

Mirosława Sztemberg-Lewandowska

WYKORZYSTANIE GŁÓWNYCH SKŁADOWYCH W ANALIZIE REGRESJI WIELORAKIEJ

1. Wstęp

Liniowa regresja wieloraka jest najczęściej stosowaną metodą statystycznej analizy wielowymiarowej. Głównym problemem pojawiającym się przy zastosowaniu tej metody jest współliniowość zmiennych objaśniających, która występuje, gdy kilka regresorów jest w zależności liniowej. O współliniowości często informują wysokie korelacje między niektórymi zmiennymi objaśniającymi.

Współliniowość w analizie regresji jest przyczyną wysokich wariancji estymowanych współczynników regresji, co powoduje, że estymowane równanie regresji jest niestabilne i błędne. W celu przezwyciężenia tego problemu proponowane są różne rozwiązania. Jedną z możliwości jest wykorzystanie w analizie regresji głównych składowych, które są ortogonalne, a zatem niewspółliniowe.

2. Główne składowe w analizie regresji

Standardowy model regresji liniowej (por. np. [1; 8; 9]) ma postać:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad (1)$$

gdzie: $\mathbf{Y}$ – wektor n obserwacji zmiennej zależnej,

$\mathbf{X}$ – macierz $(n \times p)$, której (i, j) -ty element jest wartością j -tego predyktora dla i -tej obserwacji,

$\boldsymbol{\beta}$ – wektor $(p \times 1)$ współczynników regresji,

$\boldsymbol{\varepsilon}$ – wektor $(n \times 1)$ składników losowych.

Wszystkie zmienne są wyrażone za pomocą ich średnich oraz elementy $\boldsymbol{\varepsilon}$ są niezależne i mają tę samą wariancję. Ponadto założono, że zmienne objaśniające są standaryzowane, a zatem $\mathbf{X}'\mathbf{X}$ jest proporcjonalne do macierzy korelacji predyktorów.

Główne składowe (por. np. [2; 6; 7]) wyznacza się na podstawie wzoru:

$$\mathbf{Z} = \mathbf{XA} \quad (2)$$

gdzie: $\mathbf{Z}$ – macierz ($n \times p$), której (i, k)-ty element jest wartością k -tej składowej dla i -tej zmiennej,

$\mathbf{A}$ – macierz ($p \times p$), której k -ta kolumna jest k -tym wektorem własnym macierzy $\mathbf{X}'\mathbf{X}$.

Macierz $\mathbf{A}$ jest ortogonalna ($\mathbf{AA}' = \mathbf{I}$) (por. np. [5]), zatem: $\mathbf{X}\beta = \mathbf{XAA}'\beta = \mathbf{Z}\gamma$, gdzie $\gamma = \mathbf{A}'\beta$. W związku z tym równanie (1) można przekształcić następująco:

$$\mathbf{Y} = \mathbf{Z}\gamma + \epsilon. \quad (3)$$

We wzorze (3) predyktory zostały zastąpione głównymi składowymi, w przypadku gdy w analizie wykorzystuje się m ($m < p$) głównych składowych model ten redukuje się do postaci:

$$\mathbf{Y} = \mathbf{Z}_m\gamma_m + \epsilon_m, \quad (4)$$

gdzie: γ_m – wektor ($m \times 1$), którego elementy tworzą podzbiór elementów γ ,

$\mathbf{Z}_m$ – macierz ($n \times m$) zawierająca m kolumn macierzy $\mathbf{Z}$,

ϵ_m – wektor ($n \times 1$) składników losowych.

Wykorzystanie głównych składowych w analizie regresji ma wiele zalet. Po pierwsze estymowanie γ z równania (3) jest dużo prostsze niż β z (1). Po drugie łatwiej jest zinterpretować wkład głównej składowej w równaniu regresji niż wkład pierwotnej zmiennej. Ponadto wkład i współczynniki regresji głównej składowej nie zależą od tego, które składowe są uwzględnione w analizie. Natomiast w regresji z pierwotnymi zmiennymi wkład i współczynniki regresji mogą radykalnie ulec zmianie, gdy zmienna zostanie dodana lub usunięta z analizy.

3. Zastosowanie metody głównych składowych w analizie regresji

Przeprowadzono badanie ankietowe, którego celem było określenie, jakie zmienne mają decydujący wpływ na popularność danej czekolady. Przedmiotem badania było sześć polskich czekolad mlecznych – Goplana, Wedel, Terravita, Śnieżka, Pomorzanka, Wawel – i sześć zagranicznych czekolad mlecznych – Milka, Alpen Gold, Fazer, Lindt, Nestle, Cadbury. W badaniu wzięło udział 340 studentów Wydziału Gospodarki Regionalnej i Turystyki Akademii Ekonomicznej we Wrocławiu. Respondenci oceniali tylko znane im czekolady.

W tabeli 1 zamieszczono liczby respondentów oceniających poszczególne marki czekolad.

Tabela 1. Liczba ankiet wypełnionych przez respondentów przy ocenie poszczególnych marek czekolad

Nr czekolady	Nazwa czekolady mlecznej, polskiej	Liczba ankiet	Nr czekolady	Nazwa czekolady mlecznej, zagranicznej	Liczba ankiet
1	Goplana	314	7	Milka	332
2	Wedel	317	8	Alpen Gold	324
3	Terravita	275	9	Fazer	232
4	Śnieżka	233	10	Lindt	213
5	Pomorzanka	159	11	Nestle	260
6	Wawel	205	12	Cadbury	257

Źródło: opracowanie własne.

Czekolady porównywano ze względu na 8 zmiennych zaprezentowanych na skali semantycznej (siedmiostopniowej). Krańce skali określono w postaci antoniimów. Przykładową ocenę czekolady mlecznej Goplana podaną przez jednego z respondentów przedstawiono w tab. 2.

W tabeli 3 zamieszczono mediany ocen wszystkich czekolad, a w tab. 4 – postać macierzy korelacji między zmiennymi.

Tabela 2. Przykładowa ocena czekolady Goplana dokonana przez jednego respondenta

	1	2	3	4	5	6	7	
Popularna			X					Niepopularna
Tania			X					Droga
Niski poziom słodkości				X				Wysoki poziom słodkości
Szeroki asortyment smakowy (bakaliowe, nadziewane)				X				Wąski asortyment smakowy (bakaliowe, nadziewane)
Opakowanie estetyczne			X					Opakowanie nieestetyczne
Szeroko dostępna		X						Mało dostępna
Znana marka			X					Mało znana marka
Często promowana			X					Rzadko promowana

Źródło: opracowanie własne.

Tabela 3. Mediany ocen dwunastu czekolad dla ośmiu badanych zmiennych

Symbol zmiennej	Numer czekolady											
	1	2	3	4	5	6	7	8	9	10	11	12
<i>a</i>	3	1	3	4	7	4	1	1	3,5	4	2	2
<i>b</i>	3	4	3	2	3	3	5	4	5	7	4	5
<i>c</i>	4	5	4	4	4	4	6	5	5	5	5	5
<i>d</i>	4	2	4	5	5	4	1	2	3	3	3	3
<i>e</i>	3	2	3	4	4	4	1	2	2	2	2	2
<i>f</i>	2	1	3	4	6	4	1	2	3	4	3	3
<i>g</i>	3	1	4	4	6	5	1	2	3	4	3	3
<i>h</i>	3	2	3,5	6	7	6	1	2	4	6	3	3

Źródło: opracowanie własne na podstawie przeprowadzonego badania ankietowego.

Tabela 4. Macierz korelacji między zmiennymi

Symbol zmiennej	<i>a</i>	<i>b</i>	<i>c</i>	<i>d</i>	<i>e</i>	<i>f</i>	<i>g</i>	<i>h</i>
<i>a</i>	1,000	-0,252	-0,656	0,823	0,750	0,925	0,920	0,919
<i>b</i>	-0,252	1,000	0,725	-0,610	-0,744	-0,191	-0,318	-0,192
<i>c</i>	-0,656	0,725	1,000	-0,919	-0,934	-0,592	-0,752	-0,650
<i>d</i>	0,823	-0,610	-0,919	1,000	0,920	0,793	0,869	0,808
<i>e</i>	0,750	-0,744	-0,934	0,920	1,000	0,710	0,814	0,773
<i>f</i>	0,925	-0,191	-0,592	0,793	0,710	1,000	0,952	0,932
<i>g</i>	0,920	-0,318	-0,752	0,869	0,814	0,952	1,000	0,915
<i>h</i>	0,919	-0,192	-0,650	0,808	0,773	0,932	0,915	1,000

Źródło: obliczenia własne z wykorzystaniem pakietu SPSS.

Przed przystąpieniem do analizy obliczono wskaźniki KMO i MSA_h (por. np. [3; 10]). Na podstawie wskaźnika MSA_h wyeliminowano z analizy zmienną *b* (cenę). Współczynnik KMO bez tej zmiennej wynosi 0,82, czyli jest wysoki.

W modelu regresji liniowej zmienną zależną jest pierwsza zmienna – popularność, natomiast pozostałe zmienne są zmiennymi objaśniającymi. Na podstawie tab. 4 można zauważyć, że między zmiennymi występują wysokie korelacje, co wskazuje na współliniowość.

Metodą najmniejszych kwadratów przeprowadzono analizę regresji dla zmiennej objaśnianej (popularność) i zmiennych objaśniających: poziom słodkości, asortyment smakowy, estetyka opakowania, dostępność w sprzedaży, powszechność marki i promocje. Otrzymano następujące wyniki:

- współczynnik korelacji wielorakiej $R = 0,946$,
- kwadrat współczynnika korelacji wielorakiej $R^2 = 0,895$,
- dopasowany $R^2 = 0,77$,
- błąd standardowy estymacji = 0,834,
- wartość $F(6, 5) = 7,127$ przy poziomie istotności $\alpha = 0,024$.

Dokładność oszacowania funkcji regresji liniowej ocenia się za pomocą kwadratu współczynnika korelacji wielorakiej, który jest tutaj wysoki, podobnie jak dopasowany R^2 . Istotność statystyczną modelu testuje się za pomocą rozkładu Fishera-Snedecora; otrzymany model regresji jest istotny statystycznie na poziomie istotności 0,024.

W tabeli 5 zamieszczono współczynniki regresji, ich błędy standardowe oraz graniczne poziomy istotności dla rozkładu Studenta, które określają, czy poszczególne współczynniki są istotne statystycznie. Wszystkie wielkości w ostatniej kolumnie tab. 5 są większe od 0,05, a zatem wszystkie współczynniki są nieistotne statystycznie.

Tabela 5. Współczynniki regresji liniowej

Wyszczególnienie	Symbol zmiennej	Współczynniki	Błąd standardowy	Statystyka <i>t</i>	Graniczny poziom istotności
Wyraz wolny		-4,455	12,431	-0,358	0,735
Poziom słodkości	<i>c</i>	0,602	1,984	0,303	0,774
Asortyment smakowy	<i>d</i>	0,485	0,899	0,539	0,613
Estetyka opakowania	<i>e</i>	-0,004	0,911	-0,005	0,996
Dostępność	<i>f</i>	0,139	1,041	0,134	0,899
Popularność marki	<i>g</i>	0,432	0,939	0,460	0,665
Promocje	<i>h</i>	0,315	0,404	0,779	0,471

Źródło: obliczenia własne z wykorzystaniem pakietu SPSS.

Następnie zastosowano metodę głównych składowych w celu wyodrębnienia ortogonalnych składowych, które zostaną wykorzystane w analizie regresji.

W tabeli 6 zamieszczono wyniki metody głównych składowych z rotacją ortogonalną *varimax*. Na podstawie tych wyników można zauważyć, że zmienne mają istotne ładunki tylko dla dwóch pierwszych składowych, a pozostałe składowe wyjaśniają tylko 3,73% wariancji wszystkich zmiennych, dlatego zostaną uwzględnione tylko dwie pierwsze składowe.

Trzy ostatnie zmienne mają istotne ładunki dla pierwszego czynnika, natomiast trzy pierwsze zmienne mają istotne ładunki dla drugiego czynnika. Zatem otrzymano dwa czynniki, które wyjaśniają 96% zasobu zmienności wspólnej wszystkich zmiennych:

- pierwszy czynnik jest opisany przez: dostępność, powszechność marki, promocje, czynnik ten nazwano „powszechność czekolady”;
- drugi czynnik jest scharakteryzowany przez następujące zmienne: poziom słodkości, asortyment smakowy, estetyka opakowania, czynnik ten nazwano „właściwości smakowe”.

Tabela 6. Ładunki czynnikowe uzyskane po zastosowaniu metody głównych składowych z rotacją *varimax*

Nr składowej	1	2	3	4	5	6
Zmienne						
Poziom słodkości	-0,307	-0,948	-0,019	0,015	0,080	-0,026
Asortyment smakowy	0,551	0,793	0,025	0,258	-0,019	-0,004
Estetyka opakowania	0,454	0,856	0,086	-0,026	0,230	-0,007
Dostępność	0,938	0,328	-0,053	0,072	0,041	-0,055
Popularność marki	0,841	0,517	-0,064	-0,008	-0,015	0,145
Promocje	0,865	0,404	0,294	0,018	0,044	-0,011
Procent wyjaśnionej wariancji	85,28	10,99	1,73	1,06	0,74	0,20

Pogrubioną czcionką zaznaczono ładunki czynnikowe większe od 0,65, co do wartości bezwzględnej.

Źródło: obliczenia własne z wykorzystaniem pakietu SPSS.

Warto zauważyć, że pierwszy czynnik wyjaśnia 85%, a drugi tylko 11% wariancji, więc czynnik „właściwości smakowe” jest dużo mniej ważny przy opisie zależności pomiędzy zmiennymi.

Pierwotny zbiór badanych 12 czekolad w przestrzeni sześciowymiarowej został przekształcony w zbiór 12 czekolad w dwuwymiarowej przestrzeni niezależnych czynników głównych (por. tab. 7).

Następnie przeprowadzono analizę regresji wielorakiej, w której zmienną zależną jest popularność czekolady, natomiast zmiennymi objaśniającymi są główne składowe: powszechność czekolady i właściwości smakowe. Otrzymano następujące wyniki:

- współczynnik korelacji wielorakiej $R = 0,944$,
- kwadrat współczynnika korelacji wielorakiej $R^2 = 0,892$,
- dopasowany $R^2 = 0,868$,
- błąd standardowy estymacji = 0,633,
- wartość $F(2, 9) = 37,015$ przy poziomie istotności $\alpha = 0,000$.

Tabela 7. Współrzędne badanych czekolad w układzie dwóch czynników głównych

Nr czekolady	Czekolada	Czynnik 1	Czynnik 2
1	Goplana	-1,16	1,39
2	Wedel	-1,42	-0,10
3	Terravita	-0,41	1,03
4	Śnieżka	0,36	1,24
5	Pomorzanka	1,87	0,53
6	Wawel	0,65	0,90
7	Milka	-0,86	-1,68
8	Alpen Gold	-0,79	-0,40
9	Fazer	0,28	-0,65
10	Lindt	1,40	-1,18
11	Nestle	0,04	-0,54
12	Cadbury	0,04	-0,54

Źródło: obliczenia własne z wykorzystaniem pakietu SPSS.

Podane wielkości świadczą o dobrym dopasowaniu modelu do danych i są zbliżone do wyników otrzymanych na podstawie regresji dla zmiennych pierwotnych.

W tabeli 8 zamieszczono współczynniki regresji, ich błędy standardowe oraz graniczne poziomy istotności. Z ostatniej kolumny tej tabeli wynika, że wszystkie współczynniki są istotne statystycznie przy poziomie istotności 0,004.

Model wielorakiej regresji liniowej, w którym zmiennymi objaśniającymi są główne składowe, ma następującą postać:

$$Y = 2,958 + 1,463Z_1 + 0,743Z_2. \quad (5)$$

Tabela 8. Współczynniki regresji liniowej

Wyszczególnienie	Współczynniki	Błąd standardowy	Statystyka <i>t</i>	Graniczny poziom istotności
Wyraz wolny	2,958	0,183	16,199	0,000
Powszechność czekolady	1,463	0,191	7,672	0,000
Właściwości smakowe	0,743	0,191	3,895	0,004

Źródło: obliczenia własne z wykorzystaniem pakietu SPSS.

Model regresji liniowej, w którym uwzględniono główne składowe, jest dobrze dopasowany do danych i w odróżnieniu od modelu z pierwotnymi zmiennymi ma istotne statystycznie współczynniki.

4. Podsumowanie

Metoda głównych składowych przekształca zbiór skorelowanych zmiennych w zbiór ortogonalnych głównych składowych. Dzięki temu główne składowe są pozbawione współliniowości i dlatego są wykorzystywane w liniowej analizie regresji. Jednak należy pamiętać, że zastosowanie głównych składowych w regresji ma sens tylko wtedy, gdy składowe dają się zinterpretować, dlatego często uwzględnia się tylko kilka interpretowalnych głównych składowych.

Wykorzystanie głównych składowych w analizie regresji liniowej pozwala na lepsze dopasowanie modelu do danych, a także na pozbycie się niestabilności modelu regresji otrzymanego na podstawie zmiennych pierwotnych.

Literatura

- [1] Churchill G.A., *Badania marketingowe. Podstawy metodologiczne*, PWN, Warszawa 2002.
- [2] Crawford I.M., Lomas R.A., *Factor Analysis – a Tool for Data Reduction*, „European Journal of Marketing” 1980, vol. 14, nr 7, s. 414-421.
- [3] Hair J.F., Anderson R.E., Tatham R.L., Black W.C., *Multivariate Data Analysis with Readings*, Prentice-Hall, Englewood Cliffs 1995.
- [4] Jajuga K., *Statystyczna analiza wielowymiarowa*, PWN, Warszawa 1993.
- [5] Jolliffe I.T., *Principal Component Analysis*, Springer-Verlag, New York 2002.
- [6] Kim J.O., Mueller C.W., *Introduction to Factor Analysis. What It Is and How to Do It*, Sage, Beverly Hills 1978.
- [7] Kline P. (1994), *An Easy Guide to Factor Analysis*, Routledge, London 1994.
- [8] Rószkiewicz M., *Metody ilościowe w badaniach marketingowych*, PWN, Warszawa 2002.
- [9] Walesiak M., *Metody analizy danych marketingowych*, PWN, Warszawa 1996, s. 140-149.
- [10] Zakrzewska M., *Analiza czynnikowa w budowaniu i sprawdzaniu modeli psychologicznych*, UAM, Poznań 1994.

PRINCIPAL COMPONENTS IN REGRESSION ANALYSIS

Summary

In multiple regression, one of the major difficulties is the problem of multicollinearity. To overcome this problem various approaches have been proposed. One possibility is to use principal components.

The article presents the usage of principal components in regression analysis to reduce the collective features of chocolate in order to obtain a few components, which contribute mostly to the popularity of the product on the market.

Dr Mirosława Sztemberg-Lewandowska jest pracownikiem Katedry Ekonometrii i Informatyki w Akademii Ekonomicznej we Wrocławiu.