Imputacja danych niekompletnych

Przemysław Szymczak

Wprowadzenie

Celem projektu jest zbadanie jakości różnych metod imputacji braków danych. Zbiór początkowy zawierał 467 obserwacji dla 7 zmiennych. Na sześciu z nich, będących zmiennymi niezależnymi, wprowadzone zostały sztucznie wygenerowane braki typu Missing Completly At Radom, Missing At Random oraz Not Missing At Random w dwóch wariantach. Braki imputowane zostały następującymi metodami: podstawianie średniej, k najbliższych sąsiadów oraz imputacją wielokrotną. Następnie dane wykorzystane zostały do oszacowania modeli logit, na podstawie których oceniono jakość imputacji pod względem ilości istotnych parametrów modelu oraz miar bazujących na macierzy pomyłek (trafność, precyzja, czułość). Etapy generowania braków ich imputacji oraz estymacji modeli wraz z wyznaczeniem macierzy pomyłek dokonano przy pomocy języka programowania R oraz odpowiednich pakietów (glm2, Amelia, Zelig, ModelMetrics oraz xlsx).

Generowanie braków

Braki wprowadzone zostały w dwóch wariantach A i B. W wersji A wygenerowane zostały braki typu MCAR na zmiennych wiek, kwota, staż oraz okres, z prawdopodobieństwem ich wystąpienia równym 20%. Wprowadzono także braki MAR na zmiennej dochód, które zależały od wieku tzn. obserwacje z wiekiem co najmniej 40 lat miały 50% szans na wystąpienie braku (łączny udział pustych wartości tej zmiennej wyniósł ok. 20%). Taki sam typ braków wygenerowano na zmiennej rata i zależały one od zmiennej staż tzn. obserwacje ze stażem mniejszym lub równym 13 miały 70% szans na wystąpienie braku, natomiast powyżej tylko 10%. Sposób wprowadzenia braków zobaczyć można w poniższym fragmencie kodu, zgodnie z nim utworzone zostały:

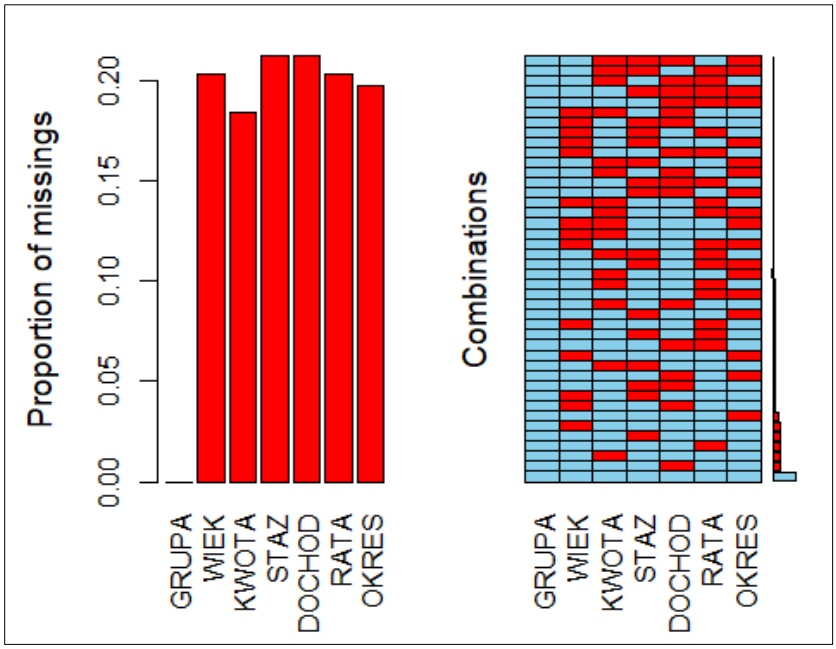
- macierz R1 – braków typu MCAR dla czterech zmiennych, które wylosowane zostały z wektora c(1,NA) z prawdopodobieństwami kolejno 0,8 i 0,2;

- macierz R2 – braków typu MAR, wylosowanych z wektora c(1,0), prawdopodobieństwo braków dla piątej zmiennej 0,5 (przy wieku >=40), dla szóstej 0,7 (przy stażu<=13) oraz 0,2 (przy stażu >13);
- macierz R_a – macierz braków w zbiorze A, numery jej kolumn odpowiadają kolumnom macierzy danych kompletnych, 1 oznacza obecność wartości, NA jej brak. Braki MCAR dodane zostały do niej na samym początku, natomiast MAR w oparciu o warunki – poprzez funkcję which, która wskazała w jakich wierszach powinien pojawić się brak – spełniona zależność od wartości innej zmiennej oraz wylosowane 0.

Utworzoną macierz braków pomnożono razy ramkę danych kompletnych, by wprowadzić do niej braki i zapisać jako zbiór A (działanie na odpowiadających sobie komórkach tabel, nie mnożenie macierzowe).

```
## Zbiór A ##
set.seed(2943)
R1 = matrix(sample(c(1,NA), wiersze*4, prob = c(0.8, 0.2), replace = TRUE), nrow=wiersze)

R2 = matrix(sample(c(1,0), wiersze, prob = c(0.5, 0.5), replace = TRUE), nrow=wiersze)
R2 = cbind(R2, sample(c(1,0), wiersze, prob = c(0.3, 0.7), replace = TRUE))
R2 = cbind(R2, sample(c(1,0), wiersze, prob = c(0.8, 0.2), replace = TRUE))
R_a = cbind(1,R1[,1:3],1,1,R1[,4])
R_a[which((dane_kompletne$WIEK>=40)&(R2[,1]==0)),5]=NA
R_a[which(((dane_kompletne$STAZ<=13)&(R2[,2]==0))|((dane_kompletne$STAZ>13)&(R2[,3
]==0))),6]=NA

zbior_A = dane_kompletne*R_a
```

Rysunek 1. Wizualizacje rozkładu braków na zmiennych w zbiorze A



Źródło: opracowanie własne w R

Braki w zbiorze A prezentują powyższe wizualizacje, na lewej zauważyć można, że na wszystkich zmiennych braki stanowiły ok. 20%, natomiast na prawej wyczytać można jak często braki pojawiały się pojedynczo (jednowymiarowy wzorzec braków) lub kilku zmiennych na raz, maksymalnie były to 4 zmienne (wielowymiarowy wzorzec).

W wariancie B wygenerowane zostały wyłącznie braki NMAR na trzech zmiennych:

- Wiek – prawdopodobieństwo braku = 70% dla osób z wiekiem co najwyżej 23 oraz 10% powyżej tej wartości.
- Dochód – prawdopodobieństwo braku = 70% dla obserwacji z dochodem wyższym niż 3200 oraz 10% z niższym lub równym 3200.
- Okres – prawdopodobieństwo braku = 70% dla obserwacji z okresem mniejszym niż 14 oraz 10% z równym lub wyższym niż 14.

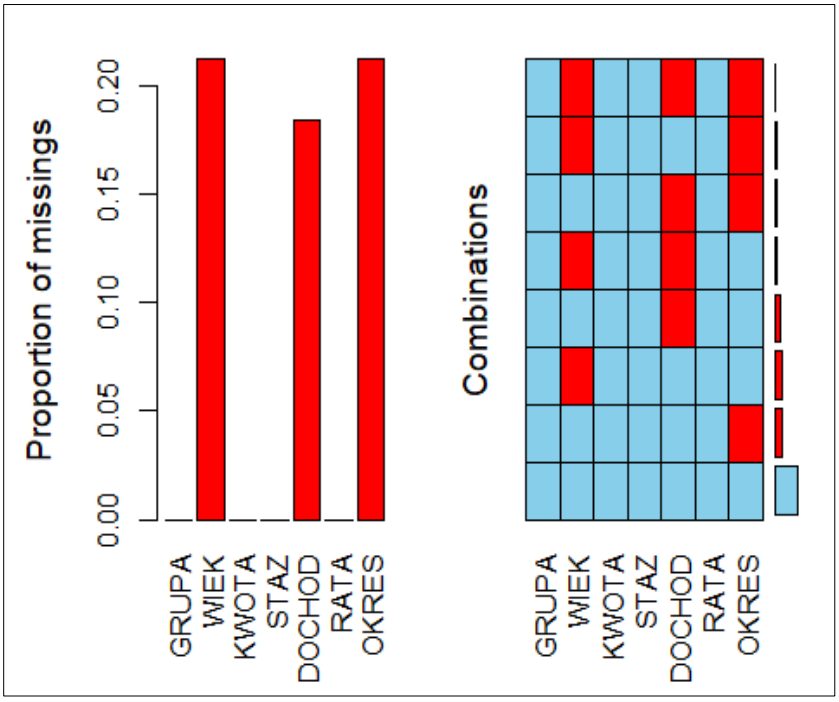
Wprowadzenie braków w zbiorze B przebiegało podobnie do wariantu A i wykonane zostało przy pomocy poniższego kodu:

```
## Zbiór B ##

set.seed(4912)
R3 = matrix(sample(c(1,0), wiersze*3, prob = c(0.3, 0.7), replace = TRUE), nrow=wiersze)
R3 = cbind(R3, matrix(sample(c(1,0), wiersze*3, prob = c(0.9, 0.1), replace = TRUE),
nrow=wiersze))
```

```
R_b = matrix(1, nrow=wiersze, ncol=kolumny)
R_b[which(((dane_kompletne$WIEK<=23)&(R3[,1]==0))|((dane_kompletne$WIEK>23)&(R3[,4]==0))),2]=NA
R_b[which(((dane_kompletne$DOCHOD>3200)&(R3[,2]==0))|((dane_kompletne$DOCHOD<=3200)&(R3[,5]==0))),5]=NA
R_b[which(((dane_kompletne$OKRES<14)&(R3[,3]==0))|((dane_kompletne$OKRES>=14)&(R3[,4]==6))),7]=NA

zbior_B = dane_kompletne*R_b
```

Jak zauważyć można na poniższych wykresach braki stanowiły ponownie ok 20%, jednak w tym wypadku maksymalna liczba braków na jednej obserwacji wynieść mogła 3 i wystąpiły takie przypadki.

Rysunek 2. Wizualizacje rozkładu braków na zmiennych w zbiorze B



Źródło: opracowanie własne w R

Przygotowane zbiory zapisane zostały jako pliki xlsx do wglądu.

Imputacja braków

Po wprowadzeniu braków i utworzeniu dwóch zbiorów przystąpiono do imputacji braków trzema wybranymi metodami. Pierwszą z nich było podstawianie w miejsce braku średniej

wartości zmiennej (wyznaczanej na podstawie dostępnych przypadków). Jest to jedna z najprostszych metod, jednak powoduje ona spłaszczenie zbioru oraz zmniejszenie zróżnicowania, gdyż pojawiają się obserwacje skoncentrowane wokół jednej wartości.

Jako drugą zastosowano metodę k najbliższych sąsiadów, która imputuje braki bazując na podobieństwie obserwacji. Imputowane wartości są podobne do tych jakie mają najbliższe przypadki. Zaletą tej metody jest jej działanie nawet w przypadku wielowymiarowego wzorca braków.

Ostatnią metodą była imputacja wielokrotna, polegająca na przeprowadzeniu imputacji m-krotnie np. w tym projekcie 5 razy, a następnie ich kombinacji w celu znalezienia optymalnego rozwiązania. Zakłada się przy tym wielowymiarowy rozkład normlany zmiennych. W tym wypadku zastosowano funkcję z pakietu Amelia, która bazuje na algorytmie maksymalizacji oczekiwań i metodzie bootstrap. Wynikiem tej imputacji jest m zbiorów, które bazują na rozkładach zmiennych ich poprzedników, a algorytm dąży do maksymalizacji logarytmu wiarygodności. Dane są imputowane do momentu uzyskania zbieżności, polegającej na zbliżaniu się imputowanych wartości do określonego optimum i przerwaniu w momencie minimalnej zmiany w stosunku do poprzedniej iteracji. Jako ograniczenie wartości imputowanych zastosowano minimum i maksimum pozostałych przypadków, dla których obserwowalne były wartości danej zmiennej. Zastosowanie wielokrotnej imputacji (MI – Multiple Imputation) zagwarantowało ujęcie niepewności z danych w procesie imputacji, kosztem tego, że metoda ta jest czarną skrzynką i nie da się zobaczyć jak przebiegał proces.

Ocena imputacji

Dane kompletne, wybrakowane oraz imputowane wykorzystane zostały do estymacji modelu regresji logistycznej w celu zbadania jakości imputacji. Została ona zmierzona ilością istotnych statystycznie parametrów modelu oraz miar wyznaczanych na podstawie macierzy pomyłek. Wyniki estymacji przedstawione zostały poniżej.

```
## dane kompletne ##
##              Estimate    Std. Error    z value  Pr(>|z|)
## (Intercept) -1.59e-01   5.97e-01   -0.27  0.78982
## WIEK        -1.88e-02   1.10e-02   -1.72  0.08552 .
## KWOTA       -4.46e-05   1.82e-04   -0.25  0.80598
## STAZ        -5.07e-03   1.74e-03   -2.91  0.00361 **
```

```
## DOCHOD -5.90e-04   1.73e-04   -3.41  0.00065 ***
## RATA        3.92e-03   2.57e-03    1.52  0.12821
## OKRES       5.68e-02   2.20e-02    2.59  0.00968 **
```

dane z brakami - Zbiór A

```
##             Estimate    Std. Error  z value  Pr(>|z|)
## (Intercept) -1.08e+00   1.18e+00   -0.92    0.36
## WIEK         5.40e-03   2.18e-02    0.25    0.80
## KWOTA       -8.39e-05   3.67e-04   -0.23    0.82
## STAZ        -6.23e-03   4.36e-03   -1.43    0.15
## DOCHOD -5.13e-04   4.19e-04   -1.23    0.22
## RATA         4.71e-03   4.67e-03    1.01    0.31
## OKRES        6.65e-02   4.50e-02    1.48    0.14
```

 # dane z brakami - Zbiór B ##

```
##              Estimate   Std. Error    z value Pr(>|z|)
## (Intercept)  0.198073   1.068066    0.19    0.853
## WIEK        -0.005419   0.014718   -0.37    0.713
## KWOTA    0.000364   0.000372    0.98    0.327
## STAZ        -0.002519   0.002155   -1.17    0.242
## DOCHOD -0.000582   0.000254   -2.29    0.022 *
## RATA        -0.003394   0.006041   -0.56    0.574
## OKRES        0.018050   0.037748    0.48    0.633
```

1. metoda - średnia - Zbiór A

```
##              Estimate     Std. Error  z value  Pr(>|z|)
## (Intercept)  0.915183   0.597633    1.53   0.1257
## WIEK        -0.026378   0.011909   -2.21   0.0268 *
## KWOTA        0.000297   0.000112    2.65   0.0081 **
## STAZ        -0.005527   0.001954   -2.83   0.0047 **
## DOCHOD -0.000575   0.000200   -2.87   0.0041 **
## RATA        -0.001693   0.001713   -0.99   0.3230
## OKRES        0.025163   0.015521    1.62   0.1050
```

1. metoda - średnia - Zbiór B

```
##              Estimate     Std. Error z value  Pr(>|z|)
## (Intercept) -3.62e-01   6.87e-01   -0.53  0.59827
## WIEK        -3.94e-03   1.25e-02   -0.31  0.75279
## KWOTA        2.10e-04   1.22e-04    1.73  0.08425 .
## STAZ        -5.75e-03   1.70e-03   -3.39  0.00071 ***
## DOCHOD -7.30e-04   2.09e-04   -3.48  0.00049 ***
## RATA         6.51e-05   1.78e-03    0.04  0.97083
## OKRES        4.50e-02   1.96e-02    2.29  0.02182 *
```

2. metoda - kNN - Zbiór A

```
##                Estimate    Std. Error  z value Pr(>|z|)
## (Intercept)  0.459357   0.578278    0.79   0.4270
## WIEK        -0.023231   0.011755   -1.98   0.0481 *
## KWOTA        0.000138   0.000138    1.00   0.3178
## STAZ        -0.005483   0.001966   -2.79   0.0053 **
## DOCHOD -0.000918   0.000207   -4.43  9.4e-06 ***
## RATA         0.001441   0.002030    0.71   0.4778
## OKRES        0.052063   0.017445    2.98   0.0028 **

## 2. metoda - kNN - Zbiór B ##
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -8.01e-01   6.33e-01   -1.27  0.20563
## WIEK         -4.23e-03   1.19e-02   -0.36  0.72142
## KWOTA         6.33e-05   1.31e-04    0.48  0.62953
## STAZ         -5.73e-03   1.75e-03   -3.27  0.00106 **
## DOCHOD -9.28e-04   2.16e-04   -4.31  1.7e-05 ***
## RATA          2.16e-03   1.86e-03    1.16  0.24610
## OKRES         7.43e-02   1.98e-02    3.74  0.00018 ***

## 3. metoda - Imputacja wielowymiarowa - Zbiór A ##
##                Value          Std. Error     t-stat    p-value
## (Intercept)   4.950e-01  0.9160201  0.5404  0.596988
## WIEK         -2.348e-02  0.0134038 -1.7517  0.086388*
## KWOTA         9.225e-05  0.0003095  0.2981  0.772470
## STAZ         -5.310e-03  0.0022886 -2.3204  0.027675**
## DOCHOD -6.722e-04  0.0002315 -2.9037  0.006252***
## RATA          1.733e-03  0.0043802  0.3956  0.701822
## OKRES         4.283e-02  0.0354233  1.2092  0.254949

## 3. metoda - Imputacja wielowymiarowa - Zbiór B ##
##                Value          Std. Error     t-stat      p-value
## (Intercept)  -8.018e-01  0.7949127 -1.0087  0.3163956
## WIEK         -8.013e-03  0.0150744 -0.5316  0.6002816
## KWOTA        -9.525e-05  0.0002033 -0.4686  0.6404228
## STAZ         -5.466e-03  0.0017893 -3.0549  0.0023338***
## DOCHOD  -7.129e-04  0.0002063 -3.4550  0.0006159***
## RATA          4.892e-03  0.0030642  1.5966   0.1147903
## OKRES         7.227e-02  0.0284614  2.5393   0.0143985**
```

Jak zauważyć można w powyższych wynikach estymacji model na danych kompletnych posiadał 3 istotne statystycznie parametry, wprowadzenie braków spowodowało spadek ich ilości nawet do 0 w przypadku zbioru A. Każda z metod imputacji spowodowała, że liczba istotnych parametrów była taka sama lub wyższa niż na danych kompletnych. Podstawianie średniej i k najbliższych sąsiadów pozwoliło osiągnąć nawet 4 istotne parametry. Należy zwrócić uwagę, że

w przypadku zbioru A po imputacji istotny stawał się parametr stojący przy zmiennej WIEK, czym nie charakteryzował się model na danych kompletnych. W wariancie A kNN jako jedyna pozwoliła uzyskać istotność parametrów takich jak w modelu na danych kompletnych, co dla zbioru B osiągnęły wszystkie z metod. Podsumowanie znaleźć można w poniższej tabeli – w kolumnach wskazano zbiór danych natomiast w wierszach brak imputacji oraz jej różne metody.

Istotne parametry	Kompletne	Zbiór A	Zbiór B
brak	3	0	1
średnia	x	4	3
kNN	x	4	3
MI	x	3	3

W celu dalszej analizy dla każdego modelu opracowano macierz pomyłek na podstawie wartości zmiennej zależnej obserwowalnych (empirycznych) oraz przewidzianych (zaklasyfikowanych) przez model. Obliczone zostały trzy miary: trafność (accuracy) – jako iloraz poprawnie zaklasyfikowanych i wszystkich obserwacji (TP+TN) / (TP+FP+FN+TN), precyzję (precision) – jako iloraz poprawnie zaklasyfikowanych wartości 1 i wszystkich zaklasyfikowanych 1 (TP) / (TP+FP) oraz czułość (recall/sensitivity) – jako iloraz poprawnie zaklasyfikowanych i wszystkich wartości 1 (TP) / (TP+FN).

		Zaklasyfikowany	
		1	0
Obserwowalny	1	TP	FN
	0	FP	TN

W poniższych tabelach podsumowujących pokazano wartości miar w zaokrągleniu do trzech miejsc po przecinku, pogrubione zostały trzy wartości najbardziej zbliżone do statystyk dla danych kompletnych. W przypadku zbioru A najczęściej metoda podstawiania średniej dawała podobne wyniki, natomiast dla zbioru B była to imputacja wielokrotna, a dalej k najbliższych sąsiadów. Sugeruje to, że najlepszą metodą jest podstawianie średnich a następnie imputacja wielokrotna.

Trafność	Kompletne	Zbiór A	Zbiór B
brak	0,649	0,651	0,584
średnia	x	**0,653**	0,625
kNN	x	0,677	**0,662**
MI	x	0,665	**0,649**

Precyzja	Kompletne	Zbiór A	Zbiór B
brak	0,655	0,633	0,581
średnia	x	**0,660**	**0,633**
kNN	x	0,688	0,672
MI	x	0,674	**0,654**

Czułość	Kompletne	Zbiór A	Zbiór B
brak	0,655	0,608	0,540
średnia	x	**0,660**	0,630
kNN	x	0,668	**0,655**
MI	x	0,662	**0,660**

Wnioski

Dane poddane zostały generowaniu braków a następnie ich imputacji trzema wybranymi metodami. Ich porównanie dokonane zostało na podstawie wyników estymacji modelu regresji logistycznej. Określenie najlepszej metody imputacji dokonać należy przez pryzmat wprowadzonych braków, metody zastosowane na zbiorze B dały obciążone wyniki, gdyż zawierał on braki typu NMAR, dlatego należy je traktować z ostrożnością. Ilość istotnych parametrów modelu wskazuje, że najpewniejsza jest metoda k najbliższych sąsiadów, ponieważ w obu zbiorach istotne okazały się być parametry takie jak przy danych kompletnych. Z kolei patrząc na miary obliczone na podstawie macierzy pomyłek za lepszą można metodę podstawiania średniej. Jednak mimo, że pod tym względem wyniki kNN różniły się bardziej to modele oszacowane na danych nią imputowanych mają lepsze zdolności predykcyjne (wyższe wartości statystyk), dlatego ją należy uznać za najlepszą metodę. Kolejny wniosek jest taki, że imputacja wielokrotna, chociaż jest bardziej zaawansowaną metodą, nie jest najlepsza, a jako dalszy kierunek badania jakości metod imputacji należałoby przeprowadzić tę analizę w wielu iteracjach/symulacjach, ponieważ powyższe było tylko jedną z nich i wnioski mogłyby się różnić, gdyby np. braki wygenerowane zostały na podstawie innego ziarna.