PRACE NAUKOWE AKADEMII EKONOMICZNEJ WE WROCŁAWIU
Nr 1107 — 2006
Badania marketingowe w przestrzeni europejskiej

Sylwester Białowąs
Akademia Ekonomiczna w Poznaniu

METODA REGRESJI W SZACOWANIU BRAKÓW DANYCH

1. Wstęp

Zajmując się badaniami marketingowymi, nie sposób ominąć tematu brakujących danych. Każdy z badaczy zdaje sobie sprawę, że niewielki odsetek brakujących danych może w określonych warunkach prowadzić do poważnych problemów z przetworzeniem danych, a co gorsza, do błędnych wniosków. Szczególnie gdy braki danych nie są rozłożone losowo, a zależą od innej zmiennej lub zespołu zmiennych. Nie sposób jednak przed analizą danych ocenić, czy występujące w pliku braki danych obciążą wyniki (rozkład braków danych nie jest losowy), czy też nie (rozkład braków danych jest losowy). Wprawdzie problem braków danych poruszany jest w literaturze w niemal każdym opracowaniu dotyczącym analizy danych, powstało także wiele opracowań na temat postępowania z brakami danych. Nie istnieją jednak sposoby idealne, które wyeliminowałyby ten problem.

Sytuację najlepiej ujmują W.G. Cochran i G.M. Cox, twierdząc, że „jedynym w pełni satysfakcjonującym rozwiązaniem w zakresie braków danych jest ich nie mieć..." [1, s. 82].

2. Sposoby imputacji

Brakujące dane mogą w dużym stopniu zmieniać rezultaty analiz. Wyłączenie z analizy braków danych często prowadzi do uzyskania wyników nieprawdziwych lub nieistotnych statystycznie. Z brakującymi danymi można postępować dwojako.

Najprostszym sposobem jest oczyszczenie pliku ze wszystkich przypadków (lub zmiennych), w których występują braki danych. Sposób taki prowadzi jednak bardzo często do nieakceptowalnego zmniejszenia liczby przypadków, ponieważ nawet jeśli braków jest stosunkowo niewiele w każdej zmiennej, to łącznie mogą one występować w dużej części przypadków. Z kolei w przypadku wyłączania

zmiennych może okazać się, że tracimy informacje ważne dla analiz. Można także przyjmować do poszczególnych analiz wszystkie przypadki bez braków danych w zmiennych, które biorą udział w analizie, w ten sposób liczba przypadków będzie większa, jednak każda analiza będzie dotyczyła nieco innej próby.

Lepszym sposobem jest uzupełnienie brakujących danych na podstawie posiadanych już informacji. Wyróżnić można kilka podstawowych metod: uzupełnienie na podstawie wartości średnich obliczonych dla ważnych obserwacji danej zmiennej, uzupełnienie na podstawie wyników analizy regresji z wykorzystaniem innych zmiennych oraz uzupełnienie na podstawie największego prawdopodobieństwa (algorytmy EM) z wykorzystaniem innych zmiennych [3, s. 28].

W dalszej części artykułu porównane zostaną dwie metody nie wymagające wysublimowanego warsztatu statystycznego – podstawienie średnich i imputacja na podstawie analizy regresji.

Zanim badacz rozpocznie jakąkolwiek procedurę statystyczną (w tym także procedurę uzupełniania braków danych), musi wyjaśnić, czy braki nie są związane z określoną regułą. Jeśli bowiem okaże się, że obiekty generujące braki danych są inne niż reszta obiektów poddanych badaniu, wyniki mogą być obarczone poważnym błędem [4, s. 410].

3. Imputacja danych – studium przypadku

W poniższym przykładzie wykorzystano plik z badań dotyczących korzystania z usług finansowych. Do prezentacji wybrano zmienną *LUSLGB* – tę liczbę wykorzystywanych usług bankowych, z której na potrzeby analizy usunięto część wartości dotyczących liczby usług bankowych (w ten sposób sztucznie tworząc braki danych). Pomocniczo na potrzeby analizy regresji dobrano trzy zmienne egzogeniczne, których wyboru dokonano na podstawie stwierdzonych wcześniej zależności między zmienną do imputacji i pozostałymi zmiennymi. Są to:

- EDU – wykształcenie (mierzone liczbą lat nauki),
- INC – średni miesięczny dochód netto,
- LZU – liczba zakładów ubezpieczeń, z których korzysta respondent.

Rozkład częstości dla zmiennej liczba usług bankowych przedstawia tab. 1.

Średnia imputowanej zmiennej wynosi 3,4 usługi, a wariancja 4,95. Spośród miar pozycyjnych warto odnotować, że mediana wynosi 3, a dominanta 2.

W pierwszym kroku wszystkie braki danych w analizowanej zmiennej zostały zastąpione wartością średnią. W ten sposób wyeliminowano braki danych, a wartość średnia nie uległa zmianie. Zastąpienie braków średnią spowodowało jednak znaczne zmniejszenie wariancji (do wartości 3,93). Kolejną niedogodnością jest pojawienie się sztucznej wartości w analizie częstości – nowo powstała wartość 3,4 przypisywana poszczególnym klientom nie może istnieć w rzeczywistości (klient nie może przecież korzystać z 3,4 usługi). Warto zauważyć także, że wartość 3,4 jest w tym przypadku także medianą i dominantą.

Tabela 1. Rozkład częstości dla zmiennej liczba usług bankowych – dane obciążone brakami

Wartości		Częstość	Procent	Procent ważnych	Procent skumulowanych
Ważne	0	14	2,2	2,8	2,8
	1	77	12,0	15,2	17,9
	2	130	20,3	25,6	43,5
	3	92	14,4	18,1	61,6
	4	62	9,7	12,2	73,8
	5	42	6,6	8,3	82,1
	6	44	6,9	8,7	90,7
	7	18	2,8	3,5	94,3
	8	11	1,7	2,2	96,5
	9	6	0,9	1,2	97,6
	10	11	1,7	2,2	99,9
	12	1	0,2	0,2	100,0
Ogółem		508	79,4	100,0	
Braki danych	Systemowe braki danych	132	20,6		
Ogółem		640	100		

Źródło: opracowanie własne.

Nowy rozkład częstości przedstawia tab. 2.

Tabela 2. Rozkład częstości dla zmiennej liczba usług bankowych – imputacja średnią

Wartości		Częstość	Procent	Procent ważnych	Procent skumulowanych
Ważne	0	14	2,2	2,2	2,2
	1	77	12,0	12,0	14,2
	2	130	20,3	20,3	34,5
	3	92	14,4	14,4	48,9
	3,4	132	20,6	20,6	69,5
	4	62	9,7	9,7	79,2
	5	42	6,6	6,6	85,8
	6	44	6,9	6,9	92,7
	7	18	2,8	2,8	95,5
	8	11	1,7	1,7	97,2
	9	6	0,9	0,9	98,1
	10	11	1,7	1,7	99,8
	12	1	0,2	0,2	100,0
Ogółem		640	100,04	100,0	

Źródło: opracowanie własne.

Alternatywą takiego rozwiązania jest, jak wcześniej wspomniano, podstawienie szacunkowych danych otrzymanych w analizie regresji. Na podstawie analizy logicznej i po statystycznym potwierdzeniu istniejących zależności do szacowania liczby usług bankowych przyjęto następujące zmienne egzogeniczne:

- średni miesięczny dochód netto *INC* – poziom pomiaru ilorazowy, jednostka 1 tys. zł,
- poziom wykształcenia *EDU* – poziom pomiaru ilorazowy, jednostka – rok nauki,
- liczba zakładów ubezpieczeń, z których korzysta respondent *LZU* – poziom pomiaru ilorazowy, jednostki naturalne.

Otrzymano następujące równanie regresji.

$$LUSLGB = -0{,}400 + 0{,}709 \cdot LZU + 0{,}161 \cdot INC + 0{,}129 \cdot EDU$$

Równanie w całości oraz wszystkie parametry są istotne statystycznie, współczynnik determinacji R^2 wynosi 0,269. Na podstawie uzyskanego równania uzupełniono braki danych dla zmiennej liczba usług bankowych. Wartość średnia zmiennej po imputacji wynosi 3,30, a wariancja 4,15. Mediana i dominanta, tak jak w poprzedniej imputacji, wynoszą odpowiednio 3 i 2.

Tabela 3. Porównanie statystyk w różnych sposobach imputacji

		Dane oryginalne	Dane z brakami	Braki zastąpione średnią	Braki zastąpione średnią i zaokrąglone	Braki uzupełnione regresją	Braki uzupełnione regresją i zaokrąglone
N	Ważne	640	508	640	640	640	640
	Braki danych	0	132	0	0	0	0
Średnia		3,22	3,40	3,40	3,39	3,30	3,29
Mediana		3	3	3,40	3	3	3
Dominanta		2	2	3,40	3	2	2
Odchylenie standardowe		2,09	2,22	1,98	1,95	2,04	2,04
Wariancja		4,38	4,95	3,93	3,79	4,15	4,17
Skośność		1,16	1,04	1,17	1,39	1,19	1,19
Kurtoza		1,37	0,85	1,85	2,14	1,59	1,58
Rozstęp		12	12	12	11	12	12
Minimum		0	0	0	1	0	0
Maksimum		12	12	12	12	12	12
Percentyle	10	1	1	1	1	1	1
	20	2	2	2	2	2	2
	30	2	2	2	2	2	2
	40	2	2	3	3	2	2
	50	3	3	3,4	3	3	3
	60	3	3	3,4	3	3	3
	70	4	4	4	4	4	4
	80	5	5	5	5	5	5
	90	6	6	6	6	6	6

Źródło: opracowanie własne.

Metoda imputacji poprzez regresję także generuje sztuczne wartości (liczby inne niż całkowite), na potrzeby analiz wymagających liczb całkowitych (np. analizy

przypadków) można uzyskane wyniki zaokrąglić. Uzyskane w ten sposób statystyki zmieniają się tylko w niewielkim stopniu – średnia wynosi 3,29, wariancja 4,17, mediana 3, a dominanta 2.

Podsumowując wszystkie imputacje, ich wyniki przedstawiono wraz z porównaniem z danymi oryginalnymi, zawierającymi komplet informacji.

Z tab. 3 wynika, że braki danych nie były rozproszone losowo. Więcej braków wystąpiło wśród respondentów korzystających z mniejszej liczby usług bankowych, średnia uzyskana ze wszystkich obserwacji jest o 0,18 usługi większa od średniej uzyskanej z pliku obciążonego brakami danych.

W analizowanym zbiorze wystąpiła ciekawa sytuacja, w której średnia szacowana na podstawie pliku z imputowanymi średnią wartościami statystycznie istotnie różni się od średniej oryginalnej, a średnia wyliczona ze zmiennej obciążonej brakami danych nie jest statystycznie istotnie różna. Średnie te są oczywiście identyczne, jednak zmieniona liczebność próby zdecydowała o różnych wynikach testu.

Tabela 4. Test średniej dla średnich uzyskanych różnymi metodami imputacji

Test dla jednej próby	Wartość testowana = 3,22				95% przedział ufności dla różnicy średnich	
	t	df	istotność dwustronna	różnica średnich	dolna granica	górna granica
Dane z brakami	1,780	507	0,076	0,176	- 0,018	0,370
Braki zastąpione średnią	2,243	639	0,025	0,176	0,022	0,329
Braki zastąpione regresją	0,044	639	0,965	0,002	- 0,082	0,086
Dane oryginalne	0,023	639	0,982	0,002	- 0,161	0,164

Źródło: opracowanie własne.

4. Podsumowanie

Zdecydowanie bliższa prawdziwej wartości jest średnia oszacowana ze zbioru, gdzie braki były imputowane metodą regresji, różnica w tym przypadku wynosi 0,08 i nie jest istotna statystycznie. Również porównanie wariancji przemawia na korzyść metody imputacji przez regresję. W zakresie miar pozycyjnych nie zaobserwowano różnic, jedynie wartość modalna dla danych uzyskanych z podstawienia średnią różni się od danych oryginalnych.

Należy także zwrócić uwagę, że oprócz imputacji braków warto analizować zależności występujące między zmiennymi w zakresie braków danych, np. poprzez stworzenie krzyżowej tabeli występowania braków danych. Pozwala to wychwycić słabości narzędzia i wywiadu. Szybko i kompleksowo (dla wielu zmiennych) można takie analizy przeprowadzać za pomocą narzędzi dostępnych w programach do analizy danych, np. w module Missing Value pakietu SPSS. Należy także zawsze pamiętać o jednoznacznym oznaczeniu braków danych (np. dokonywać imputacji

na nowej zmiennej), tak aby dane imputowane nie były traktowane jako oryginalne i aby zawsze można było wrócić do pierwotnej struktury.

Warto także pamiętać, że w niektórych badaniach brak danych może mieć znaczenie równe innym wartościom zmiennej.

Literatura

[1] Cochran W.G., Cox G.M., *Experimantal Designs*, J. Wiley, New York 1957.
[2] *Missing Data: The Hidden Problem*, SPSS White Paper, SPSS Illinois.
[3] Rubin D.B., Little R.J.A., *Statistical Analysis with Missing Data*, Wiley Series in Probability and Mathematical Statistics, John Wiley & Sons, New York 1987.
[4] Sheskin D.J., *The Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman&Hall CRC.

REGRESSION METHOD IN THE ASSESSMENT OF THE LOCK OF DATA

Summary

The text describes a problem of missing values in the data sets used for statistical analysis. Consequences of working with data sets with missing values as well as some of the most popular methods of data imputation are presented together with an example comparing two simplest methods of imputation of missing values – the average vs. the linear regression. In line with the theory, better results were achieved using the linear regression for data imputation.