

ANALIZA DYSKRYMINACYJNA

Ideą analizy dyskryminacyjnej jest rozstrzygnięcie czy grupy różnią się ze względu na średnią pewnych zmiennych, a następnie wykorzystanie tych zmiennych do przewidywania przynależności do grupy (nowe przypadki). Analiza dyskryminacyjna bada kilka zmiennych jednocześnie w celu sprawdzenia, która z nich przyczynia się do dyskryminowania grup. Zmienne te noszą nazwę **zmiennych dyskryminujących**.

Analiza dyskryminacyjna to kilka powiązanych procedur statystycznych:

- A. Procedury opisujące i interpretujące różnice międzygrupowe
- B. Procedury klasyfikacji przypadków

A. Procedury opisujące i interpretujące różnice międzygrupowe

- **Kanoniczne funkcje dyskryminacyjne**

Postać tych funkcji może być dowolna – najczęściej wykorzystuje się funkcje liniowe:

$$D_{kj} = \beta_0 + \beta_1 x_{1kj} + \dots + \beta_p x_{pjk}$$

gdzie:

p - liczba zmiennych dyskryminacyjnych

n - liczebność próby

g - liczba grup

D_{kj} - wartość kanonicznej dyskryminacyjnej funkcji dla k-tego przypadku w j-tej grupie

$k=1, \dots, n; j=1, \dots, g$

x_{ijk} - wartość i-tej zmiennej dyskryminacyjnej dla k-tego przypadku w j-tej grupie

$i = 1, \dots, p$

β_i - współczynniki kanonicznej funkcji dyskryminacyjnej wyznaczone na podstawie własności funkcji.

Kanonicznych funkcji dyskryminacyjnych może być wiele. Maksymalna liczba obliczonych funkcji, będzie równa liczbie grup minus jeden lub liczbie zmiennych w analizie, w zależności od tego, która jest mniejsza.

- **Współczynniki funkcji dyskryminacyjnych β**

surowe - służą do obliczania wartości funkcji dyskryminacyjnych nie nadają się do interpretacji,

standaryzowane - wskazują jak silny jest wpływ danej zmiennej dyskryminacyjnej na różnicowanie grup (indywidualny (częstkowy) wkład każdej zmiennej do funkcji dyskryminacyjnej, w ich wyznaczaniu brane są pod uwagę wkłady wszystkich pozostałych zmiennych).

- **Współczynniki struktury czynnikowej**

Współczynnik korelacji między pojedynczymi zmiennymi a funkcjami dyskryminacyjnymi (mówi jak silne jest powiązanie między nimi). W ich wyznaczaniu nie bierze się pod uwagę wpływu pozostałych zmiennych.

- **Wartość własna funkcji dyskryminacyjnych (λ)**

Określa udział funkcji w dyskryminowaniu grup (im większa wartość własna tym lepsza dyskryminacja). Z wartością własną związany jest **procent wariancji międzygrupowej** wyjaśniony przez daną funkcję dyskryminacyjną, a który wyznaczany jest według wzoru:

$$\frac{\lambda_i}{\sum_{i=1}^n \lambda_i} 100\%$$

a także tzw. **korelacje kanoniczne R**

$$R_i = \sqrt{\frac{\lambda_i}{1 + \lambda_i}}$$

będące miarą powiązania pomiędzy i-tą kanoniczną funkcją dyskryminacyjną a odpowiadającą jej zmienną reprezentującą grupy (powiązanie między grupami, a funkcjami dyskryminującymi).

- **Statystyka lambda Wilksa**

Statystyka stosowana do wyznaczania istotności statystycznej mocy dyskryminacyjnej (sprawdzenie istotności funkcji dyskryminacyjnych). Procedura krokowa: w pierwszym kroku ocenia się czy warto wprowadzić 1 funkcję dyskryminacyjną, w drugim usuwana jest informacja dyskryminacyjna niesiona przez 1 funkcję i testowane jest czy dość dużo zróżnicowania zostało aby wprowadzić 2 funkcję dyskryminacyjną, itd.). Im lambda Wilksa jest bliższa 0 tym lepsza moc dyskryminacyjna. Statystyka ta ma w przybliżeniu rozkład chi-kwadrat.

B. Procedury klasyfikacji przypadków

- **Funkcje klasyfikacyjne**

Nie mylić z funkcjami dyskryminacyjnymi. Funkcji klasyfikacyjnych jest tyle ile grup. Funkcje te są wykorzystywane do rozstrzygnięcia do której grupy najprawdopodobniej należy dany przypadek.

$$K_i = c_{i0} + c_{i1}X_1 + c_{i2}X_2 + \dots + c_{ip}X_p$$

C_{ij} , $j=0, \dots, p$ – współczynniki funkcji klasyfikacyjnych.

Przypadek jest klasyfikowany do grupy dla której K_i przyjmuje wartość największą.

- **Odległość Mahalanobisa**

Opiera się na miarach odległości indywidualnego przypadku od centroidy grupy. Punkt należy do grupy, dla której ta odległość jest najmniejsza.

- **Prawdopodobieństwo a priori**

Prawdopodobieństwo, że dany przypadek należy do danej grupy bez wykorzystywania wiedzy o wartościach zmiennych dyskryminacyjnych.

Zad.1.

Na podstawie 4 parametrów biochemicznych (CRP w mg/l, glukoza mg%, MCV, IgG w g/l) dokonaj analizy dyskryminacyjnej ze względu na przebieg choroby. Czy na podstawie wartości tych 4 parametrów możemy przewidzieć przebieg choroby?

Moduł **Analizy dyskryminacyjnej** w Pakiecie Statistica wywołujemy z menu **Statystyka/Wielowymiarowe techniki eksploatacyjne/Analiza dyskryminacyjna**.

1. Wyznacz istotność statystyczną mocy dyskryminacyjnej przez wybrane zmienne dyskryminacyjne (4 parametry biochemiczne)
2. Wyznacz kanoniczne funkcje dyskryminacyjne. W tym celu przejdź na zakładkę **Więcej i Wykonaj analizę kanoniczną**
 - a. ile max. funkcji możesz wyznaczyć i dlaczego?
 - b. oceń na podstawie wartości własnych, R, Lambdy Wilksa i statystyki ch-kwadrat czy funkcje dyskryminacyjne są statystycznie istotne? **Podsum: Testy chi-kwadrat kolejnych pierwiastków**
 - c. Zapisz kanoniczne funkcje dyskryminacyjne (użyj surowych współczynników funkcji dyskryminacyjnych) **Współczynniki dla zmiennych kanonicznych**
 - d. Ile mocy dyskryminacyjnej jest wyjaśnione przez pierwszą a ile przez drugą funkcję?
 - e. oceń korelacje międzygrupowe zmiennych z odpowiednimi funkcjami dyskryminacyjnymi **Struktura czynnikowa**
 - f. które grupy zostały najlepiej wyróżnione przez każdą funkcję dyskryminacyjną **Średnie zmiennych kanonicznych**
 - g. wykreśl **wykres rozrzutu wartości kanonicznych** dostępny na karcie **Wartości kanoniczne** – dyskryminacja której grupy jest najbardziej wyraźna?
3. Wyznacz funkcje klasyfikacyjne **Klasyfikacja** w oknie **Wyniki analizy funkcji dyskryminacyjnej**

- a. dokonaj klasyfikacji przypadków przy prawdopodobieństwie a priori proporcjonalnym do wielkości grup
 - b. sprawdź jak rozkładają się odległości Mahalanobisa. Jeżeli odległość między przypadkiem a centroidą jest najmniejsza dla danej grupy to przypadek jest kwalifikowany czy odrzucany do danej grupy?
 - c. policz prawdopodobieństwo a posteriori, że przypadek należy do danej grupy np. przypadek MJ
 - d. jaka jest liczba i odsetek przypadków poprawnie sklasyfikowanych w każdej grupie
- Macierz klasyfikacji**

4. Czy wszystkie zmienne powinny być wybrane do dalszej analizy? Jeżeli nie dokonaj analizy dyskryminacyjnej krokowej postępującej. Co się stanie gdy F do wprowadzenia ustawimy na $F=13.22$?

Zad 2.

Pacjentów cierpiących na nadczynność tarczycy podzielono na 3 klasy:

Klasa 1 – leczenie pomyślne

Klasa 2 – leczenie bez wyników pozytywnych

Klasa 3 – nawrót choroby po początkowym leczeniu pozytywnym

Każdy z pacjentów miał przeprowadzony test radiojodem po 3 h (j3) i po 48 (J48) oraz zmierzono związany w białku proteinowym Jod – 131 po 48 h (PBJ). Jakie są zależności między podziałem pacjentów na klasy a pomiarami. Czy jesteśmy w stanie przewidzieć wynik kuracji?

1. dokonaj przeglądu statystyk opisowych
 - a. wyznacz średnią i liczbę przypadków dla każdej z grup oraz łącznie dla wszystkich grup
 - b. wykreśl wykresy normalności oraz histogramy skategoryzowane względem grup dla każdej zmiennej
 - c. pod kątem sprawdzenia występowania obserwacji odstających wykreśl macierzowy wykres rozrzutu
2. dokonaj analizy dyskryminacyjnej krokowej postępującej
3. wyznacz istotność statystyczną mocy dyskryminacyjnej przez wybrane zmienne dyskryminacyjne
4. Wyznacz kanoniczne funkcje dyskryminacyjne. W tym celu przejdź na zakładkę **Więcej i Wykonaj analizę kanoniczną**
 - a. ile max. funkcji możesz wyznaczyć i dlaczego?
 - b. oceń na podstawie wartości własnych, R, Lambdy Wilksa i statystyki ch-kwadrat czy funkcje dyskryminacyjne są statystycznie istotne? **Podsum: Testy chi-kwadrat kolejnych pierwiastków**
 - c. Zapisz kanoniczne funkcje dyskryminacyjne (użyj surowych współczynników funkcji dyskryminacyjnych) **Współczynniki dla zmiennych kanonicznych**

- d. Ile mocy dyskryminacyjnej jest wyjaśnione przez pierwszą a ile przez drugą funkcję?
 - e. oceń korelacje międzygrupowe zmiennych z odpowiednimi funkcjami dyskryminacyjnymi
 - f. które grupy zostały najlepiej wyróżnione przez funkcję dyskryminacyjną **Średnie zmiennych kanonicznych**
 - g. wykreśl **wykreś rozrzutu wartości kanonicznych** dostępny na karcie **Wartości kanoniczne** – dyskryminacja której grupy jest najbardziej wyraźna?
3. Wyznacz funkcje klasyfikacyjne **Klasyfikacja** w oknie **Wyniki analizy funkcji dyskryminacyjnej**
- a. dokonaj klasyfikacji przypadków przy prawdopodobieństwie a priori proporcjonalnym do wielkości grup oraz jednakowe dla wszystkich grup
 - b. sprawdź jak rozkładają się odległości Mahalanobisa dla przypadków błędnie zakwalifikowanych
 - c. policz prawdopodobieństwo a posteriori, że przypadek należy do danej grupy
 - d. jaka jest liczba i odsetek przypadków poprawnie sklasyfikowanych w każdej grupie **Macierz klasyfikacji**

Założenia analizy dyskryminacyjnej:

- 1. wielowymiarowa normalność (**przegląd statystyk opisowych/wewnętrzgrupowe/wykreś normalności skategoryzowany względem grup**)
- 2. jednorodność wariancji i kowariancji (wykorzystujemy **test M Boxa** z modułu **Anova/Manova**)
- 3. brak korelacji między średnimi a wariancjami w grupach (**wykreś odchyłeń standardowych względem średnich** z modułu **Anova/Manova**)
- 4. brak współliniowości (tolerancja mniejsza niż 0,1 oznacza redundantną zmienną)
- 5. rozmiar próby 4 do 5 razy większy niż liczba zmiennych dyskryminujących
- 6. brak wartości odstających

Zad 3.

Przeprowadź analizę założeń dla danych z Zad.1.