PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 163
Ekonometria 30 2011

Beata Zmyślona
Uniwersytet Ekonomiczny we Wrocławiu

UWAGI NA TEMAT WŁASNOŚCI ESTYMATORÓW WYZNACZANYCH NA BAZIE NIEPEŁNYCH DANYCH

Streszczenie: W praktyce w przypadku niepełnych danych wykorzystuje się metody imputacji rozumiane jako metody szeroko rozumianej predykcji brakujących danych. Takie uzupełnione dane traktowane są tak, jak gdyby były obserwowalne, co powoduje niewłaściwe oszacowanie estymatora oraz wariancji estymatora, zwłaszcza jeżeli mechanizm generowania brakujących danych jest nielosowy. W artykule pokazano przykłady wpływu imputacji na własności estymatorów. Pewnym rozwiązaniem jest zastosowanie metod wielokrotnego uzupełniania danych, które uwzględniają tzw. błąd imputacji.

Słowa kluczowe: wnioskowanie statystyczne w przypadku niepełnych danych, macierz utraconej informacji na skutek brakujących danych, błąd imputacji, imputacja wielokrotna.

1. Wstęp

W badaniach społeczno-ekonomicznych stosunkowo często występuje zjawisko braku danych. Niekompletne zbiory danych stanowią problem zarówno z praktycznego, jak i z teoretycznego punktu widzenia. W celu przeciwdziałania niedogodnościom związanym z występowaniem brakujących danych albo usuwa się elementy z próby z brakującymi obserwacjami (w literaturze takie podejście nazywane jest analizą kompletnych przypadków), albo wykorzystuje się metody uzupełniania danych, czyli metody imputacji (określane jako metody szeroko pojętej predykcji danych). Analiza kompletnych przypadków sprawdza się jedynie w przypadku, gdy mechanizm powstawania brakujących danych jest losowy. W przeciwnym wypadku estymatory wyznaczane na bazie niekompletnego zbioru danych są obciążone oraz nieefektywne z powodu zredukowanego rozmiaru próby. Przy zastosowaniu metod imputacji uzupełnione dane traktowane są tak, jak gdyby były obserwowalne. Mimo iż metody te mogą w istotny sposób zmniejszać bądź też eliminować obciążenie estymatorów, mają jednak poważną wadę – zaniżają oceny wariancji estymatorów. Powoduje to, że konstrukcja przedziałów ufności dla szacowanych parametrów oraz testowanie hipotez są obarczone błędem wynikającym z tego niedoszacowania.

Metody wnioskowania statystycznego bazujące na tzw. technikach powiększania danych umożliwiają właściwe oszacowanie wariancji przez uwzględnienie tzw.

błędu imputacji. Jedną z metod powiększania danych jest metoda imputacji wielokrotnej zaproponowana przez Rubina (por. [1996; 1987]). W artykule przedstawiono, w jaki sposób za pomocą metody imputacji wielokrotnej uwzględnia się błąd imputacji przy wyznaczaniu wariancji estymatora bayesowskiego.

2. Wpływ analizy kompletnych przypadków oraz metod imputacji na własności estymatorów

W dalszej części artykułu zostaną przedstawione przykłady wpływu analizy tylko kompletnych przypadków oraz niektórych metod imputacji na obciążenie estymatora oraz jego wariancję.

W pierwszym przykładzie zakłada się, że wartości $y_1, y_2, ..., y_n$ wylosowane do próby są realizacjami zmiennej losowej o rozkładzie normalnym, czyli $Y \sim N(\mu, \sigma)$, przy czym przyjmuje się, że wariancja σ^2 zmiennej losowej Y jest znana. Wartości zmiennej losowej Y są obserwowane jedynie dla k obserwacji ($k < n$). Bez straty ogólności rozważań zbiór wartości zmiennej losowej Y dla n-elementowej próby zostaje podzielony na dwa podzbiory $\{y_{k_i}: i = 1, 2, ..., k\}$ oraz $\{y_{g_j}: j = k + 1, ..., g\}$, gdzie pierwszy zbiór odnosi się do obserwowanych, a drugi odpowiednio do nieobserwowanych wartości zmiennej Y.

Estymatorem nieobciążonym parametru μ jest średnia arytmetyczna wyznaczona na bazie obserwowanych danych: $\hat{\mu} = \overline{y}_{obs} = \frac{1}{k}\sum_{i=1}^{k} y_{k_i}$. Wariancja estymatora wyraża się następującym wzorem: $s^2_{\overline{y}_{obs}} = \frac{1}{k(k-1)}\sum_{i=1}^{k}\left(y_{k_i} - \overline{y}_{obs}\right)^2$. Jeżeli nieobserwowane wartości zmiennej Y zostaną zastąpione średnią arytmetyczną obliczoną na bazie obserwowanych wartości, czyli $y^*_{g_j} \equiv \overline{y}_{obs}$, wtedy estymatorem parametru μ wyznaczonym dla uzupełnionego zbioru danych jest średnia arytmetyczna

$$\hat{\mu} = \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{1}$$

wyznaczona dla n-elementowej próby. Jak łatwo zauważyć, średnia ta jest równa średniej arytmetycznej obliczonej tylko dla obserwowanych danych, czyli $\overline{y} = \overline{y}_{obs}$. Wariancja estymatora po uzupełnieniu danych $s^2_{\overline{y}} = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2$ może być przedstawiona jako

$$s^2_{\overline{y}} = \frac{1}{n(n-1)}\left(\sum_{i=1}^{k}\left(y_{k_i} - \overline{y}_{obs}\right)^2 + \sum_{j=k+1}^{n}\left(y^*_{g_j} - \overline{y}_{obs}\right)^2\right). \tag{2}$$

Wariancję $s^2_{\bar{y}}$ można zapisać jako $s^2_{\bar{y}} = \frac{k(k-1)}{n(n-1)} s^2_{\bar{y}_{obs}}$ ze względu na to, że drugi składnik sumy jest równy zeru. Ponieważ $s^2_{\bar{y}} < s^2_{\bar{y}_{obs}}$, zatem wariancja średniej jest niedoszacowana. Nasuwa się zatem wniosek, że uzupełnianie nieobserwowanych danych średnią arytmetyczną obliczoną na podstawie obserwowanych danych nie jest właściwą metodą, ponieważ wariancja estymatora jest niedoszacowana.

Drugi przykład pokazuje wpływ zastosowanej metody imputacji na obciążenie estymatorów parametrów modelu liniowej regresji $Y = \beta_0 + \beta_1 x + \varepsilon$. Zakłada się, że składnik losowy jest zmienną losową o rozkładzie normalnym $\varepsilon \sim N(0, \sigma)$, wariancja σ^2 jest nieznana. Jeżeli wszystkie wartości zmiennej Y oraz zmiennej X są obserwowane, wtedy estymatory parametrów β_0, β_1 oraz σ^2 wyznaczone metodą największej wiarygodności są nieobciążone i wyrażają się następującymi wzorami:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2},$$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-2},$$

gdzie $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$.

W analizowanym przykładzie przyjmuje się, że wartości zmiennej niezależnej X są obserwowane dla wszystkich elementów w próbie, natomiast wartości zmiennej zależnej Y są obserwowane tylko dla k jednostek. Zbiór wartości zmiennej losowej X dla n-elementowej próby zostaje podzielony na dwa podzbiory $\{y_{k_i}: i = 1, 2, ..., k\}$ oraz $\{y_{g_j}: j = k + 1, ..., g\}$, gdzie pierwszy zbiór odnosi się do obserwowanych, a drugi odpowiednio do nieobserwowanych wartości zmiennej Y. Przypuśćmy, że została zastosowana metoda imputacji polegająca na zastąpieniu nieobserwowanych wartości zmiennej Y średnią arytmetyczną $\bar{y}_{obs}$ obliczoną na podstawie obserwowanych danych, czyli $y^*_{g_j} \equiv \bar{y}_{obs} = \frac{1}{k} \sum_{i=1}^{k} y_{k_i}$. Estymatory parametrów β_0, β_1 oraz σ^2 wyrażają się następującymi wzorami:

$$\hat{\beta}^*_0 = \bar{y}_{obs} - \hat{\beta}^*_1 \bar{x},$$

$$\hat{\beta}_1^* = \frac{\sum_{i=1}^{k}\left(x_{k_i}-\overline{x}\right)\left(y_{k_i}-\overline{y}_{obs}\right)+\sum_{j=k+1}^{n}\left(x_{g_j}-\overline{x}\right)\left(y_{g_j}^*-\overline{y}_{obs}\right)}{\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2},$$
$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{k}\left(y_{k_i}-\hat{y}_{k_i}\right)^2+\sum_{j=k+1}^{n}\left(y_{g_j}^*-\hat{y}_g^*\right)^2}{n-2}. \tag{3}$$

Warunkowe wartości oczekiwane estymatorów danych wzorem wyrażają się w następujący sposób:

$$E\left(\hat{\beta}_0\left|\{x_i\}_{i=1}^n\right.\right) = \beta_0+\beta_1\left\{\overline{x}_k-\frac{\overline{x}\sum_{i=1}^{k}\left(x_{k_i}-\overline{x}_k\right)^2}{\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2}\right\},$$
$$E\left(\hat{\beta}_1\left|\{x_i\}_{i=1}^n\right.\right) = \beta_1\left\{\frac{\sum_{i=1}^{k}\left(x_{k_i}-\overline{x}_k\right)^2}{\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2}\right\},$$
$$E\left(\hat{\sigma}\left|\{x_i\}_{i=1}^n\right.\right) = \frac{\sigma^2}{n-2}\left\{(k-1)-\left\{\overline{x}_k-\frac{\overline{x}\sum_{i=1}^{k}\left(x_{k_i}-\overline{x}_k\right)^2}{\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2}\right\}\right\}$$
$$+\frac{\beta_1}{n-2}\left\{\sum_{i=1}^{k}\left(x_{k_i}-\overline{x}_k\right)^2-\left\{1-\frac{\sum_{i=1}^{k}\left(x_{k_i}-\overline{x}_k\right)^2}{\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2}\right\}\right\}. \tag{4}$$

Jak łatwo zauważyć, estymatory są obciążone.

W powyższych dwóch przykładach zakładano *implicite*, że mechanizm powstawania brakujących danych był losowy. W kolejnym przykładzie przyjmuje się nielosowy mechanizm powstawania brakujących danych. Przyjmuje się, że próba losowa pochodzi z rozkładu normalnego ze znaną wariancją σ^2. Wartość zmiennej Y nie jest obserwowana, jeżeli przekracza pewną stałą c. Jeśli zastosuje się analizę kompletnych przypadków i przyjmie za estymator parametru μ średnią arytmetyczną odpowiadającą tylko obserwowanym danym, to wartość tego estymatora będzie zaniżona. W takim przypadku powinna zostać zastosowana taka procedura estymacji, która uwzględniałaby mechanizm powstawania brakujących danych.

Z wyżej przedstawionych przykładów można wysunąć wniosek, że proponowane w praktyce rozwiązania dotyczące analizy niepełnych danych nie dają w wielu wypadkach właściwych rozwiązań (oceny estymatorów oraz ich wariancji są zaniżone bądź zawyżone). Obciążenie estymatorów można eliminować przez wykorzystanie takich metod imputacji, w których dane generowane są z warunkowych rozkładów brakujących danych (np. metody wykorzystujące bootstraping). Natomiast w celu poprawnego oszacowania wariancji w przypadku niekompletnych zbiorów danych należy wziąć pod uwagę błąd wynikający ze stosowania określonych technik uzupełniania danych, tzw. błąd metody imputacji. Metoda imputacji wielokrotnej umożliwia wzięcie pod uwagę błędu imputacji.

3. Wnioskowanie bayesowskie metodą imputacji wielokrotnej

Przyjmuje się, że obserwacje $\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n$ są realizacjami wektora losowego $\mathbf{Y}_i = \left(Y_{i1}, Y_{i2}, ..., Y_{ip}\right)$ dla i = 1, 2, ..., n. Rozkład wektora losowego zależy od wektora nieznanych parametrów $\boldsymbol{\theta} \subseteq \boldsymbol{\Theta} \in R^d$. W przypadku niepełnych danych wektor $\mathbf{Y}_i$ zapisuje się symbolicznie jako $\mathbf{Y}_i = \left(\mathbf{Y}_{i(obs)}, \mathbf{Y}_{i(br)}\right)$, gdzie pierwszy podwektor odnosi się do obserwowanych zmiennych, natomiast drugi do nieobserwowanych zmiennych dla i-tego elementu próby. Przy założeniu, że realizacje wektora losowego są niezależne, funkcję wiarygodności można przedstawić jako (por. [Rubin 1987; Schafer 2000])

$$p\left(\mathbf{y}|\boldsymbol{\theta}\right) = \prod_{i=1}^{n} p\left(\mathbf{y}_{i(obs)}|\boldsymbol{\theta}\right) p\left(\mathbf{y}_{i(br)}\middle|\mathbf{y}_{i(obs)}, \boldsymbol{\theta}\right), \tag{5}$$

gdzie $\mathbf{y}$ jest macierzą, której wierszami są wektory $\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n$. Wyrażenia $p\left(\mathbf{y}_{obs}|\boldsymbol{\theta}\right)$ oraz $p\left(\mathbf{y}_{i(br)}\middle|\mathbf{y}_{i(obs)}, \boldsymbol{\theta}\right)$ oznaczają odpowiednio rozkład prawdopodobieństwa podwektora $\mathbf{Y}_{i(obs)}$ odnoszącego się do obserwowanych zmiennych oraz warunkowy rozkład prawdopodobieństwa brakujących danych. W przypadku metod powiększania danych stosuje się bayesowskie podejście do wnioskowania statystycznego o parametrach, w którym nieobserwowane wielkości, tzn. parametry oraz nieobserwowane wartości zmiennych $Y_1, Y_2, ..., Y_p$, są traktowane jako zmienne losowe. Wnioskowanie o wektorze parametrów $\boldsymbol{\theta}$ bazuje na rozkładzie *a posteriori* wyrażonym następującym wzorem (por. np. [Gosh, Delampady, Samanta 2006, s. 31]):

$$p\left(\boldsymbol{\theta}|\mathbf{y}\right) = \frac{p\left(\mathbf{y}|\boldsymbol{\theta}\right) p\left(\boldsymbol{\theta}\right)}{\int p\left(\mathbf{y}|\boldsymbol{\theta}\right) p\left(\boldsymbol{\theta}\right) d\boldsymbol{\theta}}, \tag{6}$$

gdzie $p(\boldsymbol{\theta})$ jest rozkładem *a priori* wektora parametrów $\boldsymbol{\theta}$. Rozkład *a posteriori* ma interesujące własności. Jedną z nich jest jego zbieżność do gęstości wielowymiarowego rozkładu normalnego przy spełnieniu ogólnych warunków regularności typu Craméra-Rao dla gęstości rozkładów prawdopodobieństwa ([DeGroot 1981, s. 175-190;

por. Krzyśko 2000, s. 55-62]). Oznacza to, że wektor parametrów $\boldsymbol{\theta}$ ma asymptotyczny *d*-wymiarowy rozkład normalny, gdzie wektor wartości oczekiwanych tego rozkładu odpowiada wektorowi $\hat{\boldsymbol{\theta}}$ estymatorów największej wiarygodności wektora parametrów $\boldsymbol{\theta}$, natomiast macierz kowariancji jest równa odwrotnej macierzy informacji Fishera podzielonej przez *n*, co symbolicznie można zapisać w następujący sposób

$$\boldsymbol{\theta} \xrightarrow{D} N\left(\hat{\boldsymbol{\theta}}, \frac{I_{\mathbf{y}}^{-1}(\boldsymbol{\theta})}{n}\right). \tag{7}$$

Elementami macierzy informacji Fishera $I_{\mathbf{y}}(\boldsymbol{\theta}) = E\left(-\frac{\partial^2 \log p(\mathbf{y}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right)$ są drugie pochodne logarytmu funkcji wiarygodności $\frac{\partial^2 \log p(\mathbf{y}|\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j}$. Funkcja informacji Fishera określona jest jako miara informacji o parametrze $\boldsymbol{\theta}$ zawarta w próbie (por. [Daniels, Hogan 2008, s. 41; Krzyśko 2000, s. 57-58]).

W przypadku analizy niepełnych danych istotne jest ustalenie, jaka część informacji o szacowanym parametrze jest utracona na skutek brakujących danych. Macierz brakujących informacji definiowana jest zgodnie z zasadą Orcharda i Woodbury'ego w następujący sposób (por. [Orchard, Woodbury 1972]). Zdefiniujmy macierz Fishera dla obserwowanej części danych jako $I_{\mathbf{y}_{obs}}(\boldsymbol{\theta}) = E\left(-\frac{\partial^2 \log p(\mathbf{y}_{obs}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right)$. Druga pochodna logarytmu funkcji wiarygodności wyraża się jako

$$-\frac{\partial^2 \log p(\mathbf{y}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} = -\frac{\partial^2 \log p(\mathbf{y}_{obs}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} - \frac{\partial^2 \log p(\mathbf{y}_{br}|\mathbf{y}_{obs},\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}. \tag{8}$$

Wartość oczekiwana wyrażenia wyznaczona względem warunkowego rozkładu brakujących danych jest postaci

$$\begin{aligned} &-\int \frac{\partial^2 \log p(\mathbf{y}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} p(\mathbf{y}_{br}|\mathbf{y}_{obs},\boldsymbol{\theta})\, d\mathbf{y}_{br} = \\ &= -\int \left(\frac{\partial^2 \log p(\mathbf{y}_{obs}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} + \frac{\partial^2 \log p(\mathbf{y}_{br}|\mathbf{y}_{obs},\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right) p(\mathbf{y}_{br}|\mathbf{y}_{obs},\boldsymbol{\theta})\, d\mathbf{y}_{br} \end{aligned}. \tag{9}$$

Na bazie wartości oczekiwanej formułuje się zasadę, że macierz informacji Fishera $I_{\mathbf{y}}(\boldsymbol{\theta})$ jest sumą macierzy informacji Fishera dla obserwowanych danych $I_{\mathbf{y}_{obs}}(\boldsymbol{\theta})$ oraz macierzy utraconej informacji na skutek brakujących danych $I_{\mathbf{y}_{br}}(\boldsymbol{\theta})$, a zatem spełniona jest następująca równość (por. [Schafer 2000, s. 62-63]):

$$I_{\mathbf{y}}(\boldsymbol{\theta}) = I_{\mathbf{y}_{obs}}(\boldsymbol{\theta}) + I_{\mathbf{y}_{br}}(\boldsymbol{\theta}). \tag{10}$$

We wnioskowaniu bayesowskim można dodatkowo uwzględnić trzeci składnik macierzy informacji Fishera, a mianowicie informację uzyskaną z rozkładu *a priori* [Schafer 2000, s. 66-67]. W dalszych rozważaniach ten składnik pomija się, przyjmując, że dla dużych prób rozkład *a posteriori* zależy coraz bardziej od funkcji wiarygodności, a w coraz mniejszym stopniu od rozkładu *a priori*.

Oszacowanie macierzy kowariancji wektora estymatorów parametrów $\boldsymbol{\theta}$ jest równoznaczne z wyznaczeniem macierzy kowariancji rozkładu *a posteriori*, przy czym macierz informacji Fishera wyraża się wzorem.

4. Procedura wyznaczania estymatora bayesowskiego metodą imputacji wielokrotnej

Procedura wyznaczania bayesowskiego estymatora wektora parametrów $\boldsymbol{\theta}$ składa się z czterech kroków, które są wykonywane iteracyjnie ([Rubin 1996; 1987; Schafer 2000, s. 104-118]). W pierwszym kroku generowane są *m*-krotnie próbki pseudolosowe z warunkowego rozkładu brakujących danych $y_{ijk(br)} \sim p\left(\mathbf{y}_{br} \middle| \mathbf{Y}_{obs} = \mathbf{y}_{obs}, \tilde{\boldsymbol{\theta}}_{nj}\right)$, gdzie $y_{ijk(br)}$ jest wartością dla *k*-tej zmiennej odpowiadającą *i*-temu elementowi próby w *j*-tym zbiorze danych, natomiast $\tilde{\boldsymbol{\theta}}_{nj}$ jest realizacją wektora losowego wygenerowanego z rozkładu *a posteriori* wyznaczonego dla *j*-tego zbioru danych ($\tilde{\boldsymbol{\theta}}_{nj} \sim p\left(\boldsymbol{\theta}_n \middle| \mathbf{Y}_{obs} = \mathbf{y}_{obs}, \mathbf{Y}_{br} = \mathbf{y}_{br}^{j}\right)$).

W drugim kroku tworzonych jest *m* kompletnych zbiorów danych przez uzupełnienie każdej nieobserwowanej wartości *m* razy. Kompletne zbiory danych oznaczone są przez $\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, ..., \mathbf{y}^{(m)}$.

W kroku trzecim dla każdego zbioru danych wyznaczany jest estymator $\tilde{\boldsymbol{\theta}}_{nj}$ dla j=1, 2, ..., m.

Ostatecznie estymator bayesowski wyznaczany jest zgodnie z formułą

$$\overline{\boldsymbol{\theta}}_m = \frac{1}{m}\sum_{j=1}^{m} \hat{\boldsymbol{\theta}}_{nj}. \tag{11}$$

Kolejnym etapem analizy jest wyznaczenie wariancji estymatora .

5. Procedura wyznaczania wariancji estymatora bayesowskiego

W procedurze wyznaczania wariancji estymatora w przypadku niepełnych danych powinno się uwzględniać, oprócz tzw. błędu próbkowania, błąd wynikający z zastosowania metody uzupełniania danych (czyli metody imputacji). W procedurze tej wykorzystuje się wnioski z następującego twierdzenia.

Twierdzenie 1

Niech $\mathbf{Y}_1, \mathbf{Y}_2, ..., \mathbf{Y}_n$ będzie ciągiem wektorów losowych. Załóżmy, że dana jest funkcja tych wektorów $f_n\left(\mathbf{Y}_1, \mathbf{Y}_2, ..., \mathbf{Y}_n\right)$ oraz wektory losowe $\mathbf{Z}_n$ i $\xi_n \in \mathbb{R}^d$. Jeżeli

wektory $\sqrt{n}\left(\mathbf{Z}_n - f_n\left(\mathbf{Y}_1, \mathbf{Y}_2, ..., \mathbf{Y}_n\right)\right)$ oraz $\sqrt{n}\left(f_n\left(\mathbf{Y}_1, \mathbf{Y}_2, ..., \mathbf{Y}_n\right) - \boldsymbol{\xi}_n\right)$ mają asymptotyczne *d*-wymiarowe rozkłady normalne

$$\begin{array}{l} \sqrt{n}\left(\mathbf{Z}_n - f_n\left(\mathbf{Y}_1, \mathbf{Y}_2, ..., \mathbf{Y}_n\right)\right) \xrightarrow{D} N\left(\mathbf{0}, \boldsymbol{\Sigma}_1\right) \\ \sqrt{n}\left(f_n\left(\mathbf{Y}_1, \mathbf{Y}_2, ..., \mathbf{Y}_n\right) - \boldsymbol{\xi}_n\right) \xrightarrow{D} N\left(\mathbf{0}, \boldsymbol{\Sigma}_2\right) \end{array}, \tag{12}$$

wtedy

$$\sqrt{n}\left(\mathbf{Z}_n - \boldsymbol{\xi}_n\right) \xrightarrow{D} N\left(\mathbf{0}, \boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2\right). \tag{13}$$

Dowód twierdzenia znajduje się w pracy [Nielsen 2003].

Przez $\overline{\boldsymbol{\theta}}_m$ oraz $\boldsymbol{\theta}_0$ oznaczmy odpowiednio wektor estymatorów wyznaczany na bazie *m* uzupełnionych zbiorów danych dany wzorem oraz wektor prawdziwych wartości parametrów. Wektor $\left(\overline{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_0\right)$ można zapisać jako następującą sumę dwóch wektorów $\left(\overline{\boldsymbol{\theta}}_m - \hat{\boldsymbol{\theta}}_n\right) + \left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right)$, gdzie $\hat{\boldsymbol{\theta}}_n$ oznacza wektor estymatorów wyznaczony tylko na bazie obserwowanych danych. Na bazie własności asymptotycznego rozkładu *a posteriori* wektora parametrów $\boldsymbol{\theta}$ można przedstawić macierz kowariancji wektora $\sqrt{n}\left(\overline{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_0\right)$ jako następującą sumę [Nielsen 2003]

$$V\left(\sqrt{n}\left(\overline{\boldsymbol{\theta}}_m - \boldsymbol{\theta}_0\right)\right) = V\left(\sqrt{n}\left(\overline{\boldsymbol{\theta}}_m - \hat{\boldsymbol{\theta}}_n\right)\right) + V\left(\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right)\right). \tag{14}$$

Pierwszy składnik sumy odnosi się do zmienności związanej z zastosowaniem metody imputacji (tzw. błąd imputacji), natomiast drugi składnik sumy odnosi się do tzw. zmienności próbkowej. Macierz kowariancji wektora $\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right)$ jest równa odwrotnej macierzy informacji Fishera odpowiadającej obserwowanym danym, czyli $I^{-1}_{\mathbf{y}_{obs}}\left(\boldsymbol{\theta}\right)$. Problemem pozostaje wyznaczenie macierzy kowariancji wektora $\sqrt{n}\left(\overline{\boldsymbol{\theta}}_m - \hat{\boldsymbol{\theta}}_n\right)$. Krokiem pośrednim do wyznaczenia tej macierzy jest wyznaczenie macierzy kowariancji dla wektora $\sqrt{n}\left(\hat{\boldsymbol{\theta}}_{nj} - \hat{\boldsymbol{\theta}}_n\right)$, gdzie $\hat{\boldsymbol{\theta}}_{nj}$ jest wektorem wartości estymatorów wyznaczonym dla *j*-tego zbioru danych $\mathbf{y}^{(j)}$ (dla $j = 1, 2, ..., m$). Wektor wartości parametrów $\hat{\boldsymbol{\theta}}_{nj}$ uzyskuje się przez rozwiązanie następującego równania

$$\sqrt{n}\left(\frac{\partial \log p\left(\mathbf{y}^{(j)}|\boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}}\right)_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{nj}} = \mathbf{0}. \tag{15}$$

Jeżeli pochodna logarytmu funkcji wiarygodności zostanie rozpisana zgodnie z twierdzeniem Lagrange'a o wartości średniej, wtedy równanie można przedstawić w następujący sposób

$$\sqrt{n}\left(\frac{\partial \log p\left(\mathbf{y}^{(j)}|\boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}}\right)_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_n} + \sqrt{n}\left(\hat{\boldsymbol{\theta}}_{nj} - \hat{\boldsymbol{\theta}}_n\right)\left(\frac{\partial^2 \log p\left(\mathbf{y}^{(j)}|\boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{'}}\right)_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}^*_{nj}} = \mathbf{0}, \tag{16}$$

gdzie $\hat{\boldsymbol{\theta}}^*_{nj}$ jest punktem pośrednim pomiędzy $\hat{\boldsymbol{\theta}}_n$ oraz $\hat{\boldsymbol{\theta}}_{nj}$, takim że $\left(\hat{\boldsymbol{\theta}}_n < \hat{\boldsymbol{\theta}}^*_{nj} < \hat{\boldsymbol{\theta}}_{nj}\right)$.

Funkcję wiarygodności $p\left(\mathbf{y}^{(j)}|\boldsymbol{\theta}\right)$ dla *j*-tego zbioru danych można przedstawić jako sumę dwóch składników (por. [Tan, Tian, Wang Ng 2010, s. 36-37])

$$p\left(\mathbf{y}^{(j)}|\boldsymbol{\theta}\right)=p\left(\mathbf{y}_{obs}|\boldsymbol{\theta}\right)+p\left(\mathbf{y}_{br}^{(j)}|\mathbf{y}_{obs},\boldsymbol{\theta}\right), \tag{17}$$

gdzie pierwszy składnik sumy odnosi się do obserwowanych danych, natomiast drugi do uzupełnionego *j*-tego zbioru danych. Ponieważ funkcja wiarygodności jest sumą wyrażoną wzorem, zatem równanie przybiera postać

$$\begin{aligned}&\sqrt{n}\left(\frac{\partial \log p\left(\mathbf{y}_{obs}|\boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}}\right)_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_n}+\sqrt{n}\left(\frac{\partial \log p\left(\mathbf{y}_{br}^{(j)}|\mathbf{y}_{obs},\boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}}\right)_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_n}+\\&+\sqrt{n}\left(\hat{\boldsymbol{\theta}}_{nj}-\hat{\boldsymbol{\theta}}_n\right)\left(\frac{\partial^2 \log p\left(\mathbf{y}_{obs}|\boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}+\frac{\partial^2 \log p\left(\mathbf{y}_{br}^{(j)}|\mathbf{y}_{obs},\boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right)_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{nj}^{*}}=\mathbf{0}.\end{aligned} \tag{18}$$

Pochodna logarytmu funkcji wiarygodności dla obserwowanych danych obliczona w punkcie $\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_n$ równa jest zeru, a zatem pierwszy składnik sumy jest równy zeru (por. [Schafer 2000, s. 58]). Logarytm funkcji wiarygodności wyznaczony dla *j*-tego zbioru danych wygenerowanych z rozkładu zależnego od wektora parametrów $\tilde{\boldsymbol{\theta}}_{jn}$ zgodnie z twierdzeniem Lagrange'a o wartości średniej jest równy

$$\begin{aligned}&\sqrt{n}\left(\frac{\partial \log p\left(\mathbf{y}_{br}^{(j)}|\mathbf{y}_{obs},\boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}}\right)_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_n}=\\&=\sqrt{n}\left(\frac{\partial \log p\left(\mathbf{y}_{br}^{(j)}|\mathbf{y}_{obs},\boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}}\right)_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}_{jn}}+\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n-\tilde{\boldsymbol{\theta}}_{jn}\right)\left(\frac{\partial^2 \log p\left(\mathbf{y}_{br}^{(j)}|\mathbf{y}_{obs},\boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right)_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}_{jn}^{*}}.\end{aligned} \tag{19}$$

Podstawiając prawą stronę wyrażenia za drugi składnik sumy oraz uwzględniając fakt, że pierwszy składnik jest równy zeru, otrzymujemy:

$$\begin{aligned}&\sqrt{n}\left(\frac{\partial \log p\left(\mathbf{y}_{br}^{(j)}|\mathbf{y}_{obs},\boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}}\right)_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}_{jn}}+\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n-\tilde{\boldsymbol{\theta}}_{jn}\right)\left(\frac{\partial^2 \log p\left(\mathbf{y}_{br}^{(j)}|\mathbf{y}_{obs},\boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right)_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}_{jn}^{*}}+\\&+\sqrt{n}\left(\hat{\boldsymbol{\theta}}_{nj}-\hat{\boldsymbol{\theta}}_n\right)\left(\frac{\partial^2 \log p\left(\mathbf{y}_{obs}|\boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}+\frac{\partial^2 \log p\left(\mathbf{y}_{br}^{(j)}|\mathbf{y}_{obs},\boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right)_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{nj}^{*}}=\mathbf{0}.\end{aligned} \tag{20}$$

W powyższych rozważaniach przyjmuje się, że model kompletnych danych oraz model obserwowanych danych spełniają warunki regularności w sensie Craméra-Rao. Wiadomo, że wektor $\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n-\tilde{\boldsymbol{\theta}}_{jn}\right)$ ma asymptotyczny rozkład normalny

o wartości oczekiwanej równej wektorowi zerowemu, a macierzy kowariancji równej odwrotnej macierzy informacyjnej Fishera dla obserwowanych danych (por. [Gosh, Delampady, Samanta 2006, s. 6-8, 15; Nielsen 2003]).

Wykorzystując opisane powyżej założenia odnośnie do rozkładów, można pokazać, że wektor $\sqrt{n}\left(\hat{\boldsymbol{\theta}}_{nj}-\hat{\boldsymbol{\theta}}_{n}\right)$ ma asymptotycznie rozkład normalny o wartości oczekiwanej równej wektorowi zerowemu, a macierzy kowariancji równej odwrotnej macierzy informacyjnej Fishera dla obserwowanych danych minus odwrotnej macierzy informacyjnej $I_{\mathbf{y}}(\boldsymbol{\theta})$, co można zapisać jako

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_{nj}-\hat{\boldsymbol{\theta}}_{n}\right)\xrightarrow{D}N\left(\mathbf{0},\left(I^{-1}_{\mathbf{y}_{obs}}(\boldsymbol{\theta})-I^{-1}_{\mathbf{y}}(\boldsymbol{\theta})\right)\right). \tag{21}$$

Na bazie estymatorów $\hat{\boldsymbol{\theta}}_{n1},\hat{\boldsymbol{\theta}}_{n2},...,\hat{\boldsymbol{\theta}}_{nm}$ wyznaczonych dla *m* uzupełnionych zbiorów danych wyznacza się estymator $\bar{\boldsymbol{\theta}}_{m}=\frac{1}{m}\sum_{j=1}^{m}\hat{\boldsymbol{\theta}}_{nj}$. Wektor $\sqrt{n}\left(\bar{\boldsymbol{\theta}}_{m}-\hat{\boldsymbol{\theta}}_{n}\right)$ ma asymptotyczny rozkład normalny o niżej podanych parametrach [Nielsen 2003]

$$\sqrt{n}\left(\bar{\boldsymbol{\theta}}_{m}-\hat{\boldsymbol{\theta}}_{n}\right)\xrightarrow{D}N\left(\mathbf{0},\frac{1}{m}\left(I^{-1}_{\mathbf{y}_{obs}}(\boldsymbol{\theta})-I^{-1}_{\mathbf{y}}(\boldsymbol{\theta})\right)\right). \tag{22}$$

Błąd imputacji jest mierzony zatem za pomocą wariancji rozkładu . Ostatecznie wariancja estymatora $\bar{\boldsymbol{\theta}}_{m}=\frac{1}{m}\sum_{j=1}^{m}\hat{\boldsymbol{\theta}}_{j}$ wyraża się jako suma składnika odnoszącego się do zmienności próbkowej oraz błędu imputacji, co można zapisać jako [Nielsen 2003]

$$\sqrt{n}\left(\bar{\boldsymbol{\theta}}_{n}-\boldsymbol{\theta}_{0}\right)\xrightarrow{D}N\left(\mathbf{0},I^{-1}_{\mathbf{y}_{obs}}(\boldsymbol{\theta})+\frac{1}{m}\left(I^{-1}_{\mathbf{y}_{obs}}(\boldsymbol{\theta})-I^{-1}_{\mathbf{y}}(\boldsymbol{\theta})\right)\right). \tag{23}$$

W metodzie imputacji wielokrotnej macierz kowariancji jest szacowana zgodnie z podaną niżej procedurą. Najpierw dla każdego zbioru danych oblicza się macierz kowariancji $\hat{\mathbf{U}}_{(j)}$. Macierz kowariancji reprezentującą średnią zmienność wewnątrzgrupową oznaczamy przez $\bar{\mathbf{U}}_{m}=\frac{1}{m}\sum_{l=1}^{m}\hat{\mathbf{U}}_{(j)}$, natomiast macierz kowariancji reprezentującą zmienność międzygrupową przez $\mathbf{B}_{m}=\frac{1}{m-1}\sum_{l=1}^{m}\left(\hat{\boldsymbol{\theta}}_{(j)}-\bar{\boldsymbol{\theta}}_{n}\right)^{T}\left(\hat{\boldsymbol{\theta}}_{(j)}-\bar{\boldsymbol{\theta}}_{n}\right)$ Estymator macierzy kowariancji rozkładu jest sumą dwóch macierzy, a mianowicie $\mathbf{T}_{m}=\bar{\mathbf{U}}_{m}+(1+m^{-1})\mathbf{B}_{m}$ (zob. [Rubin 1996; 1987]).

6. Podsumowanie

Stosowane w praktyce metody zastępowania brakujących danych pewnymi wartościami oraz traktowanie tych wartości tak, jak gdyby były obserwowane, powoduje, że oceny estymatorów parametrów oraz wariancje tych estymatorów są źle oszacowane (albo przeszacowane albo niedoszacowane). Pewnym rozwiązaniem tego problemu jest zastosowanie metod bazujących na wielokrotnym uzupełnianiu danych.

Metoda imputacji wielokrotnej w niektórych wypadkach przeszacowuje błąd wynikający z imputacji, szczególnie wtedy, gdy mechanizm generowania brakujących danych jest nielosowy (przykłady przeszacowania tego błędu podane są w pracy [Nielsen 2003]). W takiej sytuacji dokładniejsze oszacowanie błędu imputacji uzyskuje się za pomocą metody symulacji parametrycznej. W metodzie symulacji parametrycznej zarówno ocena estymatora parametru, jak i macierz kowariancji estymatora są wyznaczane na bazie wygenerowanych za pomocą metody Monte Carlo wartości wektorów parametrów $\tilde{\boldsymbol{\theta}}_{inj}$ z rozkładu *a posteriori*. Otrzymany w taki sposób estymator bayerowski oraz jego wariancja są estymatorami Rao-Blackwella wartości oczekiwanej i macierzy kowariancji rozkładu *a posteriori* (por. [Congdon 2003; Gosh, Delampady, Samanta 2006, s. 13-14; Tan, Tian, Wang 2010, s. 23-25; Tanner, Wong 1987; Schafer 2000, s. 89-104]).

Literatura

Congdon P., *Applied Bayesian Modelling*, Wiley, New York 2003.

Daniels M., Hogan J., *Missing Data In Longitudinal Studies, Strategies for Bayesian Modeling and Sensity Analysis,* Chapman & Hall/CRC, Boca Raton, London, New York 2008.

DeGroot M., *Optymalne decyzje statystyczne,* PWN, Warszawa 1981.

Gosh J., Delampady M., Samanta T., *An Introduction to Bayesian Ananlysis, Theory and Methods,* Springer, New York 2006.

Krzyśko M., *Wielowymiarowa analiza statystyczna,* Wydawnictwo Uniwersytetu A. Mickiewicza, Poznań 2000.

Nielsen S., *Proper and improper multiple imputation,* „International Statistical Review" 2003 no 71.

Orchard T., Woodbury M., *A Missing Information Principle: Theory and Applications,* Sixth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, University of California Press 1972.

Rubin D., *Multiple imputation after 18+years,* „JASA" 1996 no 91 .

Rubin D., *Multiple Imputation for Nonresponse in Surveys*, John Willey & Sons, New York 1987.

Schafer J., *Analysis of Incomplete Multivariate Data,* Chapman & Hall, New York 2000.

Tan M., Tian G., Wang Ng K., *Bayesian Missing Data Problems,* Chapman & Hall/CRC, Boca Raton, London, New York 2010.

Tanner M., Wong W., *The calculation of posterior distributions by data augmentation*, „JASA" 1987 no 82.

NOTES ON THE ESTIMATORS PROPERTIES IN CASE OF IMPUTED DATA SETS

Summary: In current surveys practice the most common method for handling non-response item is simple imputation. The simple imputation methods are the prediction methods for missing data. The imputed values are treated as if they were observed. This results in under or overestimation of the estimator variance, especially if the missing data mechanism is non-random. Simple imputation is inappropriate when the goal is to construct test statistics and confidence of intervals. The paper shows examples of the impact of imputation on the estimates properties. One solution is to use multiple imputation methods that take into account the so-called imputation error.