko gdy*** $\tilde{\mu} = \mu$.

Dowód. Z postaci funkcji g występującej w definicji estymatora $\hat{\mu}$ wynika, że

$$\hat{\mu}_{Imp} = g\left(\tilde{\boldsymbol{X}}\right) = \frac{1}{n}\left(\sum_{i\in R} X_i + \sum_{i\in R^c} \tilde{X}_i\right),$$

co jest równoważne z (I.0.1).

Wzór (I.0.1) implikuje

$$\mathrm{E}\hat{\mu}_{Imp} = \frac{1}{n}\left(r\mathrm{E}\overline{X}_R + (n-r)\mathrm{E}\overline{\tilde{X}}\right).$$

Wzór (I.0.2) wynika z tej równości, ponieważ $\mathrm{E}\overline{X}_R = \mu$, a $\mathrm{E}\overline{\tilde{X}} = \tilde{\mu}$.

Wzór na obciążenie jest natychmiastową konsekwencją definicji obciążenia $\mathrm{B}\hat{\mu}_{Imp} = \mathrm{E}\hat{\mu}_{Imp} - \mu$ i wzoru (I.0.2).■

TWIERDZENIE I.0.2. ***Wariancja imputacyjnego estymatora średniej*** $\hat{\mu}_{Imp}$ ***ma postać***

$$\mathrm{Var}\ \hat{\mu}_{Imp} = \frac{r\sigma^2 + (n-r)\tilde{\sigma}^2 + (n-r)(n-r-1)\tilde{\rho}\tilde{\sigma}^2 + 2r(n-r)\rho\sigma\tilde{\sigma}}{n^2} \tag{I.0.3}$$

Dowód. Z postaci estymatora $\hat{\mu}_{Imp}$ i podstawowych własności wariancji otrzymujemy

$$\mathrm{Var}\,\hat{\mu}_{Imp} = \frac{1}{n^2}\left(r^2\mathrm{Var}\,\overline{X}_R + (n-r)^2\,\mathrm{Var}\,\overline{\tilde{X}} + 2r(n-r)\mathrm{Cov}\left(\overline{X}_R, \overline{\tilde{X}}\right)\right). \tag{I.0.4}$$

Ale $\mathrm{Var}\,\overline{X}_R = \frac{\sigma^2}{r}$.

Z kolei

$$\operatorname{Var}\overline{\widetilde{X}} = \frac{1}{(n-r)^2}\left(\sum_{i\in R^c}\operatorname{Var}\widetilde{X}_i + \sum_{i,\,j\in R^c,\,i\neq j}\operatorname{Cov}\left(\widetilde{X}_i,\,\widetilde{X}_j\right)\right) =$$
$$= \frac{1}{(n-r)^2}\left((n-r)\widetilde{\sigma}^2 + (n-r)(n-r-1)\widetilde{\rho}\widetilde{\sigma}^2\right).$$

A zatem

$$\operatorname{Var}\overline{\widetilde{X}} = \frac{1+(n-r-1)\widetilde{\rho}}{n-r}\widetilde{\sigma}^2. \tag{I.0.5}$$

Natomiast

$$\operatorname{Cov}\left(\overline{X}_R,\ \overline{\widetilde{X}}\right) = \frac{1}{r(n-r)}\sum_{j\in R}\sum_{i\in R^c}\operatorname{Cov}\left(X_j,\,\widetilde{X}_i\right) = \rho\sigma\widetilde{\sigma}. \tag{I.0.6}$$

Wstawiając te trzy wyrażenia do wzoru (I.0.4) otrzymujemy

$$\operatorname{Var}\hat{\mu}_{Imp} = \frac{1}{n^2}\left(r\sigma^2 + (n-r)\left(1+(n-r-1)\widetilde{\rho}\right)\sigma^2 + 2r(n-r)\rho\sigma\widetilde{\sigma}\right)$$

Ten wzór jest równoważny wzorowi (I.0.3).■

Imputacyjna estymacja wariancji polega na wykorzystaniu standardowego estymatora wariancji do próbki imputowanej $\widetilde{\boldsymbol{X}}$, czyli zgodnie z definicją I.0.1

$$S^2_{Imp} = S^2\left(\widetilde{\boldsymbol{X}}\right) = \frac{1}{n-1}\left(\sum_{j\in R}\left(X_j - \hat{\mu}_{Imp}\right)^2 + \sum_{i\in R^c}\left(\widetilde{X}_i - \hat{\mu}_{Imp}\right)^2\right). \tag{I.0.7}$$

TWIERDZENIE I.0.3. ***Imputacyjny estymator wariancji*** S^2_{Imp} ***ma postać***

$$S^2_{Imp} = \frac{1}{n-1}\left((r-1)S^2_R + (n-r-1)\widetilde{S}^2 + \frac{r(n-r)}{n}\left(\overline{X}_R - \overline{\widetilde{X}}\right)^2\right), \tag{I.0.8}$$

gdzie $S^2_R = \frac{1}{r-1}\sum_{j\in R}\left(X_j - \overline{X}_R\right)^2$ ***oraz*** $\widetilde{S}^2 = \frac{1}{n-r-1}\sum_{i\in R^c}\left(\widetilde{X}_i - \overline{\widetilde{X}}\right)^2$.

Dowód. Ze wzoru (I.0.7) i definicji $\hat{\mu}_{Imp}$ mamy

$$S_{Imp}^2 = \frac{1}{n-1}\left(\sum_{i\in R}\left(X_i - \frac{1}{n}\left(r\overline{X}_R + (n-r)\overline{\widetilde{X}} \right) \right)^2 + \sum_{j\in R^c}\left(\widetilde{X}_j - \overline{\widetilde{X}} + \frac{r}{n}\left(\overline{\widetilde{X}} - \overline{X}_R \right) \right)^2 \right).$$

Wykonując kwadratowanie pod obiema sumami i korzystając z faktu, że $\sum_{i\in R}\left(X_i - \overline{X}_R \right) = 0$ oraz $\sum_{i\in R^c}\left(\widetilde{X}_i - \overline{\widetilde{X}} \right) = 0$ otrzymujemy

$$S_{Imp}^2 = \frac{1}{n-1}\left(\sum_{i\in R}\left(X_i - \overline{X}_R \right)^2 + r\frac{(n-r)^2}{n^2}\left(\overline{X}_R - \overline{\widetilde{X}} \right)^2 + \sum_{i\in R^c}\left(\widetilde{X}_i - \overline{\widetilde{X}} \right)^2 + (n-r)\frac{r^2}{n^2}\left(\overline{X}_R - \overline{\widetilde{X}} \right)^2 \right) =$$

$$= \frac{1}{n-1}\left((r-1)S_R^2 + (n-r-1)\widetilde{S}^2 + \left(r\frac{(n-r)^2}{n^2} + (n-r)\frac{r^2}{n^2} \right)\left(\overline{X}_R - \overline{\widetilde{X}} \right)^2 \right),$$

co kończy dowód.■

TWIERDZENIE I.0.4. ***Wartość oczekiwana imputacyjnego estymatora wariancji*** S_{Imp}^2 ***ma postać***

$$\begin{aligned} \mathrm{E}S_{Imp}^2 = {} & \frac{r}{n}\sigma^2 + \frac{n-r}{n}\widetilde{\sigma}^2 - \frac{(n-r)(n-r-1)}{n(n-1)}\widetilde{\rho}\widetilde{\sigma}^2 - \\ & + 2\frac{r(n-r)}{n(n-1)}\rho\sigma\widetilde{\sigma} + \frac{r(n-r)}{n(n-1)}\left(\mu - \widetilde{\mu} \right)^2. \end{aligned} \tag{I.0.9}$$

Dowód. Mamy $\mathrm{E}S_R^2 = \sigma^2$ oraz $\mathrm{E}\widetilde{S}^2 = \sigma^2\left(1-\widetilde{\rho} \right)$. Pierwszy fakt jest powszechnie znany, a drugi wynika z następującego rachunku

$$\mathrm{E}\,\widetilde{S}^2 = \mathrm{E}\left(\frac{1}{n-r-1}\sum_{i\in R^c}\widetilde{X}_i^2 - \frac{1}{(n-r-1)(n-r)}\left(\sum_{i\in R^c}\widetilde{X}_i \right)^2 \right) = \frac{1}{n-r-1}\sum_{i\in R^c}\mathrm{E}\widetilde{X}_i^2 -$$

$$+ \frac{1}{(n-r-1)(n-r)}\left(\sum_{i\in R^c}\mathrm{E}\widetilde{X}_i^2 + \sum_{i,j\in R^c,\, i\neq j}\mathrm{E}\widetilde{X}_i\widetilde{X}_j \right) = \frac{n-r}{n-r-1}(\widetilde{\sigma}^2 + \widetilde{\mu}^2) -$$

$$+ \frac{1}{(n-r-1)(n-r)}\left((n-r)(\widetilde{\sigma}^2 + \widetilde{\mu}^2) + (n-r)(n-r-1)(\widetilde{\rho}\widetilde{\sigma}^2 + \widetilde{\mu}^2) \right) = \widetilde{\sigma}^2(1-\widetilde{\rho}).$$

Co więcej

$$\mathrm{E}\left(\overline{X}_R - \overline{\widetilde{X}}\right)^2 = \mathrm{E}\,\overline{X}_R^2 + \mathrm{E}\,\widetilde{X}^2 - 2\,\mathrm{E}\,\overline{X}_R\overline{\widetilde{X}} = \mathrm{Var}\,\overline{X}_R + \left(\mathrm{E}\,\overline{X}_R\right)^2 +$$
$$+ \mathrm{Var}\,\overline{\widetilde{X}} + \left(\mathrm{E}\,\overline{\widetilde{X}}_R\right)^2 - 2\,\mathrm{Cov}\left(\overline{X}_R, \overline{\widetilde{X}}\right) - 2\,\mathrm{E}\,\overline{X}_R\,\mathrm{E}\,\overline{\widetilde{X}}.$$

Wykorzystując wzory (I.0.5) oraz (I.0.6) otrzymujemy

$$\mathrm{E}\left(\overline{X}_R - \overline{\widetilde{X}}\right)^2 = \frac{\sigma^2}{r} + \frac{\widetilde{\sigma}^2}{n-r} + \frac{n-r-1}{n-r}\widetilde{\rho}\widetilde{\sigma}^2 - 2\rho\sigma\widetilde{\sigma} + \left(\mu - \widetilde{\mu}\right)^2.$$

Ostatecznie ze wzoru (I.0.8) i powyżej uzyskanych tożsamości wynika, że

$$(n-1)\mathrm{E}\,S_{Imp}^2 = (r-1)\sigma^2 + (n-r-1)\widetilde{\sigma}^2(1-\widetilde{\rho}) +$$
$$+ \frac{r(n-r)}{n}\left(\frac{\sigma^2}{r} + \frac{\widetilde{\sigma}^2}{n-r} + \frac{n-r-1}{n-r}\widetilde{\rho}\widetilde{\sigma}^2 - 2\rho\sigma\widetilde{\sigma} + \left(\mu - \widetilde{\mu}\right)^2\right),$$

co po uproszczeniu daje wzór (I.0.9).■

I.1. Imputacja średnią

Technika ta polega na przypisaniu każdemu nieobserwowanemu elementowi próbki średniej z części obserwowanej próbki, czyli

$$\widetilde{X}_i = \frac{1}{r}\sum_{j\in R} X_j = \overline{X}_R, \quad i \in R^c. \tag{I.1.1}$$

TWIERDZENIE I.1.1. ***Imputacyjny estymator wartości oczekiwanej w przypadku imputacji średnią ma postać***

$$\hat{\mu}_{Imp} = \overline{X}_R. \tag{I.1.2}$$

Jego wartość oczekiwana oraz wariancja wynoszą, odpowiednio

$$\mathrm{E}\,\hat{\mu}_{Imp} = \mu \quad \textbf{\textit{oraz}} \quad \mathrm{Var}\,\hat{\mu}_{Imp} = \frac{\sigma^2}{r}. \tag{I.1.3}$$

Dowód. Ze wzoru (I.1.1) wynika, że $\overline{\widetilde{X}} = \overline{X}_R$. Zatem wzór (I.0.1) prowadzi natychmiast do (I.1.2). Ponieważ $\widetilde{\mu} = \mathrm{E}\,\widetilde{X}_i = \mathrm{E}\,\overline{X}_R = \mu$, więc (I.0.2) daje pierwszy ze wzorów (I.1.3).

Dodatkowo z (I.1.1) wynika, że $\widetilde{\sigma}^2 = \operatorname{Var}\widetilde{X}_i = \operatorname{Var}\overline{X}_R = \dfrac{\sigma^2}{r}$ oraz

$$\widetilde{\rho} = \operatorname{Corr}\left(\widetilde{X}_i, \widetilde{X}_j\right) = \operatorname{Corr}\left(\overline{X}_R, \overline{X}_R\right) = 1,$$

$$\rho = \operatorname{Corr}\left(X_i, \widetilde{X}_j\right) = \operatorname{Corr}\left(X_i, \overline{X}_R\right) = \sqrt{r}\,\frac{\operatorname{Cov}\left(X_i, \overline{X}_R\right)}{\sigma^2} = \sqrt{r}\,\frac{\dfrac{\sigma^2}{r}}{\sigma^2} = \frac{1}{\sqrt{r}}.$$

Zatem zgodnie ze wzorem (I.0.3)

$$\operatorname{Var}\hat{\mu}_{Imp} = \frac{r\sigma^2 + (n-r)\dfrac{\sigma^2}{r} + (n-r)(n-r-1)\dfrac{\sigma^2}{r} + 2r(n-r)\dfrac{1}{\sqrt{r}}\sigma\dfrac{\sigma}{\sqrt{r}}}{n^2}.$$

Po uproszczeniu otrzymujemy drugi ze wzorów (I.1.3).■

Oczywiście wzory (I.1.3) wynikają wprost z postaci (I.1.2) estymatora imputacyjnego. W dowodzie, wyprowadzając je odpowiednio ze wzorów (I.0.2) oraz (I.0.3), chcieliśmy zaznaczyć uniwersalność proponowanego podejścia.

TWIERDZENIE I.1.2. ***Imputacyjny estymator wariancji w przypadku imputacji średnią ma postać***

$$S^2_{Imp} = \frac{r-1}{n-1}S^2_R, \tag{I.1.4}$$

a jego wartość oczekiwana wynosi

$$\mathrm{E}S^2_{Imp} = \frac{r-1}{n-1}\sigma^2. \tag{I.1.5}$$

Nieobciążony estymator wariancji estymatora $\hat{\mu}_{Imp}$ ***ma postać***

$$v^2\left(\hat{\mu}_{Imp}\right) = \frac{n-1}{r(r-1)}S^2_{Imp}. \tag{I.1.6}$$

Dowód. Zgodnie ze wzorem (I.1.1) mamy $\widetilde{S}^2 = 0$ oraz $\overline{\widetilde{X}} = \overline{X}_R$. Zatem (I.1.4) wynika bezpośrednio ze wzoru (I.0.8).

Natomiast zgodnie ze wzorem (I.0.9) na wartość oczekiwaną imputacyjnego estymatora wariancji mamy

$$\mathrm{E}S_{Imp}^2 = \frac{r}{n}\sigma^2 + \frac{n-r}{n}\frac{\sigma^2}{r} - \frac{(n-r)(n-r-1)}{n(n-1)}\frac{\sigma^2}{r} - 2\frac{r(n-r)}{n(n-1)}\frac{1}{\sqrt{r}}\sigma\frac{\sigma}{\sqrt{r}},$$

skąd po uproszczeniach otrzymuje się wzór (I.1.5).

Wzór (I.1.6) wynika z porównania wzoru (I.1.5) i drugiego ze wzorów (I.1.3).■

Wzór (I.1.5) wynika również bezpośrednio z postaci (I.1.4), ponieważ S_R^2 jest nieobciążonym estymatorem wariancji σ^2. Wyprowadziliśmy go jednak ze wzoru (I.0.9), aby, podobnie jak w przypadku wzorów (I.0.2) i (I.0.3) (użytych do wyprowadzenia (I.1.3)), wskazać na jego uniwersalność.

Zauważmy, że ze wzoru (I.1.5) wynika, iż estymator S_{Imp}^2 jest obciążonym estymatorem wariancji.

I.2. Imputacja typu hot-deck

I.2.1. Losowanie spośród respondentów

Technika ta polega na przypisaniu każdemu nieobserwowanemu elementowi próbki elementu wylosowanego spośród elementów obserwowanych. W najprostszej sytuacji jest to losowanie proste ze zwracaniem.

Niech K_i będzie numerem elementu wylosowanym dla *i*-tego nierespondenta, $i \in R^c$. Wtedy zmienne losowe K_i, $i \in R^c$, są niezależne i mają ten sam rozkład $\mathrm{P}(K_i = j) = \frac{1}{r}$, $j \in R$, $i \in R^c$ oraz wektory losowe $\boldsymbol{K}=(K_i, i \in R^c)$ i $\boldsymbol{X}_R=(X_j, j \in R)$ są niezależne.

W konsekwencji

$$\widetilde{X}_i = X_{K_i}, \ i \in R^c. \tag{I.2.1}$$

Uwaga. W przypadku obserwacji dwuwymiarowych $\big((X_j, Y_j), j \in R, \ Y_i, i \in R^c\big)$ stosowana bywa metoda wyboru najbliższego sąsiada. Polega ona na ustaleniu dla każdego $i \in R^c$ takiego $K_i \in R$, dla którego odległość $|Y_{K_i} - Y_i|$ jest najmniejsza spośród $|Y_j - Y_i|$, $j \in R$. Przy założeniu, że wektory losowe $\boldsymbol{w}_i := (X_i, Y_i)$, $i = 1, \ldots, n$, są niezależne i mają jednakowe rozkłady (takie jak rozkład pary (X, Y)), zmienne losowe K_i, $i \in R^c$, mają rozkład jednostajny tak jak w opisanym wyżej modelu, natomiast wektory losowe $\boldsymbol{K}$ i $\boldsymbol{X}_R$ nie są niezależne. Zwracamy

uwagę, że założenie niezależności dotyczy par, tzn. wektory losowe $\boldsymbol{w}_i$, $i = 1, ..., n$, są łącznie niezależne, natomiast dla każdego ustalonego $i \in \{1, ..., n\}$ zmienne losowe X_i oraz Y_i (będące składowymi wektora $\boldsymbol{w}_i$) nie muszą być niezależne; najlepiej, gdy są np. dobrze skorelowane. Za takim podejściem przemawia następujące uzasadnienie: jeżeli obserwowane wartości Y_j i Y_i, gdzie $i \in R^c$, $j \in R$, są bliskie, to przy znacznej korelacji zmiennych X i Y można oczekiwać, że nieobserwowana wartość X_i jest bliska obserwowanej wartości X_j.

TWIERDZENIE I.2.1. ***Imputacyjny estymator wartości oczekiwanej w przypadku imputacji typu hot-deck polegającej na losowaniu respondentów ma postać***

$$\hat{\mu}_{Imp} = \frac{1}{n}\left(\sum_{j \in R} X_j + \sum_{i \in R^c} X_{K_i} \right). \tag{I.2.2}$$

Jego wartość oczekiwana oraz wariancja wynoszą, odpowiednio

$$\mathrm{E}\,\hat{\mu}_{Imp} = \mu \quad \textbf{\textit{oraz}} \quad \mathrm{Var}\,\hat{\mu}_{Imp} = \frac{\sigma^2}{r}\left(1 + \frac{(n-r)(r-1)}{n^2} \right). \tag{I.2.3}$$

Dowód. Zgodnie z (I.2.1) mamy $\overline{\widetilde{X}} = \dfrac{1}{n-r} \sum_{i \in R^c} X_{K_i}$, a zatem wzór (I.0.1) implikuje (I.2.2).

Zauważmy, że dla dowolnego $i \in R^c$

$$\mathrm{E}\left(X_{K_i} \middle| \boldsymbol{K} \right) = \mathrm{E}\, X = \mu \quad \text{oraz} \quad \mathrm{Var}\left(X_{K_i} \middle| \boldsymbol{K} \right) = \mathrm{Var}\left(X \right) = \sigma^2$$

i w konsekwencji

$$\hat{\mu} = \mathrm{E}\, \widetilde{X}_i = \mathrm{E}\, X_{K_i} = \mathrm{E}\, \mathrm{E}\left(X_{K_i} \middle| \boldsymbol{K} \right) = \mu \tag{I.2.4}$$

oraz

$$\widetilde{\sigma}^2 = \mathrm{Var}\, X_{K_i} = \mathrm{Var}\, \mathrm{E}\left(X_{K_i} \middle| \boldsymbol{K} \right) + \mathrm{E}\, \mathrm{Var}\left(X_{K_i} \middle| \boldsymbol{K} \right) = \sigma^2. \tag{I.2.5}$$

Ze wzoru (I.2.4) wynika, że $E\,\overline{\widetilde{X}} = \mu$, zgodnie ze wzorem (I.0.2), mamy więc pierwszy ze wzorów (I.2.3), czyli estymator $\hat{\mu}_{Imp}$ jest nieobciążonym estymatorem średniej.

Podobnie dla dowolnych $i, j \in R^c$ oraz $l \in R$ mamy

$$\mathrm{Cov}\left(X_{K_i}, X_{K_j} \middle| \boldsymbol{K}\right) = \sigma^2 \, \mathrm{I}\left(K_i = K_j\right) \quad \text{oraz} \quad \mathrm{Cov}\left(X_l, X_{K_j} \middle| \boldsymbol{K}\right) = \sigma^2 \, \mathrm{I}\left(K_j = l\right).$$

Stąd wynika, że dla dowolnych $i, j \in R^c$

$$\widetilde{\rho} = \mathrm{Corr}\left(\widetilde{X}_i, \widetilde{X}_j\right) = \frac{\mathrm{Cov}\left(X_{K_i}, X_{K_j}\right)}{\sigma^2} =$$
$$= \frac{\mathrm{E}\,\mathrm{Cov}\left(X_{K_i}, X_{K_j} \middle| \boldsymbol{K}\right) + \mathrm{Cov}\left(\mathrm{E}\left(X_{K_i} \middle| \boldsymbol{K}\right), \mathrm{E}\left(X_{K_j} \middle| \boldsymbol{K}\right)\right)}{\sigma^2}.$$

Ponieważ $\mathrm{Cov}\left(\mathrm{E}\left(X_{K_i} \middle| \boldsymbol{K}\right), \mathrm{E}\left(X_{K_j} \middle| \boldsymbol{K}\right)\right) = 0,$ z powyższego wzoru otrzymujemy

$$\widetilde{\rho} = \mathrm{P}\left(K_i = K_j\right) = \frac{1}{r}. \tag{I.2.6}$$

Z kolei dla dowolnych $j \in R$ oraz $i \in R^c$

$$\rho = \mathrm{Corr}\left(X_j, \widetilde{X}_i\right) = \frac{\mathrm{Cov}\left(X_j, X_{K_i}\right)}{\sigma^2} =$$
$$= \frac{\mathrm{E}\,\mathrm{Cov}\left(X_j, X_{K_i} \middle| \boldsymbol{K}\right) + \mathrm{Cov}\,\mathrm{E}\left(X_j \middle| \boldsymbol{K}\right), \mathrm{E}\left(X_{K_i} \middle| \boldsymbol{K}\right)}{\sigma^2}.$$

Ponieważ $\mathrm{Cov}\ \mathrm{E}\left(X_j \middle| \boldsymbol{K}\right), \mathrm{E}\left(X_{K_i} \middle| \boldsymbol{K}\right) = 0,$ z powyższego wzoru otrzymujemy

$$\rho = \mathrm{P}\left(K_i = j\right) = \frac{1}{r}. \tag{I.2.7}$$

Wykorzystując wzory (I.2.4—I.2.7), zgodnie ze wzorem (I.0.3), mamy

$$\mathrm{Var}\,\hat{\mu}_{Imp} = \frac{r\sigma^2 + (n-r)\sigma^2 + (n-r)(n-r-1)\frac{\sigma^2}{r} + 2r(n-r)\frac{\sigma^2}{r}}{n^2}.$$

Po uproszczeniach otrzymujemy drugi ze wzorów (I.2.3).■

TWIERDZENIE I.2.2. ***Imputacyjny estymator wariancji w przypadku imputacji typu hot-deck polegającej na losowaniu respondentów ma postać***

$$S^2_{Imp} = \frac{1}{n-1}\left((r-1)S^2_R + \sum_{i \in R^c}\left(X_{K_i} - \frac{1}{n-r}\sum_{j \in R^c} X_{K_j}\right)^2 + \frac{r(n-r)}{n}\left(\overline{X}_R - \frac{1}{n-r}\sum_{j \in R^c} X_{K_j}\right)^2\right),$$

a jego wartość oczekiwana wynosi

$$\mathrm{E}\, S_{Imp}^2 = \frac{r-1}{r}\frac{n(n-1)+r}{n(n-1)}\sigma^2. \tag{I.2.8}$$

Nieobciążony imputacyjny estymator wariancji estymatora $\hat{\mu}_{Imp}$ ma postać

$$v^2(\hat{\mu}_{Imp}) = \frac{n-1}{n(r-1)}\frac{n(n-1)+r+r(n-r)}{n(n-1)+r}S_{Imp}^2. \tag{I.2.9}$$

Dowód. Wzór na S_{Imp}^2 jest natychmiastową konsekwencją wzoru (I.2.1) oraz ogólnego wzoru na estymator imputacyjny wariancji (I.0.8). Z kolei wstawiając wzory (I.2.4)—(I.2.7) do wzoru (I.0.9) otrzymujemy

$$\begin{aligned}\mathrm{E}\, S_{Imp}^2 &= \frac{r}{n}\sigma^2 + \frac{n-r}{n}\sigma^2 - \frac{(n-r)(n-r-1)}{n(n-1)}\frac{\sigma^2}{r} - 2\frac{r(n-r)}{n(n-1)}\frac{\sigma^2}{r} = \\ &= \left(1 - \frac{(n-r)(n+r-1)}{n(n-1)r}\right)\sigma^2.\end{aligned}$$

Ostatnie wyrażenie jest równoważne ze wzorem (I.2.8).

Wzór (I.2.9) wynika z porównania wzoru (I.2.8) oraz drugiego ze wzorów (I.2.3).■

Zatem dla dużych wartości *r* mamy $\mathrm{E}\, S_{Imp}^2 \approx \sigma^2$, czyli estymator wariancji jest w przybliżeniu nieobciążony.

I.2.2. Losowanie z rozkładu normalnego

Zakładamy, że próbka pochodzi z rozkładu normalnego, czyli X_i, $i \in R$, są niezależnymi zmiennymi losowymi o rozkładzie normalnym $N(\mu,\sigma^2)$. Zmienne imputowane generowane są w sposób niezależny z rozkładu $N\left(\overline{X}_R, S_R^2\right)$, tzn. $\widetilde{X}_j \big| \boldsymbol{X}_R \sim N\left(\overline{X}_R, S_R^2\right)$, $j \in R^c$, oraz $\widetilde{X}_j, j \in R^c$, są warunkowo niezależne pod warunkiem $\boldsymbol{X}_R$. Zatem

$$\mathrm{E}\,\widetilde{X}_j = \mathrm{E}\mathrm{E}\left(\widetilde{X}_j \big| \boldsymbol{X}_R\right) = \mathrm{E}\,\overline{X}_R = \mu$$

oraz

$$\mathrm{Var}\,\widetilde{X}_j = \mathrm{Var}\,\mathrm{E}\left(\widetilde{X}_j \big| \boldsymbol{X}_R\right) + \mathrm{E}\,\mathrm{Var}\left(\widetilde{X}_j \big| \boldsymbol{X}_R\right) = \mathrm{Var}\,\overline{X}_R + \mathrm{E}\, S_R^2 = \sigma^2\,\frac{1+r}{r}.$$

Ponieważ

$$\mathrm{Cov}\left(X_i, \widetilde{X}_j \middle| \boldsymbol{X}_R\right) = 0 \quad \text{oraz} \quad \mathrm{Cov}\left(\widetilde{X}_i, \widetilde{X}_j \middle| \boldsymbol{X}_R\right) = 0.$$

więc

$$\mathrm{Cov}\left(X_i, \widetilde{X}_j\right) = \mathrm{Cov}\left(\mathrm{E}\left(X_i \middle| \boldsymbol{X}_R\right), \mathrm{E}\left(\widetilde{X}_j \middle| \boldsymbol{X}_R\right)\right) = \mathrm{Cov}\left(X_i, \overline{X}_R\right) = \frac{\sigma^2}{r}$$

oraz

$$\mathrm{Cov}\left(\widetilde{X}_i, \widetilde{X}_j \middle| \boldsymbol{X}_R\right) = \mathrm{Cov}\left(\mathrm{E}\left(\widetilde{X}_i \middle| \boldsymbol{X}_R\right), \mathrm{E}\left(\widetilde{X}_j \middle| \boldsymbol{X}_R\right)\right) = \mathrm{Var}\,\overline{X}_R = \frac{\sigma^2}{r}.$$

TWIERDZENIE I.2.3. ***Dla imputacyjnego estymatora wartości oczekiwanej w przypadku imputacji typu hot-deck polegającej na losowaniu z rozkładu normalnego wartość oczekiwana oraz wariancja wynoszą odpowiednio***

$$\mathrm{E}\hat{\mu}_{Imp} = \mu \quad \textbf{\textit{oraz}} \quad \mathrm{Var}\,\hat{\mu}_{Imp} = \frac{\sigma^2}{r}\left(1 + \frac{(n-r)r}{n^2}\right). \tag{I.2.10}$$

Natomiast

$$\mathrm{E}S^2_{Imp} = \left(1 - \frac{n-r}{n\,(n-1)}\right)\sigma^2. \tag{I.2.11}$$

Nieobciążony estymator wariancji estymatora $\hat{\mu}_{Imp}$ ***ma postać***

$$\hat{v}^2(\hat{\mu}_{Imp}) = \frac{n-1}{rn}\frac{n^2 + (n-r)r}{n(n-1) - (n-r)} S^2_{Imp}. \tag{I.2.12}$$

Dowód. Wzory (I.2.10) są odpowiednio natychmiastową konsekwencją wzorów (I.0.2) i (I.0.3), natomiast wzór (I.2.11) wynika wprost ze wzoru (I.0.9). Wzór (I.2.12) jest z kolei konsekwencją wzoru (I.2.11) i drugiego ze wzorów (I.2.10).■

I.3. Imputacja regresyjna

Zakładamy, że pary (X_i, Y_i), $i = 1, 2, ..., n,$ tworzące próbkę pierwotną mają jednakowe rozkłady, takie jak para (X, Y) i są niezależne. Próbka obserwowana ma postać $\left(X_i, Y_i\right), i \in R, Y_j, j \in R^c$ czyli zmienna *Y*-owa jest obserwowana

w całej próbce pierwotnej, natomiast zmienna X-owa — jedynie dla respondentów. Jeśli znany jest rozkład warunkowy $X|Y$, to można wartość zmiennej losowej $\widetilde{X}_j$ generować z rozkładu warunkowego $X_j|Y_j$, $j \in R^c$. Wtedy własności statystyczne próbki imputowanej $\widetilde{\boldsymbol{X}}$ nie różnią się od własności próbki pierwotnej $\boldsymbol{X}$. Jednak najczęściej rozkład warunkowy może być jedynie estymowany na podstawie obserwacji $(X_i, Y_i), i \in R$. W takiej sytuacji precyzyjny opis własności estymatorów imputacyjnych jest trudny, w szczególności zależy od metody estymacji rozkładu warunkowego.

Zamiast estymacji rozkładu warunkowego można wykorzystać model regresyjny, który jest konstrukcją teoretyczną znajdującą zastosowanie w takiej sytuacji.

Niech $\mathrm{E}X = \mu$, $\mathrm{E}Y = v$, $\mathrm{Var}(X) = \sigma^2$, $\mathrm{Var}(Y) = \tau^2$. Niech χ będzie współczynnikiem korelacji zmiennych X i Y. Niech λ będzie współczynnikiem regresji X względem Y, tzn. $\lambda = \chi \dfrac{\sigma}{\tau}$.

W najprostszej sytuacji, gdy znane są średnia v oraz współczynnik regresji λ, technika imputacji regresyjnej polega na przypisaniu nierespondentom wartości predyktora regresyjnego, tzn.

$$\widetilde{X}_j = \lambda(Y_j - v) + \overline{X}_R,\ j \in R^c.$$

TWIERDZENIE I.3.1. ***Imputacyjny estymator wartości oczekiwanej w przypadku imputacji regresyjnej ma postać***

$$\hat{\mu}_{Imp} = \overline{X}_R + \frac{n-r}{n}\lambda\left(\overline{Y}_{R^c} - v\right), \tag{I.3.1}$$

gdzie $\overline{Y}_{R^c} = \dfrac{1}{n-r}\sum_{j \in R^c} Y_j$.

Jego wartość oczekiwana oraz wariancja wynoszą odpowiednio

$$\mathrm{E}\,\hat{\mu}_{Imp} = \mu \quad \textbf{\textit{oraz}} \quad \mathrm{Var}\,\hat{\mu}_{Imp} = \sigma^2\left(\frac{1}{r} + \chi^2\,\frac{n-r}{n^2}\right). \tag{I.3.2}$$

Dowód. Ponieważ

$$\overline{\widetilde{X}} = \frac{1}{n-r}\sum_{j \in R^c}\widetilde{X}_j = \overline{X}_R + \lambda\left(\overline{Y}_{R^c} - v\right), \tag{I.3.3}$$

więc wzór (I.3.1) wynika wprost ze wzoru (I.0.1).

Z kolei

$$\mathrm{E}\left(\overline{Y}_{R^c} - v\right) = 0.$$

a zatem pierwszy ze wzorów (I.3.2) wynika natychmiast z (I.3.1).

Ponieważ Y_j oraz $\overline{X}_R$ są niezależne dla $j \in R^c$, to

$$\widetilde{\sigma}^2 = \mathrm{Var}\,\widetilde{X}_j = \mathrm{Var}\left(\lambda\left(Y_j - v\right)\right) + \mathrm{Var}\left(\overline{X}_R\right) = \sigma^2\left(\chi^2 + \frac{1}{r}\right).$$

Dla $i \in R, j \in R^c$

$$\mathrm{Cov}\left(X_i, \widetilde{X}_j\right) = \mathrm{Cov}\left(X_i, \overline{X}_R\right) = \frac{\sigma^2}{r},$$

więc

$$\rho = \frac{\frac{\sigma^2}{r}}{\sigma^2\sqrt{\chi^2 + \frac{1}{r}}} = \frac{1}{\sqrt{r(1 + r\chi^2)}}.$$

Dla $i, j \in R^c$

$$\mathrm{Cov}\left(\widetilde{X}_i, \widetilde{X}_j\right) = \mathrm{Var}\left(\overline{X}_R\right) = \frac{\sigma^2}{r},$$

więc

$$\widetilde{\varrho} = \frac{\frac{\sigma^2}{r}}{\sigma^2\left(\chi^2 + \frac{1}{r}\right)} = \frac{1}{1 + r\chi^2}$$

Ponieważ $\widetilde{\varrho}\,\widetilde{\sigma}^2 = \frac{\sigma^2}{r}$ oraz $\rho\sigma\widetilde{\sigma} = \frac{\sigma^2}{r}$, więc wstawiając otrzymane wyżej wyrażenia na $\widetilde{\sigma}^2$, ρ oraz $\widetilde{\varrho}$ do wzoru (I.0.3) mamy

$$\mathrm{Var}\hat{\mu}_{Imp} = \frac{r\sigma^2 + (n-r)\sigma^2\left(\chi^2 + \frac{1}{r}\right) + (n-r)(n-r-1)\frac{\sigma^2}{r} + 2r(n-r)\frac{\sigma^2}{r}}{n^2} =$$

$$= \frac{\sigma^2}{n^2}\left(\frac{n^2}{r} + (n-r)\chi^2\right),$$

co jest równoważne z drugim ze wzorów (I.3.2).■

Uwaga. W odróżnieniu od wcześniej rozważanych przypadków w imputacji regresyjnej wartości imputowane zależą od współczynnika regresji λ oraz od wartości oczekiwanej v. Najczęściej nie są one znane. Wtedy we wzorach na zmienne imputowane, na estymator imputacyjny średniej oraz na estymator imputacyjny wariancji (poniżej) zamiast λ oraz v należy wstawić ich estymatory otrzymane na podstawie obserwowanej części próbki (X_i, Y_i), $i \in R$, Y_j, $j \in R^c$.

TWIERDZENIE I.3.2. ***Imputacyjny estymator wariancji w przypadku imputacji regresyjnej ma postać***

$$S^2_{Imp} = \frac{1}{n-1}\left((r-1)S^2_R + \lambda^2\left((n-r-1)S^2_{R^c,\,Y} + \frac{(n-r)r}{n}\left(\overline{Y}_{R^c} - v\right)^2 \right)\right), \qquad \text{(I.3.4)}$$

a jego wartość oczekiwana wynosi

$$\mathrm{E}\, S^2_{Imp} = \sigma^2\left(\frac{r-1}{n-1} + \left(1 - \frac{r}{n}\right)\chi^2\right). \qquad \text{(I.3.5)}$$

Nieobciążony imputacyjny estymator wariancji estymatora $\hat{\mu}_{Imp}$ ma postać

$$\hat{v}^2(\hat{\mu}_{Imp}) = \frac{\frac{1}{r} + \frac{n-r}{n^2}\chi^2}{\frac{r-1}{n-1} + \frac{n-r}{n}\chi^2} S^2_{Imp}. \qquad \text{(I.3.6)}$$

Dowód. Ze wzoru (I.3.1) i definicji $\widetilde{S}^2$ mamy

$$\widetilde{S}^2 = \lambda^2 S^2_{R^c,\,Y}, \text{ gdzie } S^2_{R^c,\,Y} = \frac{1}{n-r-1}\sum_{j \in R^c}\left(Y_j - \overline{Y}_{R^c}\right)^2.$$

Ponownie wykorzystując wzór (I.3.1) otrzymujemy

$$\left(\overline{X}_R - \overline{\widetilde{X}}\right)^2 = \lambda^2\left(\overline{Y}_{R^c} - v\right)^2.$$

Wstawiając oba powyższe wyrażenia na $\widetilde{S}^2$ oraz na $\left(\overline{X}_R - \overline{\widetilde{X}}\right)^2$ do wzoru (I.0.8) otrzymujemy (I.3.4).

Natomiast wstawiając wyrażenia

$$\widetilde{\sigma}^2 = \sigma^2\left(\chi^2 + \frac{1}{r}\right),\ \widetilde{\varrho}\,\widetilde{\sigma}^2 = \frac{\sigma^2}{r} \text{ oraz } \rho\sigma\widetilde{\sigma} = \frac{\sigma^2}{r}$$

do wzoru (I.0.9) na wartość oczekiwaną S^2_{Imp}, po łatwych uproszczeniach, otrzymujemy (I.3.5).

Wzór (I.3.6) wynika z porównania wzoru (I.3.5) oraz drugiego ze wzorów (I.3.2).■

II. DETERMINISTYCZNY ZBIÓR BRAKÓW OBSERWACJI — IMPUTACJA WIELOKROTNA

II.0. Ogólny schemat imputacji wielokrotnej

Imputacja wielokrotna została wprowadzona przez Rubina w monografii z 1987 r. Poniżej przytaczamy podstawowe zasady tej metody, zwane zasadami Rubina (*Rubin's rules*), przedstawione np. na stronach 39 i 40 monografii Carpenter i Kenward (2013).

Imputacja wielokrotna polega na generowaniu kilku próbek imputacyjnych

$$\widetilde{\boldsymbol{X}}^{(l)} = \left(X_i, i \in R, \widetilde{X}^{(l)}_j,\ j \in R^c\right),\ l = 1, ..., m,$$

na podstawie obserwowanej części próbki $\boldsymbol{X}_R = \left(X_i, i \in R\right)$.

Definicja II.0.1. ***Estymatorem wielokrotnej imputacji Rubina parametru*** θ ***(odpowiadającym estymatorowi*** $\hat{\theta} = g(\boldsymbol{X})$***) nazywamy średnią z estymatorów imputacyjnych dla próbek*** $\widetilde{\boldsymbol{X}}^{(l)},\ l = 1, ..., m,$ ***czyli***

$$\hat{\theta}_{MImp} = \frac{1}{m}\sum_{l=1}^{m}\hat{\theta}^{(l)}_{Imp}.$$

Estymator Rubina wariancji estymatora $\hat{\theta}_{MImp}$ ***ma postać***

$$\hat{v}^2_{MImp} = \overline{U}_m + \frac{m+1}{m}B_m,$$

gdzie

$$\overline{U}_m := \frac{1}{m}\sum_{l=1}^{m}\hat{v}^{(l),2}_{Imp}$$

oznacza średnią estymatorów wariancji estymatorów imputacyjnych, natomiast

$$B_m := \frac{1}{m-1}\sum_{l=1}^{m}\left(\hat{\theta}_{Imp}^{(l)} - \hat{\theta}_{MImp}\right)^2 \tag{II.0.1}$$

jest wariancją empiryczną estymatorów imputacyjnych.

Poniżej interesować się będziemy estymacją średniej μ w modelu opisanym w poprzednim rozdziale. Zakładać będziemy, że $\widetilde{\boldsymbol{X}}^{(l)}$, $l = 1, ..., m$, są warunkowo niezależne pod warunkiem $\boldsymbol{X}_R$. Oczywiście jest to równoważne temu, że imputowane części próbek $\left(\widetilde{X}_j^{(l)}, j \in R^c\right)$, $l = 1, ..., m$, są warunkowo niezależne pod warunkiem $\boldsymbol{X}_R$. Zakłada się również, że $\mathrm{E}\,\widetilde{X}_j^{(l)} = \widetilde{\mu}, \mathrm{Var}\,\widetilde{X}_j^{(l)} = \widetilde{\sigma}^2$ nie zależą od $j \in R^c$ oraz od $l = 1, ..., m$, $\mathrm{Corr}\left(\widetilde{X}_i^{(l)}, \widetilde{X}_j^{(l)}\right) = \widetilde{\rho}$ nie zależy od i, $j \in R^c$ oraz od $l = 1, ..., m$, a $\mathrm{Corr}\left(X_i, \widetilde{X}_j^{(l)}\right) = \rho$ nie zależy od $i \in R$, $j \in R^c$ oraz $l = 1, ..., m$.

Dodatkowo zakłada się warunek liniowości regresji

$$\mathrm{E}\left(\widetilde{X}_j^{(l)} \middle| \boldsymbol{X}_R\right) = \alpha\,\overline{X}_R + \beta$$

dla dowolnych $j \in R^c$ oraz $l = 1, ..., m$. Współczynniki α i β można wyliczyć korzystając z zależności

$$\widetilde{\mu} = \mathrm{E}\mathrm{E}\left(\widetilde{X}_j^{(l)} \middle| \boldsymbol{X}_R\right) = \alpha\,\mu + \beta$$

oraz

$$\widetilde{\mu}\mu + \rho\sigma\widetilde{\sigma} = \mathrm{E}\,\widetilde{X}_j^{(l)} X_i = \mathrm{E}\,X_i \mathrm{E}\left(\widetilde{X}_j^{(l)} \middle| \boldsymbol{X}_R\right) = \mathrm{E}\,X_i\left(\alpha\,\overline{X}_R + \beta\right) = \alpha\left(\frac{\sigma^2}{r} + \mu^2\right) + \beta\mu.$$

Rozwiązując powyższy układ równań względem α i β otrzymujemy

$$\mathrm{E}\left(\widetilde{X}_j^{(l)} \middle| \boldsymbol{X}_R\right) = r\rho\frac{\widetilde{\sigma}}{\sigma}\left(\overline{X}_R - \mu\right) + \widetilde{\mu} \tag{II.0.2}$$

dla dowolnych $j \in R^c$ oraz $l = 1, \ldots, m$.

Na podstawie próbek imputacyjnych konstruowane są estymatory

$$\hat{\mu}_{Imp}^{(l)} = \frac{1}{n}\left[r\,\overline{X}_R + (n-r)\,\overline{\widetilde{X}}^{(l)}\right],\ l = 1, \ldots, m,$$

a estymator finalny jest ich średnią

$$\hat{\mu}_{MImp} = \frac{1}{m}\sum_{l=1}^{m}\hat{\mu}_{Imp}^{(l)} = \frac{r}{n}\overline{X}_R + \frac{n-r}{nm}\sum_{l=1}^{m}\overline{\widetilde{X}}^{(l)}. \tag{II.0.3}$$

TWIERDZENIE II.0.1. ***Wartość oczekiwana estymatora wielokrotnej imputacji dla średniej wynosi***

$$\mathrm{E}\hat{\mu}_{MImp} = \frac{r}{n}\mu + \frac{n-r}{n}\widetilde{\mu}. \tag{II.0.4}$$

Jego wariancja ma postać

$$\mathrm{Var}(\hat{\mu}_{MImp}) = \\ = \frac{mr\sigma^2 + (n-r)\left(1+(n-r-1)\widetilde{\rho}\right)\widetilde{\sigma}^2 + 2mr(n-r)\,\varrho\sigma\widetilde{\sigma} + (m-1)(n-r)^2 r\varrho^2\widetilde{\sigma}^2}{mn^2}. \tag{II.0.5}$$

Dowód. Wzór (II.0.4) wynika wprost z faktu, że dla każdego z estymatorów $\hat{\mu}_{Imp}^{(l)}$, $l = 1, \ldots, m$, obowiązuje wzór (I.0.2).

Znajdziemy teraz wariancję estymatora wielokrotnej imputacji średniej

$$\mathrm{Var}\,\hat{\mu}_{MImp} = \frac{1}{m^2}\left(\sum_{l=1}^{m}\mathrm{Var}\,\hat{\mu}_{Imp}^{(l)} + \sum_{k\neq l}^{m}\mathrm{Cov}\left(\hat{\mu}_{Imp}^{(k)},\ \hat{\mu}_{Imp}^{(l)}\right)\right).$$

Ale zgodnie ze wzorem (I.0.3)

$$\mathrm{Var}\,\hat{\mu}_{Imp}^{(l)} = \frac{1}{n^2}\left(r\sigma^2 + (n-r)\left(1+(n-r-1)\widetilde{\rho}\right)\widetilde{\sigma}^2 + 2r(n-r)\,\varrho\,\sigma\widetilde{\sigma}\right).$$

Z kolei

$$\mathrm{Cov}\left(\hat{\mu}_{Imp}^{(k)},\ \hat{\mu}_{Imp}^{(l)}\right) = \frac{1}{n^2}\left(r^2\,\mathrm{Var}\left(\overline{X}_R\right) + (n-r)^2\,\mathrm{Cov}\left(\overline{\widetilde{X}}^{(k)}, \overline{\widetilde{X}}^{(l)}\right) + \right. \\ \left. + r(n-r)\left(\mathrm{Cov}\left(\overline{X}_R, \overline{\widetilde{X}}^{(k)}\right) + \mathrm{Cov}\left(\overline{X}_R, \overline{\widetilde{X}}^{(l)}\right)\right)\right).$$

Wykorzystując wzór (I.0.6) otrzymujemy

$$\mathrm{Cov}\left(\hat{\mu}_{Imp}^{(k)},\ \hat{\mu}_{Imp}^{(l)}\right) = \frac{1}{n^2}\left(r\sigma^2 + (n-r)^2\,\mathrm{Cov}\left(\overline{\widetilde{X}}^{(k)},\ \overline{\widetilde{X}}^{(l)}\right) + 2r(n-r)\,\varrho\,\sigma\widetilde{\sigma}\right).$$

Ale

$$\mathrm{Cov}\left(\overline{\widetilde{X}}^{(k)},\, \overline{\widetilde{X}}^{(l)}\right) = \mathrm{E}\mathrm{Cov}\left(\overline{\widetilde{X}}^{(k)},\, \overline{\widetilde{X}}^{(l)}\middle|\boldsymbol{X}_R\right) + \mathrm{Cov}\left(\mathrm{E}\left(\overline{\widetilde{X}}^{(k)}\middle|\boldsymbol{X}_R\right),\, \mathrm{E}\left(\overline{\widetilde{X}}^{(l)}\middle|\boldsymbol{X}_R\right)\right).$$

Z warunkowej niezależności wynika, że pierwszy składnik po prawej stronie jest równy zeru, natomiast

$$\mathrm{Cov}\left(\mathrm{E}\left(\overline{\widetilde{X}}^{(k)}\middle|\boldsymbol{X}_R\right),\, \mathrm{E}\left(\overline{\widetilde{X}}^{(l)}\middle|\boldsymbol{X}_R\right)\right) = \frac{1}{(n-r)^2}\sum_{i,\,j\in R^c}\mathrm{Cov}\left(\mathrm{E}\left(\widetilde{X}_i^{(k)}\middle|\boldsymbol{X}_R\right),\, \mathrm{E}\left(\widetilde{X}_j^{(l)}\middle|\boldsymbol{X}_R\right)\right).$$

Z liniowości regresji (II.0.2) otrzymujemy dla dowolnych $i, j \in R^c$

$$\mathrm{Cov}\left(\mathrm{E}\left(\widetilde{X}_i^{(k)}\middle|\boldsymbol{X}_R\right),\, \mathrm{E}\left(\widetilde{X}_j^{(l)}\middle|\boldsymbol{X}_R\right)\right) = r^2\varrho^2\frac{\widetilde{\sigma}^2}{\sigma^2}\mathrm{Var}\left(\overline{X}_R\right) = r\varrho^2\widetilde{\sigma}^2.$$

Zatem

$$\mathrm{Cov}\left(\hat{\mu}_{Imp}^{(k)}, \hat{\mu}_{Imp}^{(l)}\right) = \frac{1}{n^2}\left(r\sigma^2 + (n-r)^2\, r\varrho^2\,\widetilde{\sigma}^2 + 2r(n-r)\,\varrho\,\sigma\widetilde{\sigma}\right). \tag{II.0.6}$$

Ostatecznie

$$\begin{aligned}\mathrm{Var}\,\hat{\mu}_{MImp} &= \frac{1}{mn^2}\left(r\sigma^2 + (n-r)\left(1+(n-r-1)\widetilde{\rho}\right)\widetilde{\sigma}^2 + 2r(n-r)\;\varrho + \sigma\widetilde{\sigma}\right)\\ &\quad + \frac{m-1}{mn^2}\left(r\sigma^2 + (n-r)^2\, r\varrho^2\,\widetilde{\sigma}^2 + 2r(n-r)\,\varrho\,\sigma\widetilde{\sigma}\right),\end{aligned}$$

co po łatwych uproszczeniach pozwala otrzymać wzór (II.0.5).■

Zgodnie z drugą częścią Definicji II.0.1 estymatorem Rubina wariancji estymatora $\hat{\mu}_{MImp}$ nazywamy statystykę określoną wzorem

$$\hat{v}_{MImp}^2 = \overline{U}_m + \frac{m+1}{m}B_m, \tag{II.0.7}$$

gdzie

$$\overline{U}_m := \frac{1}{mn}\sum_{l=1}^{m}\left(S_{Imp}^{(l)}\right)^2, \tag{II.0.8}$$

przy czym

$$\left(S_{Imp}^{(l)}\right)^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\widetilde{X}_i^{(l)} - \hat{\mu}_{Imp}^{(l)}\right)^2,\ l = 1, ..., m,$$

natomiast

$$B_m := \frac{1}{m-1}\sum_{l=1}^{m}\left(\hat{\mu}_{Imp}^{(l)} - \hat{\mu}_{MImp}\right)^2. \tag{II.0.9}$$

Estymator (II.0.7), zwany estymatorem Rubina, na ogół nie jest nieobciążony.

TWIERDZENIE II.0.2. ***Wartość oczekiwana estymatora Rubina wynosi***

$$\mathrm{E}\,\hat{v}_{MImp}^2 = \frac{r\sigma^2 + (n-r)\widetilde{\sigma}^2 - \dfrac{(n-r)\left[(n-r-1)\widetilde{\rho}\widetilde{\sigma}^2 + 2r\rho\sigma\widetilde{\sigma} - r(\mu-\widetilde{\mu})^2\right]}{n-1}}{n^2} + \\ + \frac{m+1}{m}\frac{(n-r)\left[1 + (n-r-1)\widetilde{\varrho} - (n-r)r\varrho^2\right]}{n^2}\widetilde{\sigma}^2. \tag{II.0.10}$$

Dowód. Zauważmy najpierw, że

$$\mathrm{E}\overline{U}_m = \frac{1}{mn}\sum_{l=1}^{m}\mathrm{E}\left(S_{Imp}^{(l)}\right)^2 = \frac{1}{n}\mathrm{E}\left(S_{Imp}^{(1)}\right)^2.$$

A zatem zgodnie ze wzorem (I.0.9) otrzymujemy

$$\mathrm{E}\overline{U}_m = \frac{1}{n^2}\left(r\sigma^2 + (n-r)\widetilde{\sigma}^2 - \frac{(n-r)\left[(n-r-1)\widetilde{\rho}\widetilde{\sigma}^2 + 2r\rho\sigma\widetilde{\sigma} - r(\mu-\widetilde{\mu})^2\right]}{n-1}\right), \tag{II.0.11}$$

czyli pierwszy składnik we wzorze (II.0.10).

Z kolei, jeśli $\mathrm{E}\hat{\mu}_{Imp}^{(l)} = \mathrm{E}\hat{\mu}_{MImp} = v,\ l = 1, ..., m,$ to

$$\sum_{l=1}^{m}\mathrm{E}\left(\hat{\mu}_{Imp}^{(l)} - \hat{\mu}_{MImp}\right)^2 = \sum_{l=1}^{m}\mathrm{E}\left(\hat{\mu}_{Imp}^{(l)} - v\right)^2 - 2\mathrm{E}\left(\hat{\mu}_{MImp} - v\right)\sum_{l=1}^{m}\left(\hat{\mu}_{Imp}^{(l)} - v\right) + \\ + m\mathrm{E}\left(\hat{\mu}_{MImp} - v\right)^2 = m\left(\mathrm{Var}\,\hat{\mu}_{Imp}^{(1)} - \mathrm{Var}\,\hat{\mu}_{MImp}\right).$$

Ale

$$\operatorname{Var}\hat{\mu}_{MImp} = \frac{1}{m^2}\left(\sum_{l=1}^{m}\operatorname{Var}\hat{\mu}_{Imp}^{(l)} + \sum_{l,k=1,l\neq k}^{m}\operatorname{Cov}\left(\hat{\mu}_{Imp}^{(l)},\,\hat{\mu}_{Imp}^{(k)}\right)\right) =$$
$$= \frac{\operatorname{Var}\hat{\mu}_{Imp}^{(1)} + (m-1)\operatorname{Cov}\left(\hat{\mu}_{Imp}^{(1)},\hat{\mu}_{Imp}^{(2)}\right)}{m}.$$

W konsekwencji

$$\sum_{l=1}^{m}\mathrm{E}\left(\hat{\mu}_{Imp}^{(l)} - \hat{\mu}_{MImp}\right)^2 = m\left(\operatorname{Var}\hat{\mu}_{Imp}^{(1)} - \frac{\operatorname{Var}\hat{\mu}_{Imp}^{(1)} + (m-1)\operatorname{Cov}\left(\hat{\mu}_{Imp}^{(1)},\,\hat{\mu}_{Imp}^{(2)}\right)}{m}\right) =$$
$$= (m-1)\left(\operatorname{Var}\hat{\mu}_{Imp}^{(1)} + \operatorname{Cov}\left(\hat{\mu}_{Imp}^{(1)},\,\hat{\mu}_{Imp}^{(2)}\right)\right).$$

Zatem

$$\mathrm{E}B_m = \frac{1}{m-1}\sum_{l=1}^{m}\mathrm{E}\left(\hat{\mu}_{Imp}^{(l)} - \hat{\mu}_{MImp}\right)^2 = \operatorname{Var}\hat{\mu}_{Imp}^{(1)} + \operatorname{Cov}\left(\hat{\mu}_{Imp}^{(1)},\hat{\mu}_{Imp}^{(2)}\right).$$

Stosując w powyższej równości wzory (I.0.3) i (II.0.6) otrzymujemy

$$\mathrm{E}B_m = \frac{(n-r)\left[1 + (n-r-1)\widetilde{\varrho} - (n-r)r\varrho^2\right]}{n^2}\widetilde{\sigma}^2, \tag{II.0.12}$$

czyli drugą część wzoru (II.0.10), bez mnożnika $\frac{m+1}{m}$. ■

II.1. Imputacja wielokrotna typu hot-deck

II.1.1. Losowanie spośród respondentów

Każdą z próbek imputacyjnych $\widetilde{\boldsymbol{X}}^{(l)}$, $l = 1, \ldots, m$, tworzymy stosując metodę hot-deck opisaną w podrozdziale I.2. Niech $\boldsymbol{K}^{(l)} = \left(K_j^{(l)},\, j \in R^c\right)$ będzie wektorem numerów elementów wylosowanych dla kolejnych nierespondentów przy tworzeniu próbki $\widetilde{\boldsymbol{X}}^{(l)}$. Zakładamy, że wektory losowe $\boldsymbol{K}^{(l)}$, $l = 1, \ldots, m$, oraz $\boldsymbol{X}_R$ są niezależne. Wtedy próbki imputacyjne $\widetilde{\boldsymbol{X}}^{(l)}$, $l = 1, \ldots, m$, są warunkowo niezależne pod warunkiem $\boldsymbol{X}_R$.

Zgodnie z teorią rozwiniętą dla jednokrotnej imputacji hot-deck mamy

$$\widetilde{\mu} = \mu, \quad \widetilde{\sigma}^2 = \sigma^2, \quad \widetilde{\rho} = \rho = \frac{1}{r}. \tag{II.1.1}$$

Co więcej, z niezależności $\boldsymbol{K}^{(l)}$ i $\boldsymbol{X}_R$ wynika, że dla $j \in R^c$

$$\mathrm{E}\left(\widetilde{X}_j^{(l)} \middle| \boldsymbol{X}_R\right) = \mathrm{E}\left(X_{K_j^{(l)}} \middle| \boldsymbol{X}_R\right) = \overline{X}_R,$$

czyli $\alpha = 1$ oraz $\beta = 0$ w ogólnym wzorze na liniową regresję z podrozdziału II.0.

TWIERDZENIE II.1.1. ***Estymator wielokrotnej imputacji ma postać***

$$\hat{\mu}_{MImp} = \frac{r}{n}\overline{X}_R + \frac{1}{nm}\sum_{l=1}^{m}\sum_{j \in R^c} X_{K_j^{(l)}}.$$

Jego wartość oczekiwana wynosi

$$\mathrm{E}\hat{\mu}_{MImp} = \mu,$$

a jego wariancja ma postać

$$\mathrm{Var}\,\hat{\mu}_{MImp} = \frac{\sigma^2}{r}\left(1 + \frac{(n-r)(r-1)}{mn^2}\right). \tag{II.1.2}$$

Dowód. Wzór na postać estymatora wielokrotnej imputacji wynika wprost z ogólnej postaci (II.0.3) oraz wzoru (I.2.2) z twierdzenia I.2.1, natomiast wzór na jego wartość oczekiwaną jest konsekwencją wzoru (II.0.4) oraz równości $\widetilde{\mu} = \mu$.

Zgodnie ze wzorem (II.0.5) mamy

$$\mathrm{Var}\,\hat{\mu}_{MImp} = \frac{mr\sigma^2 + (n-r)\left(1 + \frac{n-r-1}{r}\right)\sigma^2 + 2m(n-r)\sigma^2 + (m-1)\frac{(n-r)^2}{r}\sigma^2}{mn^2} =$$

$$= \frac{\sigma^2}{n^2}\left(r + 2(n-r) + \frac{n^2}{r} - 2n + r\right) + \frac{\sigma^2}{mn^2}\left(n - r + \frac{(n-r)^2}{r} - \frac{n-r}{r} - \frac{(n-r)^2}{r}\right) =$$

$$= \frac{\sigma^2}{r} + \frac{\sigma^2(n-r)(r-1)}{mn^2 r},$$

co daje wzór (II.1.2).■

TWIERDZENIE II.1.2. ***Obciążenie estymatora Rubina wariancji*** $\hat{v}^2_{MImp}$ ***wynosi***

$$\mathrm{B}\hat{v}^2_{MImp} = \mathrm{E}\hat{v}^2_{MImp} - \mathrm{Var}\,\hat{\mu}_{MImp} = -\frac{(n-r)\left[n(n-r+1)+2(r-1)\right]}{n^2(n-1)r}\sigma^2. \quad \text{(II.1.3)}$$

Estymator nieobciążony wariancji estymatora $\hat{\mu}_{MImp}$ ***ma postać***

$$\hat{v}^2_* = \frac{n}{r}\overline{U}_m + \left(\frac{1}{m} + \frac{n(n+r-1)}{(n-1)r(r-1)}\right)B_m.$$

Dowód. Wstawiając do (II.0.11) wartości podane w (II.1.1) otrzymujemy

$$\mathrm{E}\overline{U}_m = \frac{\sigma^2}{n^2}\left(n - \frac{(n-r)(n+r-1)}{r(n-1)}\right).$$

Z kolei wstawiając do (II.0.12) te same wartości z (II.1.1) otrzymujemy

$$\mathrm{E}B_m = \sigma^2\,\frac{(n-1)(r-1)}{rn^2}. \quad \text{(II.1.4)}$$

Zatem

$$\mathrm{E}\hat{v}^2_{MImp} = \mathrm{E}\overline{U}_m + \mathrm{E}B_m + \frac{1}{m}\mathrm{E}B_m = \frac{\sigma^2(r-1)\left(2n(n-1)+2r-nr\right)}{rn^2(n-1)} + \frac{\sigma^2(r-1)(n-r)}{mn^2r}.$$

Obciążenie wynosi więc

$$\mathrm{E}\hat{v}^2_{MImp} - \mathrm{Var}\,\hat{\mu}_{MImp} = \frac{\sigma^2(r-1)\left(2n(n-1)+2r-nr\right)}{rn^2(n-1)} - \frac{\sigma^2}{r},$$

co daje po uproszczeniach wzór (II.1.3).

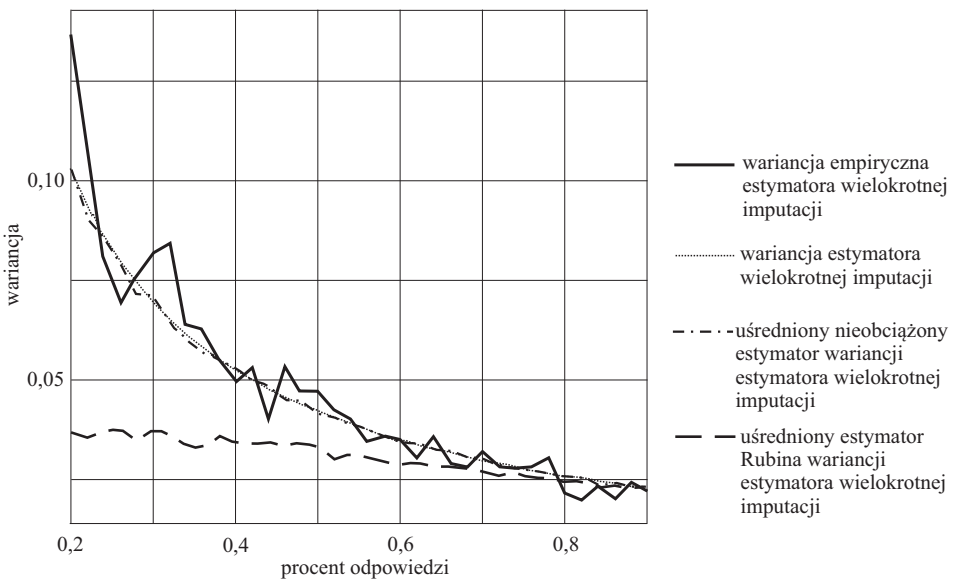
Zauważmy, że

$$\frac{n}{r}\mathrm{E}\overline{U}_m + \frac{n(n+r-1)}{(n-1)r(r-1)}\mathrm{E}B_m = \frac{\sigma^2}{r}.$$

Nieobciążoność estymatora $\hat{v}^2_*$ wynika z powyższego wzoru oraz (II.1.4) i wzoru (II.1.2) na wariancję estymatora $\hat{\mu}_{MImp}$. ■

Symulacyjna ilustracja wyników zawartych w twierdzeniach II.1.1 i II.1.2 przedstawiona jest na wykr. 1. Szczegółowy opis procedury symulacyjnej znajduje się w aneksie.

Wykr. 1. WYNIKI ESTYMACJI WARIANCJI ESTYMATORA ŚREDNIEJ DLA WIELOKROTNEJ IMPUTACJI (HOT-DECK)



Ź r ó d ł o: obliczenia własne (zobacz aneks).

Wniosek II.1.3. ***Asymptotycznie względne obciążenie estymatora Rubina*** $\hat{v}^2_{MImp}$ ***wynosi***

$$\lim_{n\to\infty} \frac{\mathrm{E}\hat{v}^2_{MImp} - \mathrm{Var}\,\hat{\mu}_{MImp}}{\mathrm{Var}\,\hat{\mu}_{MImp}} = -1. \tag{II.1.5}$$

Dowód. Ze wzoru (II.1.2) mamy

$$\lim_{n\to\infty} \mathrm{Var}\,\hat{\mu}_{MImp} = \frac{\sigma^2}{r}.$$

Z kolei wzór (II.1.3) implikuje

$$\lim_{n\to\infty} \left(\mathrm{E}\hat{v}^2_{MImp} - \mathrm{Var}\,\hat{\mu}_{MImp}\right) = -\frac{\sigma^2}{r}. \blacksquare$$

Uwaga. Oczywiście istnieją też inne niż $\hat{v}_*^2$ nieobciążone estymatory wariancji estymatora wielokrotnej imputacji $\hat{\mu}_{MImp}$. W szczególności estymatory mające postać kombinacji liniowej statystyk $\overline{U}_m$ i B_m. Analiza tej klasy estymatorów jest równoważna rozważaniu estymatorów wariancji $\hat{\mu}_{MImp}$ postaci

$$\hat{v}_{(\alpha,\beta)}^2 = \alpha\overline{U}_m + \beta B_m + \frac{1}{m}B_m,$$

przy warunku

$$\beta = \frac{1}{n-r}\left(\frac{n^2}{r-1} - \frac{n(n-1)+r}{(n-r)(n-1)}\alpha\right) \tag{II.1.6}$$

zapewniającym nieobciążoność. Istotnym zagadnieniem jest znalezienie optymalnych wartości α, β, tzn. takich, przy których wariancja estymatora $\hat{v}_{(\alpha,\beta)}^2$ osiąga minimum. Rozwiązanie analityczne jest trudne, wymaga znajomości trzecich i czwartych momentów. Możliwe jest natomiast numeryczne badanie wariancji estymatora $\hat{v}_{(\alpha,\beta)}^2$ jako funkcji zmiennej α, przy ograniczeniu (II.1.6).

II.1.2. Losowanie z rozkładu normalnego

Korzystamy ze wzorów ogólnych, przy czym

$$\widetilde{\mu} = \mu, \quad \widetilde{\sigma}^2 = \sigma^2\frac{1+r}{r}, \quad \varrho = \frac{1}{\sqrt{r(r+1)}}, \quad \widetilde{\varrho} = \frac{1}{r+1}.$$

TWIERDZENIE II.1.4. ***Wartość oczekiwana estymatora imputacji wielokrotnej przy losowaniu z rozkładu normalnego wynosi***

$$\mathrm{E}\hat{\mu}_{MImp} = \mu,$$

a jego wariancja ma postać

$$\mathrm{Var}\hat{\mu}_{MImp} = \frac{\sigma^2}{r}\left(1 + \frac{(n-r)r}{mn^2}\right). \tag{II.1.7}$$

Dowód. Powyższe wzory wynikają wprost ze wzorów (II.0.4) i (II.0.5) z twierdzenia II.1.1. ■

TWIERDZENIE II.1.5. ***Obciążenie estymatora Rubina wariancji*** $\hat{v}^2_{MImp}$ ***wynosi***

$$\mathrm{B}\hat{v}^2_{MImp} = \mathrm{E}\hat{v}^2_{MImp} - \mathrm{Var}\,\hat{\mu}_{MImp} = -\frac{(n-r)\big[(n-1)(n-r)-1\big]}{n^2(n-1)r}\sigma^2. \qquad \text{(II.1.8)}$$

Estymator nieobciążony wariancji estymatora $\hat{\mu}_{MImp}$ ***ma postać***

$$\hat{v}^2_* = \frac{n}{r}\overline{U}_m + \left(\frac{1}{m} + \frac{n}{(n-1)r}\right)B_m.$$

Dowód. Ze wzorów (II.0.11) oraz (II.0.12) wynika, że

$$\mathrm{E}\overline{U}_m = \frac{\sigma^2}{n^2}\left(n - \frac{n-r}{n-1}\right) \quad \text{oraz} \quad \mathrm{E}B_m = \frac{\sigma^2(n-r)}{n^2}.$$

Zatem

$$\mathrm{E}\hat{v}^2_{MImp} - \mathrm{Var}\,\hat{\mu}_{MImp} = \frac{\sigma^2}{n^2}\left(n + \frac{(n-r)(r-2)}{r(n-1)}\right) + \frac{m+1}{m}\frac{\sigma^2(n-r)}{n^2} - \frac{\sigma^2}{r}\left(1 + \frac{(n-r)r}{mn^2}\right)$$

skąd po prostych przekształceniach otrzymujemy wzór (II.1.8).

Ponieważ

$$\mathrm{Var}\,\hat{\mu}_{MImp} - E\frac{1}{m}B_m = \frac{\sigma^2}{r},$$

więc wystarczy znaleźć α i β takie, że $\alpha\mathrm{E}\overline{U}_m + \beta\mathrm{E}B_m = \frac{\sigma^2}{r}$. Przyjmujemy $\alpha = \frac{n}{r}$.

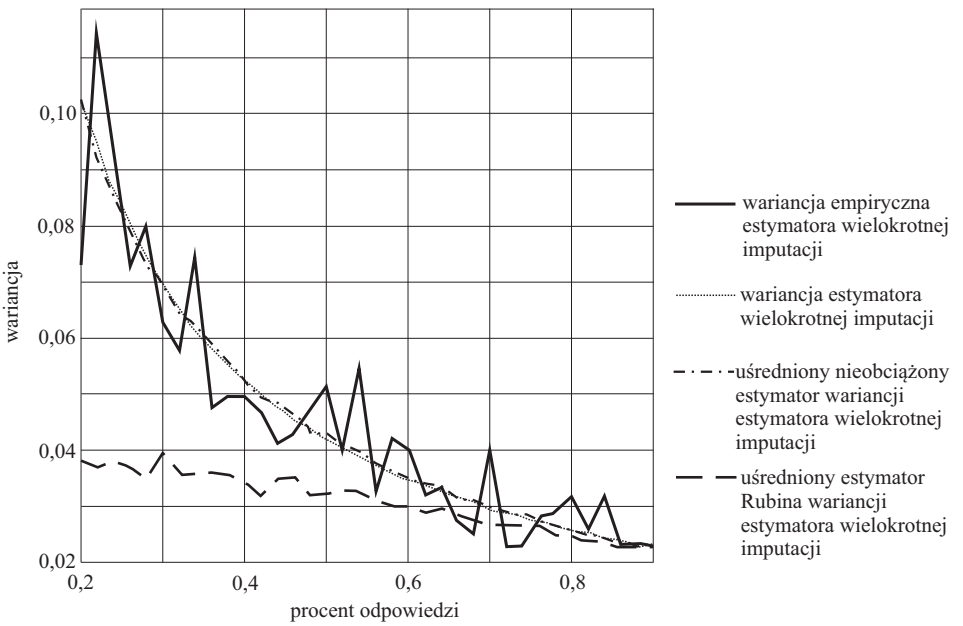
Wtedy nietrudno wyliczyć, że $\beta = -\frac{n(r-2)}{(n-1)r^2}$. ■

Podobnie jak we wniosku II.1.3 i w tym przypadku asymptotyczne obciążenie wynosi

$$\lim_{n\to\infty}\frac{\mathrm{E}\hat{v}^2_{MImp} - \mathrm{Var}\,\hat{\mu}_{MImp}}{\mathrm{Var}\,\hat{\mu}_{MImp}} = -1.$$

Symulacyjna ilustracja wyników zawartych w twierdzeniach II.1.4 i II.1.5 przedstawiona jest na wykr. 2. Szczegółowy opis procedury symulacyjnej znajduje się w aneksie.

Wykr. 2. WYNIKI ESTYMACJI WARIANCJI ESTYMATORA ŚREDNIEJ DLA WIELOKROTNEJ IMPUTACJI (HOT-DECK) Z ROZKŁADU NORMALNEGO



Ź r ó d ł o: jak przy wykr. 1.

III. LOSOWY ZBIÓR BRAKÓW OBSERWACJI — IMPUTACJA JEDNOKROTNA

III.0. Ogólny schemat imputacyjny

Niech wektor losowy $\boldsymbol{X} = (X_1, \ldots, X_n)$ oznacza pierwotną próbkę, natomiast $\boldsymbol{J} = (J_1, \ldots, J_n)$ niech oznacza wektor losowy, w którym składowa $J_i = 1$, jeśli zmienna X_i jest obserwowana oraz $J_i = 0$, jeśli zmienna X_i nie jest obserwowana. W konsekwencji zbiór $\boldsymbol{R} = \{i \in \{1, ..., n\} : J_i = 1\}$ jednostek, dla których obserwacja jest dostępna jest losowy.

Przy założeniu, że mechanizm pojawiania się braków odpowiedzi nie jest zależny od wartości badanych zmiennych (MCAR, *missing completely at random*), czyli że wektory losowe $\boldsymbol{X}$ i $\boldsymbol{J}$ są (statystycznie) niezależne, do analizy własności rozkładu warunkowego $\boldsymbol{X}$ pod warunkiem $\boldsymbol{J}$ można wygodnie stosować metodologię rozwiniętą w rozdziałach I i II, czyli w sytuacji, gdy zbiór $\boldsymbol{R}$ jest deterministyczny. Ostateczne wyniki otrzymuje się poprzez uśrednianie

rezultatów dla rozkładu warunkowego względem rozkładu wektora ***J***. Szczegółowo jest to opisane dla imputacji jednokrotnej w bieżącym rozdziale i dla imputacji wielokrotnej w rozdziale kolejnym.

Konstrukcja wektora ***J*** wymaga przeanalizowania pewnej subtelności technicznej. Losowość braków obserwacji modelowana jest pierwotnie przez zmienne losowe (indykatory odpowiedzi) $\delta_1, ..., \delta_n$, przy czym $\delta_i = 1$, jeśli obserwacja X_i jest dostępna, a $\delta_i = 0$, jeśli obserwacja X_i nie jest dostępna (brak odpowiedzi). Zakładamy, że zmienne $\delta_1, ..., \delta_n$ są niezależne oraz że wektory $\boldsymbol{X}$ i $\boldsymbol{\delta} = (\delta_1, ..., \delta_n)$ są również niezależne. Wprowadzamy oznaczenie p_i na prawdopodobieństwo *i*-tej odpowiedzi, $\mathrm{P}(\delta_i = 1) = p_i$, dla każdego $i = 1, \ldots, n$. Niech $|\boldsymbol{\delta}| = \sum_{i=1}^{n} \delta_i$ oznacza liczbę elementów obserwowanych w próbce. Jeśli w danej próbce nie ma żadnej dostępnej obserwacji, tzn. $|\boldsymbol{\delta}| = 0$, to badamy kolejną próbkę i postępujemy tak do momentu, gdy znajdziemy próbkę, dla której $|\boldsymbol{\delta}| > 0$. Dopiero na zakończenie takiej procedury wprowadzamy wektor ostatecznych braków odpowiedzi ***J*** — jeśli obserwacja X_i jest dostępna, przyjmujemy $J_i = 1$, a jeśli obserwacja X_i nie jest dostępna, przyjmujemy $J_i = 0$, $i = 1, \ldots, n$. Czyli zamiast oryginalnego wektora $\boldsymbol{\delta}$ o składowych niezależnych, braki odpowiedzi opisane są za pomocą wektora $\boldsymbol{J} = (J_1, \ldots, J_n)$, którego składowe nie są niezależne. Tym niemniej wektory losowe ***X*** i ***J*** są niezależne. Zauważmy, że

$$\mathrm{P}(J_1 = \varepsilon_1, ..., J_n = \varepsilon_n) = \mathrm{P}(\delta_1 = \varepsilon_1, ..., \delta_n = \varepsilon_n \,\|\, |\boldsymbol{\delta}| > 0) = \frac{\prod_{i=1}^{n} p_i^{\varepsilon_i} (1 - p_i)^{1-\varepsilon_i}}{1 - \prod_{i=1}^{n} (1 - p_i)},$$

dla $\varepsilon_i \in \{0,1\}$, $i = 1, \ldots, n$, takich, że $\sum_{i=1}^{n} \varepsilon_i > 0$. Wtedy $|\boldsymbol{J}| = |\boldsymbol{\delta}| > 0$ oraz

$$\mathrm{P}(|\boldsymbol{J}| = k) = \frac{\sum_{1 \le i_1, ..., i_k \le n} \prod_{l=1}^{k} p_{i_l} \prod_{j \notin \{i_1, ..., i_k\}} (1 - p_j)}{1 - \prod_{i=1}^{n} (1 - p_i)}, \quad k = 1, ..., n.$$

W szczególnym przypadku, gdy prawdopodobieństwo odpowiedzi jest stałe, tzn. $p_i = p$, $i = 1, \ldots, n$,

$$\mathrm{P}(J_1 = \varepsilon_1, ..., J_n = \varepsilon_n) = \frac{p^{\sum_{i=1}^{n} \varepsilon_i} (1 - p)^{n - \sum_{i=1}^{n} \varepsilon_i}}{1 - (1 - p)^n}.$$

Wtedy zmienna losowa $\left|\boldsymbol{\delta}\right|$ ma rozkład dwumianowy, $b(n, p)$, tzn.

$$\mathrm{P}\left(\left|\boldsymbol{\delta}\right| = k\right) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, ..., n,$$

a zmienna losowa $\left|\boldsymbol{J}\right|$ ma rozkład dwumianowy ucięty w zerze, $b_+(n, p)$, tzn.

$$\mathrm{P}\left(\left|\boldsymbol{J}\right| = k\right) = \binom{n}{k} \frac{p^k (1-p)^{n-k}}{q_n}, \quad k = 1, ..., n,$$

gdzie

$$q_n = 1 - (1-p)^n.$$

Gdy prawdopodobieństwo braków odpowiedzi jest jednakowe i wynosi p, wtedy

$$\mathrm{E}\, g\left(\left|\boldsymbol{J}\right|\right) = \frac{1}{q_n} \sum_{i=1}^{n} g(i) \binom{n}{i} p^i (1-p)^{n-i},$$

w szczególności

$$\mathrm{E}\left|\boldsymbol{J}\right| = \frac{np}{q_n} \quad \text{oraz} \quad \mathrm{E}\frac{1}{\left|\boldsymbol{J}\right|} = \frac{1}{q_n} \sum_{i=1}^{n} \frac{1}{i} \binom{n}{i} p^i (1-p)^{n-i}.$$

Uwaga. W pracy Szabłowski, Wesołowski i Wieczorkowski (1996) pokazano, że $\mathrm{E}\frac{1}{\left|\boldsymbol{J}\right|} \cong \frac{1}{np}$. Dokładniejsze przybliżenie można znaleźć np. w pracach Marciniak i Wesołowski (1999) albo Rempała (2004).

Zmienne imputacyjne oznaczamy symbolem $\widetilde{X}_i$, $i = 1, \ldots, n$. Estymator imputacyjny średniej ma postać

$$\hat{\mu}_{Imp} = \frac{1}{n} \sum_{i=1}^{n} \left(X_i J_i + \widetilde{X}_i (1 - J_i)\right) = \frac{1}{n} \left(\sum_{j \in R} X_i + \sum_{j \in R^c} \widetilde{X}_i \right). \tag{III.0.1}$$

Natomiast imputacyjny estymator wariancji ma postać

$$S^2_{Imp} = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i J_i + \widetilde{X}_i (1 - J_i) - \hat{\mu}_{Imp}\right)^2. \tag{III.0.2}$$

III.1. Imputacja średnią

Technika ta polega na przypisaniu każdemu nieobserwowanemu elementowi próbki średniej z obserwowanej części próbki, czyli

$$\widetilde{X}_i = \frac{1}{|\boldsymbol{J}|}\sum_{j=1}^{n} X_j J_j =: \overline{X}_{\boldsymbol{J}}. \tag{III.1.1}$$

TWIERDZENIE III.1.1. ***Imputacyjny estymator wartości oczekiwanej w przypadku imputacji średnią ma postać***

$$\hat{\mu}_{Imp} = \overline{X}_{\boldsymbol{J}}. \tag{III.1.2}$$

Jego wartość oczekiwana oraz wariancja wynoszą odpowiednio

$$\mathrm{E}\hat{\mu}_{Imp} = \mu \quad \textbf{\textit{oraz}} \quad \mathrm{Var}\,\hat{\mu}_{Imp} = \sigma^2 \mathrm{E}\frac{1}{|\boldsymbol{J}|}. \tag{III.1.3}$$

Dowód. Ze wzorów (III.0.1) i (III.0.2) wynika, że

$$\hat{\mu}_{Imp} = \frac{1}{n}\sum_{i=1}^{n}\left(J_i X_i + (1-J_i)\overline{X}_{\boldsymbol{J}}\right) = \overline{X}_{\boldsymbol{J}} + \frac{1}{n}\sum_{i=1}^{n} J_i X_i - \frac{1}{n}\overline{X}_{\boldsymbol{J}}\sum_{i=1}^{n} J_i,$$

co prowadzi bezpośrednio do wzoru (III.1.2).

Z pierwszego ze wzorów (I.1.3) mamy

$$\mathrm{E}\left(\hat{\mu}_{Imp}\middle|\boldsymbol{J}\right) = \mu,$$

a zatem

$$\mathrm{E}\hat{\mu}_{Imp} = \mathrm{E}\mathrm{E}\left(\hat{\mu}_{Imp}\middle|\boldsymbol{J}\right) = \mu.$$

Z kolei

$$\mathrm{Var}\,\hat{\mu}_{Imp} = \mathrm{Var}\mathrm{E}\left(\hat{\mu}_{Imp}\middle|\boldsymbol{J}\right) + \mathrm{E}\mathrm{Var}\left(\hat{\mu}_{Imp}\middle|\boldsymbol{J}\right) = \mathrm{E}\mathrm{Var}\left(\hat{\mu}_{Imp}\middle|\boldsymbol{J}\right),$$

ponieważ $\mathrm{E}\left(\hat{\mu}_{Imp}\middle|\boldsymbol{J}\right)$ nie jest losowa. Z drugiego ze wzorów (I.1.3) mamy

$$\mathrm{Var}\left(\hat{\mu}_{Imp}\middle|\boldsymbol{J}\right) = \frac{\sigma^2}{|\boldsymbol{J}|}.$$

Czyli zachodzi drugi ze wzorów (III.1.3).■

TWIERDZENIE III.1.2. ***Imputacyjny estymator wariancji w przypadku imputacji średnią ma postać***

$$S_{Imp}^2 = \frac{1}{n-1}\sum_{i=1}^{n} J_i\left(X_i - \overline{X}_J\right)^2, \tag{III.1.4}$$

a jego wartość oczekiwana wynosi

$$\mathrm{E}S_{Imp}^2 = \frac{\mathrm{E}|\boldsymbol{J}|-1}{n-1}\sigma^2. \tag{III.1.5}$$

Nieobciążony imputacyjny estymator wariancji estymatora $\hat{\mu}_{Imp}$ ***ma postać***

$$v^2\left(\hat{\mu}_{Imp}\right) = \frac{(n-1)\mathrm{E}\dfrac{1}{|\boldsymbol{J}|}}{\mathrm{E}|\boldsymbol{J}|-1} S_{Imp}^2. \tag{III.1.6}$$

W przypadku $p_i = p$, $i = 1, \ldots, n$

$$\mathrm{E}S_{Imp}^2 = \frac{np-q_n}{q_n(n-1)}\sigma^2 \quad \textbf{\textit{oraz}} \quad v^2\left(\hat{\mu}_{Imp}\right) = \frac{q_n(n-1)\mathrm{E}\dfrac{1}{|\boldsymbol{J}|}}{np-q_n} S_{Imp}^2. \tag{III.1.7}$$

Dowód. Zgodnie ze wzorami (III.0.2) i (III.1.2) mamy

$$S_{Imp}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i J_i + \overline{X}_J(1-J_i) - \overline{X}_J\right)^2 = \frac{1}{n-1}\sum_{i=1}^{n} J_i\left(X_i - \overline{X}_J\right)^2.$$

Zgodnie ze wzorem (I.1.5)

$$\mathrm{E}\left(S_{Imp}^2 \middle| \boldsymbol{J}\right) = \frac{|\boldsymbol{J}|-1}{n-1}\sigma^2.$$

Zatem równość $\mathrm{E}S_{Imp}^2 = \mathrm{E}\mathrm{E}\left(S_{Imp}^2 \middle| \boldsymbol{J}\right)$ pociąga wzór (III.1.5).■

Zauważmy, że ze wzoru (III.1.5) wynika, iż estymator S_{Imp}^2 jest obciążonym estymatorem wariancji.

III.2. Imputacja typu hot-deck

III.2.1. Losowanie spośród respondentów

Niech K_i będzie numerem elementu wylosowanym dla i-tej jednostki, gdy $i \notin R$. Wtedy zmienne losowe K_i $i \notin R$, są warunkowo niezależne pod warunkiem $\boldsymbol{J}$ oraz $\mathrm{P}\left(K_i = j \middle| \boldsymbol{J}\right) = \frac{1}{|\boldsymbol{J}|}$, $j \in R$. Co więcej, wektory losowe $\boldsymbol{K}_{R^c} = (K_i, i \in R^c)$ i $\boldsymbol{X}_R = \left(X_j, j \in R\right)$ są warunkowo niezależne pod warunkiem $\boldsymbol{J}$.

W konsekwencji

$$\widetilde{X}_i = X_{K_i}, \quad i \in \{1, \ldots, n\}. \tag{III.2.1}$$

TWIERDZENIE III.2.1. ***Imputacyjny estymator wartości oczekiwanej w przypadku imputacji typu hot-deck polegającej na losowaniu respondentów ma postać***

$$\hat{\mu}_{Imp} = \frac{1}{n}\left(\sum_{j \in R} X_j + \sum_{i \in R^c} X_{K_i}\right). \tag{III.2.2}$$

Jego wartość oczekiwana oraz wariancja wynoszą odpowiednio

$$\mathrm{E}\hat{\mu}_{Imp} = \mu \ \textbf{\textit{oraz}} \ \mathrm{Var}\hat{\mu}_{Imp} = \frac{\sigma^2}{n}\left(1 + \frac{1}{n} + (n-1)\mathrm{E}\frac{1}{|\boldsymbol{J}|} - \frac{\mathrm{E}|\boldsymbol{J}|}{n}\right). \tag{III.2.3}$$

W przypadku $p_i = p$, $i = 1, \ldots, n$

$$\mathrm{Var}\hat{\mu}_{Imp} = \frac{\sigma^2}{n}\left(1 + \frac{1}{n} + (n-1)\mathrm{E}\frac{1}{|\boldsymbol{J}|} - \frac{p}{q_n}\right). \tag{III.2.4}$$

Dowód. Z twierdzenia I.2.1 wynika, że $\mathrm{E}\left(\hat{\mu}_{Imp} \middle| \boldsymbol{J}\right) = \mu$. W konsekwencji, warunkując względem $\boldsymbol{J}$ otrzymujemy pierwszy ze wzorów (III.2.3).

Ze wzoru (I.2.3) otrzymujemy

$$\mathrm{Var}\left(\hat{\mu}_{Imp} \middle| \boldsymbol{J}\right) = \frac{\sigma^2}{n}\left(1 + \frac{1}{n} + \frac{n-1}{|\boldsymbol{J}|} - \frac{|\boldsymbol{J}|}{n}\right).$$

Ponieważ warunkowa wartość oczekiwana $\mathrm{E}\left(\hat{\mu}_{Imp}\middle|\boldsymbol{J}\right)=\mu$ jest nielosowa, więc

$$\mathrm{Var}\,\hat{\mu}_{Imp}=\mathrm{E}\mathrm{Var}\left(\hat{\mu}_{Imp}\middle|\boldsymbol{J}\right),$$

skąd wynika drugi ze wzorów (III.2.3). Wzór (III.2.4) jest natychmiastową konsekwencją (III.2.3) oraz wzoru na wartość oczekiwaną w rozkładzie $b_+(n, p)$.■

TWIERDZENIE III.2.2. ***Imputacyjny estymator wariancji w przypadku imputacji typu hot-deck polegającej na losowaniu respondentów ma postać***

$$S^2_{Imp}=\frac{\left(\left|\boldsymbol{J}\right|-1\right)S^2_R+\left(n-\left|\boldsymbol{J}\right|-1\right)S^2_{R^c}+\frac{\left|\boldsymbol{J}\right|\left(n-\left|\boldsymbol{J}\right|\right)}{n}\left(\overline{X}_R-\overline{X}_{R^c}\right)^2}{n-1},$$

gdzie $S^2_{R^c}=\dfrac{1}{n-\left|\boldsymbol{J}\right|-1}\sum\limits_{i\in R^c}\left(X_{K_i}-\overline{X}_{R^c}\right)^2$ ***oraz*** $\overline{X}_{R^c}=\dfrac{1}{n-\left|\boldsymbol{J}\right|}\sum\limits_{j\in R^c}X_{K_j}$.

Jego wartość oczekiwana wynosi

$$\mathrm{E}S^2_{Imp}=\left(1+\frac{\mathrm{E}\left|\boldsymbol{J}\right|-1}{n(n-1)}-\mathrm{E}\frac{1}{\left|\boldsymbol{J}\right|}\right)\sigma^2. \tag{III.2.5}$$

Nieobciążony imputacyjny estymator wariancji estymatora $\hat{\mu}_{Imp}$ ***ma postać***

$$v^2\left(\hat{\mu}_{Imp}\right)=\frac{n+1+n(n-1)\mathrm{E}\dfrac{1}{\left|\boldsymbol{J}\right|}-\mathrm{E}\left|\boldsymbol{J}\right|}{n(n-1)-1+\mathrm{E}\left|\boldsymbol{J}\right|-n(n-1)\mathrm{E}\dfrac{1}{\left|\boldsymbol{J}\right|}}\frac{n-1}{n}S^2_{Imp}. \tag{III.2.6}$$

W przypadku gdy $p_i = p$, $i = 1, \ldots, n$

$$\mathrm{E}S^2_{Imp}=\left(1-\frac{1}{n(n-1)}+\frac{p}{q_n(n-1)}-\mathrm{E}\frac{1}{\left|\boldsymbol{J}\right|}\right)\sigma^2$$

oraz $$v^2\left(\hat{\mu}_{Imp}\right)=\frac{n+1+n(n-1)\mathrm{E}\dfrac{1}{\left|\boldsymbol{J}\right|}-\dfrac{np}{q_n}}{n(n-1)-1+\dfrac{np}{q_n}-n(n-1)\mathrm{E}\dfrac{1}{\left|\boldsymbol{J}\right|}}\frac{n-1}{n}S^2_{Imp}.$$

Dowód. Postać estymatora S^2_{Imp} wynika wprost z postaci podanej w twierdzeniu I.2.2, natomiast ze wzoru (I.2.8) mamy

$$\mathrm{E}\left(S^2_{Imp}\middle|\boldsymbol{J}\right)=\left(1-\frac{1}{n(n-1)}+\frac{|\boldsymbol{J}|}{n(n-1)}-\frac{1}{|\boldsymbol{J}|}\right)\sigma^2.$$

Więc (III.2.5) wynika z równości $\mathrm{E}S^2_{Imp}=\mathrm{E}\mathrm{E}\left(S^2_{Imp}\middle|\boldsymbol{J}\right)$. Łącząc wzory (III.2.4) i (III.2.5) otrzymujemy postać (III.2.6) nieobciążonego estymatora wariancji estymatora $\hat{\mu}_{Imp}$. ■

III.2.2. Losowanie z rozkładu normalnego

Stosując wykorzystywaną już wcześniej technikę warunkowania przez obserwowaną część próbki $\boldsymbol{X}_R$ bezpośrednio z twierdzenia I.2.3 otrzymujemy pełny opis standardowych własności imputacyjnego estymatora w tym przypadku.

TWIERDZENIE III.2.3. ***Dla imputacyjnego estymatora wartości oczekiwanej w przypadku imputacji hot-deck polegającej na losowaniu z rozkładu normalnego wartość oczekiwana oraz wariancja wynoszą odpowiednio***

$$\mathrm{E}\,\hat{\mu}_{Imp}=\mu \quad \textbf{\textit{oraz}} \quad \mathrm{Var}\,\hat{\mu}_{Imp}=\sigma^2\left(\mathrm{E}\frac{1}{|\boldsymbol{J}|}+\frac{n-\mathrm{E}|\boldsymbol{J}|}{n^2}\right). \tag{III.2.7}$$

Natomiast

$$\mathrm{E}S^2_{Imp}=\left(1-\frac{n-\mathrm{E}|\boldsymbol{J}|}{n(n-1)}\right)\sigma^2. \tag{III.2.8}$$

Nieobciążony estymator wariancji estymatora $\hat{\mu}_{Imp}$ ma postać

$$v^2\left(\hat{\mu}_{Imp}\right)=\frac{n-1}{n}\frac{n^2\mathrm{E}\frac{1}{|\boldsymbol{J}|}+n-\mathrm{E}|\boldsymbol{J}|}{n(n-1)-n+\mathrm{E}|\boldsymbol{J}|}S^2_{Imp}. \tag{III.2.9}$$

W przypadku gdy $p_i = p$, $i = 1, \ldots, n$

$$\mathrm{Var}\,\hat{\mu}_{Imp}=\sigma^2\left(\mathrm{E}\frac{1}{|\boldsymbol{J}|}+\frac{q_n-p}{n}\right) \quad \textbf{\textit{oraz}} \quad \mathrm{E}\,S^2_{Imp}=\left(1-\frac{q_n-p}{(n-1)q_n}\right)\sigma^2.$$

Ponadto nieobciążony estymator wariancji estymatora $\hat{\mu}_{Imp}$ ***ma postać***

$$v^2\left(\hat{\mu}_{Imp}\right)=\frac{n-1}{n}\frac{\left(n\mathrm{E}\frac{1}{|\boldsymbol{J}|}+1\right)q_n-p}{(n-2)q_n+p}S^2_{Imp}.$$

Dowód. Wzór (III.2.7) jest bezpośrednią konsekwencją wzoru (I.2.10) oraz tego, że $\mathrm{E}\left(\hat{\mu}_{Imp}|\boldsymbol{J}\right)=\mu,$ natomiast wzór (III.2.8) wynika wprost ze wzoru (I.2.11) przez warunkowanie względem $\boldsymbol{J}$. Wzór (III.2.9) na nieobciążony estymator wariancji powstaje z kombinacji wzorów (III.2.7) i (III.2.8).■

III.3. Imputacja regresyjna

W modelu z podrozdziału I.3 zakładamy, że odpowiedzi i braki odpowiedzi opisane są wektorem $\boldsymbol{J}$.

TWIERDZENIE III.3.1. ***Imputacyjny estymator wartości oczekiwanej w przypadku imputacji regresyjnej ma postać***

$$\hat{\mu}_{Imp}=\overline{X}_R+\frac{n-|\boldsymbol{J}|}{n}\lambda\left(\overline{Y}_{R^c}-v\right), \tag{III.3.1}$$

gdzie $\overline{Y}_{R^c}=\frac{1}{n-|\boldsymbol{J}|}\sum_{i\in R^c}Y_i.$

Jego wartość oczekiwana oraz wariancja wynoszą odpowiednio

$$\mathrm{E}\,\hat{\mu}_{Imp}=\mu \;\; \textbf{\textit{oraz}} \;\; \mathrm{Var}\,\hat{\mu}_{Imp}=\left(\mathrm{E}\frac{1}{|\boldsymbol{J}|}+\chi^2\frac{n-\mathrm{E}|\boldsymbol{J}|}{n^2}\right)\sigma^2. \tag{III.3.2}$$

W przypadku $p_i = p$, $i = 1, \ldots, n$

$$\mathrm{Var}\,\hat{\mu}_{Imp}=\left(\mathrm{E}\frac{1}{|\boldsymbol{J}|}+\chi^2\frac{q_n-p}{nq_n}\right)\sigma^2.$$

Dowód. Wzór (III.3.1) wynika wprost ze wzoru (I.3.1).

Z pierwszego ze wzorów (I.3.2) mamy $\mathrm{E}\left(\hat{\mu}_{Imp}\middle|\boldsymbol{J}\right)=\mu$, więc pierwszy ze wzorów (III.3.2) jest konsekwencją identyczności $\mathrm{E}\hat{\mu}_{Imp}=\mathrm{EE}\left(\hat{\mu}_{Imp}\middle|\boldsymbol{J}\right)$. Ponieważ $\mathrm{VarE}\left(\hat{\mu}_{Imp}\middle|\boldsymbol{J}\right)=0$, więc

$$\mathrm{Var}\,\hat{\mu}_{Imp}=\mathrm{EVar}\left(\hat{\mu}_{Imp}\middle|\boldsymbol{J}\right),$$

a z drugiego ze wzorów (I.3.2) mamy

$$\mathrm{Var}\left(\hat{\mu}_{Imp}\middle|\boldsymbol{J}\right)=\sigma^2\left(\frac{1}{|\boldsymbol{J}|}+\chi^2\frac{n-|\boldsymbol{J}|}{n^2}\right).$$

Stąd wynika drugi ze wzorów (III.3.2).■

TWIERDZENIE III.3.2. ***Imputacyjny estymator wariancji w przypadku imputacji regresyjnej ma postać***

$$S^2_{Imp}=\frac{\left(|\boldsymbol{J}|-1\right)S^2_R+\lambda^2\left(\left(n-|\boldsymbol{J}|-1\right)S^2_{R^c,Y}+\frac{\left(n-|\boldsymbol{J}|\right)|\boldsymbol{J}|}{n}\left(\overline{Y}_{R^c}-v\right)^2\right)}{n-1},$$

gdzie $S^2_{R^c,Y}=\frac{1}{n-|\boldsymbol{J}|-1}\sum_{i\in R^c}\left(Y_i-\overline{Y}_{R^c}\right)^2$.

Jego wartość oczekiwana wynosi

$$\mathrm{E}S^2_{Imp}=\left(\chi^2-\frac{1}{n-1}+\mathrm{E}|\boldsymbol{J}|\left(\frac{1}{n-1}-\frac{\chi^2}{n}\right)\right)\sigma^2. \qquad \text{(III.3.3)}$$

Nieobciążony imputacyjny estymator wariancji estymatora $\hat{\mu}_{Imp}$ ***ma postać***

$$v^2\left(\hat{\mu}_{Imp}\right)=\frac{\mathrm{E}\frac{1}{|\boldsymbol{J}|}+\frac{n-\mathrm{E}|\boldsymbol{J}|}{n^2}\chi^2}{\chi^2-\frac{1}{n-1}+\mathrm{E}|\boldsymbol{J}|\left(\frac{1}{n-1}-\frac{\chi^2}{n}\right)}S^2_{Imp}. \qquad \text{(III.3.4)}$$

W przypadku $p_i = p$, $i = 1, \ldots, n$

$$\mathrm{E}S_{Imp}^2 = \left(\chi^2 - \frac{1}{n-1} + \frac{np}{q_n}\left(\frac{1}{n-1} - \frac{\chi^2}{n} \right) \right)\sigma^2$$

oraz

$$v^2\left(\hat{\mu}_{Imp}\right) = \frac{\mathrm{E}\frac{1}{|\boldsymbol{J}|} + \frac{q_n - p}{nq_n}\chi^2}{\chi^2 - \frac{1}{n-1} + \frac{np}{q_n}\left(\frac{1}{n-1} - \frac{\chi^2}{n} \right)} S_{Imp}^2 .$$

Dowód. Zgodnie ze wzorem (I.3.4) mamy

$$S_{Imp}^2 = \frac{\left(|\boldsymbol{J}|-1\right)S_R^2 + \lambda^2\left(\left(n-|\boldsymbol{J}|-1\right)S_{R^c,Y}^2 + \frac{\left(n-|\boldsymbol{J}|\right)|\boldsymbol{J}|}{n}\left(\overline{Y}_{R^c} - v\right)^2 \right)}{n-1}.$$

Wzór (I.3.5) implikuje

$$\mathrm{E}\left(S_{Imp}^2 \middle| \boldsymbol{J}\right) = \sigma^2\left(\frac{|\boldsymbol{J}|-1}{n-1} + \left(1 - \frac{|\boldsymbol{J}|}{n}\right)\chi^2 \right),$$

co prowadzi natychmiast do (III.3.3). Wzór (III.3.4) jest natychmiastową konsekwencją wzorów (III.3.2) i (III.3.3).■

IV. LOSOWY ZBIÓR BRAKÓW OBSERWACJI — IMPUTACJA WIELOKROTNA

IV.1. Losowanie spośród respondentów

Zgodnie ze wzorem (II.0.3) estymator imputacji wielokrotnej jest średnią z estymatorów imputacyjnych

$$\hat{\mu}_{MImp} = \frac{1}{m}\sum_{l=1}^{m} \hat{\mu}_{Imp}^{(l)},$$

czyli zgodnie z twierdzeniem II.1.1

$$\hat{\mu}_{MImp} = \frac{|\boldsymbol{J}|}{n}\overline{X}_R + \frac{1}{nm}\sum_{l=1}^{m}\sum_{j\in R^c} X_{K_j^{(l)}}.$$

TWIERDZENIE IV.1.1. ***Wartość oczekiwana estymatora wielokrotnej imputacji wynosi***

$$\mathrm{E}\,\hat{\mu}_{MImp} = \mu,$$

a wariancja ma postać

$$\mathrm{Var}\hat{\mu}_{MImp} = \frac{\sigma^2}{mn^2}\left(n+1-\mathrm{E}|\boldsymbol{J}|-n\mathrm{E}\frac{1}{|\boldsymbol{J}|}\right)+\sigma^2\mathrm{E}\frac{1}{|\boldsymbol{J}|}. \qquad \text{(IV.1.1)}$$

W przypadku $p_i = p$, $i = 1, \ldots, n$

$$\mathrm{Var}\,\hat{\mu}_{MImp} = \frac{\sigma^2}{mn^2}\left(n+1-\frac{np}{q_n}-n\mathrm{E}\frac{1}{|\boldsymbol{J}|}\right)+\sigma^2\mathrm{E}\frac{1}{|\boldsymbol{J}|}.$$

Dowód. Ze wzoru na wartość oczekiwaną estymatora imputacyjnego hot-deck wynika, że

$$\mathrm{E}\left(\hat{\mu}_{MImp}\middle|\boldsymbol{J}\right)=\mu.$$

Więc wzór na wartość oczekiwaną estymatora $\hat{\mu}_{Imp}$ jest konsekwencją identyczności

$$\mathrm{E}\,\hat{\mu}_{MImp} = \mathrm{E}\mathrm{E}\left(\hat{\mu}_{MImp}\middle|\boldsymbol{J}\right).$$

Zgodnie ze wzorem (II.1.2) mamy

$$\mathrm{Var}\left(\hat{\mu}_{MImp}\middle|\boldsymbol{J}\right)=\frac{\sigma^2}{mn}\left(\frac{mn-1}{|\boldsymbol{J}|}+\frac{n+1}{n}-\frac{|\boldsymbol{J}|}{n}\right).$$

Ponieważ

$$\mathrm{Var}\,\hat{\mu}_{MImp} = \mathrm{E}\mathrm{Var}\left(\hat{\mu}_{MImp}\middle|\boldsymbol{J}\right)+\mathrm{Var}\mathrm{E}\left(\hat{\mu}_{MImp}\middle|\boldsymbol{J}\right)$$

i drugi składnik po prawej stronie jest równy zero otrzymujemy wzór (IV.1.1).■

TWIERDZENIE IV.1.2. ***Obciążenie estymatora Rubina wariancji*** $\hat{v}^2_{MImp}$ ***wynosi***

$$\mathrm{B}\hat{v}^2_{MImp}=\mathrm{E}\hat{v}^2_{MImp}-\mathrm{Var}\,\hat{\mu}_{MImp}=-\frac{\sigma^2}{n(n-1)}\left(\mathrm{E}|\boldsymbol{J}|+n(n+1)\mathrm{E}\frac{1}{|\boldsymbol{J}|}-2n+1\right). \quad \text{(IV.1.2.)}$$

Estymator nieobciążony wariancji estymatora $\hat{\mu}_{Imp}$ ma postać

$$\hat{v}_*^2 = \frac{\mathrm{E}\frac{1}{|\boldsymbol{J}|}}{1-\mathrm{E}\frac{1}{|\boldsymbol{J}|}}\left((n-1)\overline{U}_m - B_m\right) + \frac{1}{m}B_m.$$

Dowód. Dla estymatora Rubina wariancji $\hat{v}_{MImp}^2$, określonego wzorami (II.0.7)—(II.0.9), w przypadku wielokrotnej imputacji hot-deck zgodnie ze wzorem z dowodu twierdzenia II.1.2 mamy

$$\mathrm{E}\left(\overline{U}_m\middle|\boldsymbol{J}\right) = \frac{\sigma^2\left(|\boldsymbol{J}|-1\right)\left(n(n+1)-|\boldsymbol{J}|\right)}{|\boldsymbol{J}|n^2(n-1)} \quad \textit{oraz} \quad \mathrm{E}\left(B_m\middle|\boldsymbol{J}\right) = \frac{\sigma^2\left(|\boldsymbol{J}|-1\right)\left(n-|\boldsymbol{J}|\right)}{|\boldsymbol{J}|n^2}.$$

Zatem

$$\mathrm{E}\left(\hat{v}_{MImp}^2\middle|\boldsymbol{J}\right) = \frac{\sigma^2\left(|\boldsymbol{J}|-1\right)\left(2n-|\boldsymbol{J}|\right)}{|\boldsymbol{J}|n(n-1)} + \frac{\sigma^2\left(\boldsymbol{J}-1\right)\left(n-\boldsymbol{J}\right)}{|\boldsymbol{J}|mn^2}.$$

W konsekwencji

$$\mathrm{E}\hat{v}_{MImp}^2 = \frac{\sigma^2}{n(n-1)}\left(2n-1-\mathrm{E}|\boldsymbol{J}|-2n\mathrm{E}\frac{1}{|\boldsymbol{J}|}\right) + \frac{\sigma^2}{mn^2}\left(n+1-\mathrm{E}|\boldsymbol{J}|-n\mathrm{E}\frac{1}{|\boldsymbol{J}|}\right).$$

Ostatecznie, zgodnie ze wzorem (III.2.6) obciążenie estymatora $\hat{v}_{MImp}^2$ wynosi

$$\mathrm{E}\hat{v}_{MImp}^2 - \mathrm{Var}\,\hat{\mu}_{MImp} = \frac{\sigma^2}{n(n-1)}\left(2n-1-\mathrm{E}|\boldsymbol{J}|-2n\mathrm{E}\frac{1}{|\boldsymbol{J}|}\right) - \sigma^2\mathrm{E}\frac{1}{|\boldsymbol{J}|},$$

co po uproszczeniach daje wzór (IV.1.2).

Ponieważ $\mathrm{E}B_m = \frac{\sigma^2}{n^2}\left(n+1-\mathrm{E}|\boldsymbol{J}|-n\mathrm{E}\frac{1}{|\boldsymbol{J}|}\right)$, więc ze wzoru (IV.1.1) wynika, że wystarczy pokazać równość

$$\mathrm{E}\left((n-1)\overline{U}_m - B_m\right) = \sigma^2\left(1-\mathrm{E}\frac{1}{|\boldsymbol{J}|}\right),$$

a to wynika wprost ze wzorów na $\overline{U}_m$ i B_m. ■

Wniosek IV.1.3. ***Asymptotycznie względne obciążenie estymatora Rubina*** $\hat{v}^2_{MImp}$ ***wynosi***

$$\lim_{n\to\infty}\frac{\mathrm{E}\hat{v}^2_{MImp}-\mathrm{Var}\,\hat{\mu}_{MImp}}{\mathrm{Var}\,\hat{\mu}_{MImp}}=-\frac{(1-p)^2}{1+\dfrac{p(1-p)}{m}}. \tag{IV.1.3}$$

Dowód. Ze wzorów (IV.1.1) i (IV.1.2) wynika, że wzór na względne obciążenie można napisać w postaci

$$K_n=-\frac{\mathrm{E}|\boldsymbol{J}|+n(n+1)\mathrm{E}\dfrac{1}{|\boldsymbol{J}|}-2n+1}{\mathrm{E}\dfrac{n(n-1)}{|\boldsymbol{J}|}+\dfrac{n-1}{mn}\left(n+1-\mathrm{E}|\boldsymbol{J}|-n\mathrm{E}\dfrac{1}{|\boldsymbol{J}|}\right)}.$$

Ponieważ (Marciniak, Wesołowski, 1999)

$$\lim_{n\to\infty}n\mathrm{E}\frac{1}{|\boldsymbol{J}|}=\frac{1}{p},\quad\text{czyli}\quad\lim_{n\to\infty}\mathrm{E}\frac{1}{|\boldsymbol{J}|}=0,$$

więc dzieląc licznik i mianownik przez n i wykorzystując fakt, że

$$\lim_{n\to\infty}\frac{1}{n}\mathrm{E}|\boldsymbol{J}|=p$$

otrzymujemy

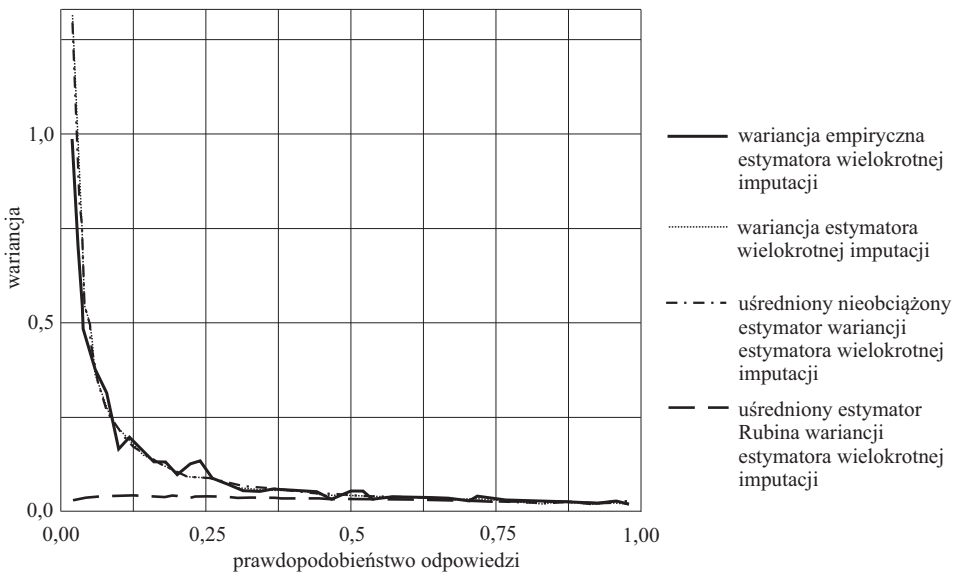
$$\lim_{n\to\infty}K_n=-\frac{\lim\limits_{n\to\infty}\dfrac{1}{n}\mathrm{E}|\boldsymbol{J}|+\lim\limits_{n\to\infty}(n+1)\mathrm{E}\dfrac{1}{|\boldsymbol{J}|}-2+\lim\limits_{n\to\infty}\dfrac{1}{n}}{\lim\limits_{n\to\infty}(n-1)\mathrm{E}\dfrac{1}{|\boldsymbol{J}|}+\dfrac{1}{m}\left(\lim\limits_{n\to\infty}\dfrac{n-1}{n}\right)\left(1+\lim\limits_{n\to\infty}\dfrac{1}{n}-\lim\limits_{n\to\infty}\dfrac{1}{n}\mathrm{E}|\boldsymbol{J}|-\lim\limits_{n\to\infty}\mathrm{E}\dfrac{1}{|\boldsymbol{J}|}\right)}=$$

$$=\frac{p+\dfrac{1}{p}-2}{\dfrac{1}{p}+\dfrac{1}{m}(1-p)},$$

co po uproszczeniach daje wzór (IV.1.3).■

Symulacyjna ilustracja wyników zawartych w twierdzeniach IV.1.1 i IV.1.2 przedstawiona jest na wykr. 3. Szczegółowy opis procedury symulacyjnej znajduje się w aneksie.

Wykr. 3. WYNIKI ESTYMACJI WARIANCJI ESTYMATORA ŚREDNIEJ DLA WIELOKROTNEJ IMPUTACJI (HOT-DECK) PRZY LOSOWYM BRAKU OBSERWACJI



Ź r ó d ł o: jak przy wykr. 1.

IV.2. Losowanie z rozkładu normalnego

TWIERDZENIE IV.2.1. ***Wartość oczekiwana estymatora imputacji wielokrotnej przy losowaniu z rozkładu normalnego wynosi***

$$\mathrm{E}\hat{\mu}_{MImp} = \mu,$$

a jego wariancja ma postać

$$\mathrm{Var}\,\hat{\mu}_{MImp} = \sigma^2\left(\mathrm{E}\frac{1}{|\boldsymbol{J}|} + \frac{n - \mathrm{E}|\boldsymbol{J}|}{mn^2}\right). \qquad \text{(IV.2.1)}$$

Dowód. Powyższe wzory wynikają wprost ze wzorów z twierdzenia II.1.4.■

TWIERDZENIE IV.2.2. ***Obciążenie estymatora Rubina wariancji*** $\hat{v}^2_{MImp}$ ***wynosi***

$$\mathrm{B}\hat{v}^2_{MImp} = \mathrm{E}\hat{v}^2_{MImp} - \mathrm{Var}\,\hat{\mu}_{MImp} = -\frac{\sigma^2}{n^2}\left(\frac{n-\mathrm{E}|\boldsymbol{J}|}{n-1} - n^2\mathrm{E}\frac{1}{|\boldsymbol{J}|} - \mathrm{E}|\boldsymbol{J}| + 2n\right). \quad \text{(IV.2.2)}$$

Estymator nieobciążony wariancji estymatora $\hat{\mu}_{MImp}$ ***ma postać***

$$\hat{v}^2_* = \left(n\mathrm{E}\frac{1}{|\boldsymbol{J}|}\right)\overline{U}_m + \left(\frac{1}{m} + \frac{n}{n-1}\mathrm{E}\frac{1}{|\boldsymbol{J}|}\right)B_m.$$

Dowód. Zgodnie ze wzorami (II.1.1) i (II.1.2) mamy

$$\mathrm{E}\overline{U}_m = \frac{\sigma^2}{n^2}\left(n - \frac{n-\mathrm{E}|\boldsymbol{J}|}{n-1}\right) \quad \text{oraz} \quad \mathrm{E}B_m = \frac{\sigma^2}{n^2}\left(n - \mathrm{E}|\boldsymbol{J}|\right).$$

W konsekwencji

$$\mathrm{E}\hat{v}^2_{MImp} - \mathrm{Var}\,\hat{\mu}_{MImp} = \frac{\sigma^2}{n^2}\left(n - \frac{n-\mathrm{E}|\boldsymbol{J}|}{n-1} + n - \mathrm{E}|\boldsymbol{J}|\right) - \sigma^2\mathrm{E}\frac{1}{|\boldsymbol{J}|},$$

co prowadzi do wzoru (IV.2.2).

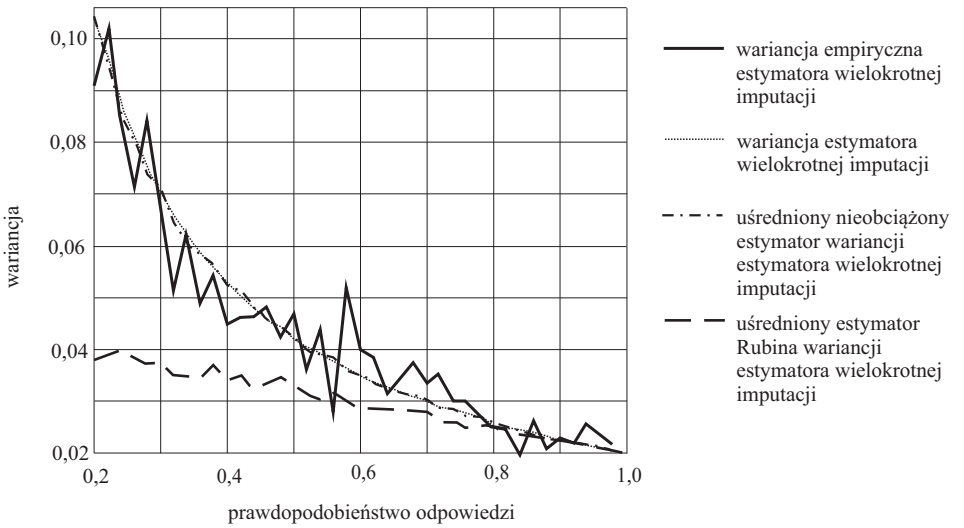
Ze wzoru (IV.2.1) wynika, że wystarczy znaleźć liczby α i β takie, że

$$\alpha\mathrm{E}\overline{U}_m + \beta\mathrm{E}B_m = \sigma^2\mathrm{E}\frac{1}{|\boldsymbol{J}|}.$$

Wtedy $\alpha\overline{U}_m + \left(\beta + \frac{1}{m}\right)B_m$ jest poszukiwanym estymatorem nieobciążonym. Przyjmując $\alpha = n\mathrm{E}\frac{1}{|\boldsymbol{J}|}$ z powyższej równości otrzymujemy $\beta = \frac{n}{n-1}\mathrm{E}\frac{1}{|\boldsymbol{J}|}$. ■

Symulacyjna ilustracja wyników zawartych w twierdzeniach IV.2.1 i IV.2.2 przedstawiona jest na wykr. 4. Szczegółowy opis procedury symulacyjnej znajduje się w aneksie.

Wykr. 4. WYNIKI ESTYMACJI WARIANCJI ESTYMATORA ŚREDNIEJ DLA WIELOKROTNEJ IMPUTACJI (HOT-DECK) Z ROZKŁADU NORMALNEGO I LOSOWYM BRAKU OBSERWACJI



Ź r ó d ł o: jak przy wykr. 1.

V. KONKLUZJE

W powyższym opracowaniu przedstawiono podstawowe idee i modele imputacji, w tym imputacji wielokrotnej na podstawie próbki niezależnych obserwacji o jednakowym rozkładzie, w której występują braki. W szczególności rozważono imputację średnią, imputacje typu hot-deck polegające na losowaniu respondentów oraz losowaniu z rozkładu normalnego z parametrami estymowanymi na podstawie obserwowanej części próbki oraz imputację regresyjną. Każda z tych metod jest szczególnym przypadkiem podejścia ogólnego analizowanego w rozdziałach I i III. W przypadku imputacji wielokrotnej wskazano, że popularny estymator wariancji, zwany estymatorem Rubina, nie ma dobrych własności, w szczególności jest obciążony. Zaproponowano wersje nieobciążone estymatorów typu Rubina w przypadku imputacji typu hot-deck metodą losowania respondentów i losowania z rozkładu normalnego. W artykule nie analizowano bayesowskiego modelu obserwacji, który jest ważnym elementem, zaproponowanego przez Rubina, podejścia wykorzystującego imputację wielokrotną. Planujemy przedstawić matematyczne podstawy imputacji w modelu bayesowskim w kolejnym opracowaniu.

prof. dr hab. Jacek Wesołowski — *GUS, Politechnika Warszawska*
mgr Jakub Tarczyński — *GUS*

LITERATURA

Andridge, R.R., Little, R.J.A. (2010). A review of hot deck imputation survey non-response. *International Statistical Revue*, vol. 78, no. 1, s. 40—64.

Carpenter, J.R., Kenward, M.G. (2013). *Multiple Imputation and its Application*. Wiley, Chichester.

de Waal, T., Pannekoek, J., Scholtus, S. (2011). *Handbook of Statistical Data Editing and Imputation*. Wiley, New York.

Donders, A.R.T., van der Heijden, G.J.M.G., Stijnen, T., Moons, K.G.M. (2006). Review: a gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology,* vol. 59, no. 10, s. 1087—1091.

Garson, G.D. (2012). Missing Values Analysis & Data Imputation. *Statistical Associate Publishers*, Asheboro.

Horton, N.J., Lipsitz, S.R. (2001). Multiple imputation in practice: comparison of software packages for regression models with missing data. *American Statistician, v*ol. 55, no. 3, s. 244—254.

Little, R.J.A., Rubin, D.B. (2002). *Statistical Analysis with Missing Data*. Wiley, New York.

Marciniak, E., Wesołowski, J. (1999). Asymptotic Eulerian expansions for binomial and negative binomial reciprocals. *Proceedings of The American Mathematical Society, v*ol. 127, s. 3329—3338.

Misztal, M. (2012). Imputation of missing data using R package. *Acta Universitatis Lodziensis, Folia Oeconomica,* vol. 269, s. 131—144.

Norazian Ramli, M.N., Yahaya, A.S., Ramli, N.A., Yusof, N.F.F.M., Abdullah, M.M.A. (2013). Roles of imputation methods for filling the missing values: a review. *Advances in Environmental Biology,* vol. 7, no. 12, s. 3861—3869.

Rempała, G.A. (2004). Asymptotic factorial powers expansions for binomial and negative binomial reciprocals. *Proceedings of The American Mathematical Society, v*ol. 32, no. 1, s. 261—272.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York.

van Buuren, S. (2012). *Flexible Imputation of Missing Data.* Chapman&Hall/CRC, London.

van Buuren, S., Groothuis-Oudshoorn, K. (2011). mice: multivariate imputation by chained equations in R. *Journal of Statistical Software, v*ol. 45, no. 3, s. 1—67, http://www.jstatsoft.org/.

Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman and Hall, Boston.

Szabłowski, P.J., Wesołowski, J., Wieczorkowski, R. (1996). Estymacja w podpopulacjach. *Wiadomości Statystyczne*, nr 7, s. 1—13.

ANEKS

Badania symulacyjne

Poniżej przedstawiony został opis badań symulacyjnych przeprowadzonych nad rozważanymi w pracy estymatorami wariancji wykorzystującymi imputację wielokrotną metodą typu hot-deck.

Symulacje polegały na $L = 100$-krotnym powtórzeniu następującego algorytmu:

1) generowano dane pochodzące z centralnego rozkładu normalnego z wariancją $\sigma^2 = 4$;
2) usuwano (losowo bądź nie, w zależności od rozważanego modelu) obserwacje z wygenerowanych danych;
3) generowano $m = 5$ próbek imputacyjnych w sposób określony przez rozważany model imputacyjny;

4) wyliczano estymator $\hat{\mu}_{MImp}$ oraz jego estymatory wariancji (Rubina — $\hat{v}^2_{MImp}$ i nieobciążony — $\hat{v}^2_*$) na podstawie wygenerowanych próbek imputacyjnych.

Na wszystkich rysunkach przedstawiono wykresy następujących wielkości:

— $\widehat{\mathrm{Var}}\left(\hat{\mu}_{MImp}\right)$ — wariancji empirycznej estymatora wielokrotnej imputacji $\hat{\mu}_{MImp}$, obliczanej na podstawie L powtórzeń eksperymentu

$$\widehat{\mathrm{Var}}\left(\hat{\mu}_{MImp}\right)=\frac{1}{L-1}\sum_{l=1}^{L}\left(\hat{\mu}^{(l)}_{MImp}-\frac{1}{L}\sum_{j=1}^{L}\hat{\mu}^{(j)}_{MImp}\right)^2,$$

gdzie $\hat{\mu}^{(l)}_{MImp}$ oznacza wartość estymatora wielokrotnej imputacji $\hat{\mu}_{MImp}$ w l-tym powtórzeniu eksperymentu, l=1, …, L;

— $\mathrm{Var}\hat{\mu}_{MImp}$ — wariancji estymatora wielokrotnej estymacji, obliczanej na podstawie wzorów: II.1.2 (wykr. 1), II.1.7 (wykr. 2), IV.1.1 w wersji $p_i = p$ (wykr. 3) oraz IV.2.1 (wykr. 4);

— $\overline{\hat{v}^2_*}$ — uśrednionej, na podstawie L powtórzeń eksperymentu, wartości $\hat{v}^2_*$ nieobciążonego estymatora wariancji estymatora wielokrotnej estymacji:

$$\overline{\hat{v}^2_*}=\frac{1}{L}\sum_{l=1}^{L}\hat{v}^2_{*,l},$$

gdzie $\hat{v}^2_{*,l}$ oznacza wartość estymatora $\hat{v}^2_*$ w l-tym powtórzeniu eksperymentu, $l = 1, \ldots, L$;

— $\overline{\hat{v}^2_{MImp}}$ — uśrednionej, na podstawie L powtórzeń eksperymentu, wartości $\hat{v}^2_{MImp}$ estymatora Rubina wariancji estymatora wielokrotnej imputacji:

$$\overline{\hat{v}^2_{MImp}}=\frac{1}{L}\sum_{l=1}^{L}\hat{v}^2_{MImp,l},$$

gdzie $\hat{v}^2_{MImp,l}$ oznacza wartość estymatora Rubina $\hat{v}^2_{MImp}$ w l-tym powtórzeniu eksperymentu, $l = 1, \ldots, L$.

Summary. *The article presents the basics of imputation methodology (including the methodology of multiple imputation), focusing on understanding its mathematical background. We analyze the situation when observations in the original sample are independent random variables with identical distributions,*

and response or its lack is modeled by a random mechanism which is independent of observations. In particular, we point out to problems that arise when the standard Rubin estimate of the multiple imputation variance estimator is used. A possible improvement of this popular estimator is indicated. The starting point of the analysis is when the appearance of response deficiencies is caused by a deterministic mechanism.

Keywords: imputation, multiple imputation, imputation estimator, Rubin estimator, mean imputation, hot-deck imputation, regression imputation.

Резюме. *В статье представлены основы импутационной методологии (в том числе методологии многократной импутации). Внимание в статье сосредоточено на прояснении математической стороны вопросов. Проанализирована ситуация, когда наблюдения формирующие оригинальную выборку являются независимыми случайными величинами с одинаковыми распределениями, а отсутствие ответов появляется случайно независимо от наблюдения. В частности статья указывает на проблемы, которые возникают когда используется стандартная оценка Рубина дисперсии оценки многократной импутации. В статье указано также на возможное улучшение этой популярной оценки. Отправной точкой анализа является ситуация, когда отсутствие ответов обясняет детерминический механизм.*

Ключевые слова: импутация, многократная импутация, импутационная оценка, оценка Рубина, импутация среднем, импутация типа hot-deck, регрессионная импутация.