Statystyka II wykłady

Piotr Pokarowski
pokar@mimuw.edu.pl
http://www.mimuw.edu.pl/~pokar

Agnieszka Prochenka
ap220756@students.mimuw.edu.pl

20 lutego 2012

Streszczenie. Wykład poświęcony jest eksploracyjnej analizie danych i predykcji statystycznej.

Słowa kluczowe:

Projekt współfinansowany przez Unię Europejską w ramach Europejskiego Funduszu Społecznego.

Skład w systemie LaTeX, z wykorzystaniem m.in. pakietów beamer oraz listings. Szablony podręcznika i prezentacji: Piotr Krzyżanowski, koncept: Robert Dąbrowski.

Spis treści

Rozdział 1

Miary zależności i miary bliskości między zmiennymi

W rozdziale tym przedstawione zostaną wskaźniki liczbowe do analizy danych jedno- i dwuwymiarowych. Przypomniane zostaną miary rozrzutu oraz omówione miary zależności i miary bliskości między zmiennymi (cechami).

Główna różnica:

- *miary zależności*: minimalne dla zmiennych niezależnych, maksymalne dla zmiennych identycznych;

- *miary bliskości* (odległości, zróżnicowania): minimalne dla zmiennych identycznych;

Wygodnie jest podzielić zmienne na ilościowe (liczbowe), porządkowe i jakościowe (nominalne). Miary wprowadzone dla danych liczbowych, wykorzystujące wartości liczbowe, mają zastosowanie tylko dla nich. Miary dla zmiennych porządkowych nadają się również dla zmiennych liczbowych, ponieważ otrzymujemy je przez zamianę wartości cechy na kolejne liczby naturalne $1, 2, \ldots, n$ lub ułamki jednostajnie rozłożone na odcinku $[0, 1]$, czyli $\frac{i}{n} - \frac{1}{2n}$, gdzie $i = 1, 2, \ldots, n$. Miary dla zmiennych jakościowych są oparte na gęstościach i mają zastosowanie do wszystkich zmiennych.

Miary zależności i bliskości zostaną podzielone na symetryczne i niesymetryczne (zależność czy błąd nie muszą być relacjami symetrycznymi). Podsumowanie znajduje się w poniższej tabelce:

	Zmienne ilościowe	Zmienne porządkowe	Zmienne jakościowe
Miary zależności symetryczne	Korelacja	Korelacja rang, Współczynnik Kendalla	Wspólna Informacja
Miary zależności niesymetryczne			Współczynnik Goodmana-Kruskala
Miary odległości symetryczne	Błąd średniokwadratowy $\mathbb{E}(X-Y)^2$		
Miary odległości niesymetryczne	$\frac{\mathbb{E}(X-Y)^2}{\mathbb{E}Y^2}$		$\chi^2(p,q)$, $\mathrm{H}(p\|\|q)$

W dalszej części skryptu będziemy oznaczać wielkimi literami, np. X, Y zarówno zmienne losowe jak i ich realizacje. Rozróżnienie będzie wynikać z kontekstu.

Definicja 1.1. *Próbą* będziemy nazywali n realizacji zmiennej losowej: $X = (X_1, \ldots, X_n)$. *Średnią z próby* będziemy oznaczać jako:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad .$$

1.1 Zmienne ilościowe

Dla zmiennych ilościowych możemy zdefiniować **kwantyle**:

Definicja 1.2. *Kwantyl rzędu* p, $p \in (0,1)$:

$$\varphi_p(X) = F_X^{-1}(p) = \inf\{x : F_X(x) \geqslant p\},$$

gdzie F_X oznacza dystrybuantę. Kwantyle rzędu $\frac{1}{4}, \frac{1}{2}$, i $\frac{3}{4}$ nazywamy *kwartylami*, z czego kwantyl rzędu $\frac{1}{2}$ to *mediana*.

Estymatorami kwantyli dla próby $X = (X_1, \ldots, X_n)$ są kwantyle próbkowe.

Definicja 1.3. *Kwantyle próbkowe rzędu* p *dla próby* n*-elementowej*, $p \in (0,1)$:

$$\widehat{\varphi}_p(X) = \frac{1}{2}(X_{\lfloor p(n+1):n \rfloor} + X_{\lceil p(n+1):n \rceil}),$$

gdzie $X_{k:n}$ oznacza k-ty element statystyki pozycyjnej: po uszeregowaniu niemalejąco wartości $X_i \quad i = 1, \ldots, n, \quad X_{k:n}$ oznacza k-tą wartość z n-elementowego ciągu. Funkcja ceiling zwraca najmniejszą liczbę całkowitą mniejszą od danej, a funkcja floor największą liczbę całkowitą mniejszą.

1.1.1 Miary rozrzutu

- **Wariancja**:

wersja populacyjna

$$\mathrm{Var}(X) = \mathbb{E}(X - \mathbb{E}X)^2 \quad ;$$

wersja próbkowa

$$\mathrm{var}(X) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \quad ;$$

wersja próbkowa nieobciążona

$$\mathrm{var}_0(X) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \quad .$$

- **Odchylenie standardowe**:

  wersja populacyjna
  $$\sigma(X) = \sqrt{\mathrm{Var}(X)} \quad ;$$
  wersja próbkowa
  $$\hat{\sigma}(X) = \sqrt{\mathrm{var}(X)} \quad ;$$
  wersja próbkowa niebciążona
  $$\hat{\sigma}_0(X) = \sqrt{\mathrm{var}_0(X)} \quad .$$

- **MAD** (*Median of Absolute Deviation*):

  $$\mathrm{MAD}(X) = \widehat{\varphi}_{\frac{1}{2}}(X - \widehat{\varphi}_{\frac{1}{2}}(X)).$$

  Dla rozkładu normalnego, MAD$\approx 1,35\sigma$.

- **IQR** (*Interquartile Range*)

  $$\mathrm{IQR}(X) = \widehat{\varphi}_{\frac{3}{4}}(X) - \widehat{\varphi}_{\frac{1}{4}}(X) \quad .$$

  Dla rozkładu normalnego IQR$\approx 0,6745\sigma$.

1.1.2 Miary zależności

- **Korelacja** pomiędzy zmiennymi X i Y:

  $$\mathrm{Cor}(X,Y) = \frac{\mathbb{E}(X - \mathbb{E}X)(Y - \mathbb{E}Y)}{\sigma(X)\sigma(Y)} \quad .$$

  Dla próby n-elemnentowej: $(X_1, Y_1), \ldots, (X_n, Y_n)$ korelacja próbkowa:

  $$\mathrm{cor}(X,Y) = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\widehat{\sigma}(X)\widehat{\sigma}(Y)} \quad ;$$

korelacja próbkowa nieobciążona:

$$\mathrm{cor}_0(X,Y) = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\widehat{\sigma}_0(X)\widehat{\sigma}_0(Y)} \quad ,$$

$$\mathrm{cor}_0(X,Y) = \mathrm{cor}(X,Y).$$

1.2 Zmienne porządkowe

Zmienne porządkowe to takie, dla których nie są ważne wartości, ale kolejność, w jakiej są ustawione. Z pojęciem zmiennej porządkowej ściśle wiąże się pojęcie *rangi*. Nadanie rang obserwacjom uniezależnia je od skali.

Definicja 1.4. *Rangi* dla obserwacji w próbie $X = (X_1, \ldots, X_n)$:

$$R_i = R(X_i) = \#\{j : X_j < X_i\} + \frac{1 + \#\{j : X_j = X_i\}}{2} \quad .$$

Przykład 1.1. Dla $X = (2; 3; 2,5; 2,5; 1,5)$ rangi są równe:

X	2	3	2,5	2,5	1,5
R	2	5	3,5	3,5	1

1.2.1 Miary zależności

- **Korelacja rang** (Spearmana):

$$\mathrm{cor}_R(X,Y) = \mathrm{cor}(R(X), R(Y)) \quad ,$$

gdzie $R(X) = (R(X_1), R(X_2), \ldots, R(X_n))$, $R(Y) = (R(Y_1), R(Y_2), \ldots, R(Y_n))$.

Stwierdzenie 1.1. *Załóżmy, że X, Y mają rozkłady o ciągłych i ściśle rosnących dystrybuantach. Wtedy:*

1. $\frac{1}{n}\sum_{i=1}^{n} R_i = \frac{n+1}{2}$;
2. $\mathrm{var(R(X))} = \frac{1}{n-1}\sum_{i=1}^{n}(R(X_i) - \overline{R})^2 = \frac{n(n+1)}{12}$;
3. $\mathrm{cor}_R(X,Y) = \frac{12}{n(n^2-1)}\sum_{i=1}^{n} R(X_i)R(Y_i) - \frac{3(n+1)}{n-1}$;
4. $\mathrm{cor}_R(X,Y) \xrightarrow[n\to\infty]{p.n} \mathrm{Cor}(F_X(X), F_Y(Y))$;
5. Jeżeli (X,Y) ma 2-wymiarowy rozkład normalny,

$$\mathrm{Cor}(F_X(X), F_Y(Y)) = \frac{6}{\pi}\arcsin\left(\frac{\mathrm{Cor}(X,Y)}{2}\right).$$

- **Współczynnik Kendala** zależności między X a Y:

  Załóżmy, że X_1, X_2 to zmienne losowe niezależne i o takim samym rozkładzie co X, Y_1, Y_2 niezależne o takim samym rozkładzie co Y. Niech:

$$X_{12} = \begin{cases} 1, & X_1 > X_2 \ ; \\ 0, & X_1 = X_2 \ ; \\ -1, & X_1 < X_2 \ . \end{cases} \qquad Y_{12} = \begin{cases} 1, & Y_1 > Y_2 \ ; \\ 0, & Y_1 = Y_2 \ ; \\ -1, & Y_1 < Y_2 \ . \end{cases}$$

$$\tau_k(X,Y) = \mathrm{Cor}(X_{12}, Y_{12}) =$$

$$= \frac{\mathbb{P}((X_1 - X_2)(Y_1 - Y_2) > 0) - \mathbb{P}((X_1 - X_2)(Y_1 - Y_2) < 0)}{\sqrt{\mathbb{P}(X_1 \neq X_2)\mathbb{P}(Y_1 \neq Y_2)}}.$$

Jeśli X, Y mają ciągłe dystrybuanty, to

$$\tau_k(X,Y) = \mathbb{P}((X_1 - X_2)(Y_1 - Y_2) = 1) - \mathbb{P}((X_1 - X_2)(Y_1 - Y_2) = -1).$$

Wersja próbkowa τ_k:

$$\widehat{\tau}_k = \frac{2}{n(n-1)} \sum_{1 \leqslant i \leqslant j \leqslant n} \mathrm{sgn}((X_i - X_j)(Y_i - Y_j)).$$

Uwaga 1.1. W programie R: `cor(X, Y, method = c("pearson", "kendall", "spearman"))` z domyślnie ustawioną opcją `"pearson"`.

1.3 Zmienne jakościowe

W tej części omówione zostaną miary rozrzutu, zależności i bliskości oparte na gęstościach prawdopodobieństwa, wykorzystywane przede wszytkim do analizy cech jakościowych.

1.3.1 Miary rozrzutu

- **Entropia** dla gęstości p o nośniku Ω:

$$\mathrm{H}(p) = -\int_\Omega [\ln p(v)] p(v) dv.$$

  Jeśli X - zmienna losowa o gęstości p_X, to $\mathrm{H}(X) := \mathrm{H}(p_X)$.

Uwaga 1.2. Różnice i podobieństwa między $\mathrm{H}(X)$ i $\mathrm{Var}(X)$:

1. Załóżmy, że $0 \leqslant X \leqslant 1$. Wtedy

$$\mathrm{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}X)^2 \leqslant \mathbb{E}X - (\mathbb{E}X)^2 \leqslant \frac{1}{4}.$$

Zatem wariancja jest największa dla rozkładu dwupunktowego: $p_0 = \frac{1}{2} = p_1$. Entropia natomiast jest największa dla rozkładu jednostajnego.

2. Załóżmy teraz, że $X \sim \mathcal{N}(\mu, \sigma^2)$. Mamy:

$$\ln p_X(x) = \ln\left(\frac{1}{\sqrt{2\pi}\sigma}\right) - \frac{(x-\mu)^2}{2\sigma^2};$$

$$\mathrm{Var}(X) = \int (x-\mu)^2 p_X(x)dx = \sigma^2;$$

$$\mathrm{H}(X) = -\ln\left(\frac{1}{\sqrt{2\pi}\sigma}\right) + \frac{1}{2} = const + \frac{1}{2}\ln\sigma^2.$$

- **Współczynnik Giniego** dla gęstości p:

$$\mathrm{V}(p) = \int_\Omega [1 - p(v)]p(v)dv = 1 - \int_\Omega p^2(v)dv.$$

Jeśli X - zmienna losowa o gęstości p_X, to $\mathrm{V}(X) = \mathrm{V}(p_X)$.

V(p) jest liniowym (rozwinięcie Taylora dla logarytmu naturalnego: $\ln(x) \approx x - 1$) przybliżeniem H(p).

1.3.2 Miary bliskości

Dla prostoty ograniczymy się w dalszej części wykładu do rozkładów dyskretnych zadanych gęstościami p i q o wspólnym nośniku Ω.

- **Odległość Kullbacka-Leiblera** (względna entropia):

$$\mathrm{H}(p||q) = \sum_{i\in\Omega} \left(\ln\frac{p_i}{q_i}\right) p_i \quad .$$

Stwierdzenie 1.2. *Własności Odległości Kullbacka-Leiblera (entropii względnej):*

1. $\mathrm{H}(p||q) \geqslant 0$;
2. $\mathrm{H}(p||q) = 0 \Leftrightarrow p = q$.

Dowód. Skorzystajmy z nierówności: $\ln(x) \leqslant x - 1$:

$$\mathrm{H}(p||q) = \sum_{i\in\Omega} \left(\ln \frac{p_i}{q_i}\right) p_i = -\sum_{i\in\Omega} \left(\ln \frac{q_i}{p_i}\right) p_i \geqslant \sum_{i\in\Omega} \left(1 - \frac{q_i}{p_i}\right) p_i = 0.$$

Stąd otrzymujemy (1). Równość w ostatniej nierówności jest równoważna warunkowi $\frac{q_i}{p_i} = 1$ dla wszystkich i, otrzymujemy (2).

□

- **Odległość χ^2** między rozkładami dyskretnymi zadanymi gęstościami p i q o wspólnym nośniku Ω:

$$\chi^2(p,q) = \sum_{i\in\Omega} \left(\frac{p_i - q_i}{p_i}\right)^2 p_i = \sum_{i\in\Omega} \frac{(p_i - q_i)^2}{p_i} \quad . \tag{1.1}$$

  Odległość χ^2 jest kwadratowym (rozwinięcie Taylora dla logarytmu: $\ln(1+x) \approx x - \frac{1}{2}x^2$, gdzie za x wstawiamy $\frac{q_i - p_i}{p_i}$) przybliżeniem $\mathrm{H}(p||q)$: $\mathrm{H}(p||q) \approx \frac{1}{2}\chi^2(p,q)$.

1.3.3 Miary zależności

Niech X, Y będą zmiennymi o rozkładzie dyskretnym, niekoniecznie o tym samym nośniku Ω. Ponadto zdefiniujmy:

$p_{ij} = \mathbb{P}(X = i, Y = j)$;

$p_{j|i} = \mathbb{P}(Y = j | X = i)$;

$p_{i.} = \mathbb{P}(X = i)$;

$p_{.j} = \mathbb{P}(Y = j)$;

warunkowy współczynnik Giniego $\mathrm{V}(Y|X = i) = 1 - \sum_j p_{j|i}^2$;

warunkową entropię $\mathrm{H}(Y|X = i) = -\sum_j (\ln p_{j|i}) p_{j|i}$.

Zauważmy, że:

$$\mathbb{E}\mathrm{V}(Y|X) = \sum_i \mathrm{V}(Y|X = i) p_{i.} = 1 - \sum_i p_{i.} \sum_j p_{j|i}^2;$$

$$\mathbb{E}\mathrm{H}(Y|X) = \sum_i \mathrm{H}(Y|X = i) p_{i.}.$$

- **Współczynnik Goodmana-Kruskala** (mówi on o tym, jak zmienił się rozrzut po zaobserwowaniu cechy X):

$$\tau(Y|X) = \frac{\mathrm{V}(Y) - \mathbb{E}(\mathrm{V}(Y|X))}{\mathrm{V}(Y)}.$$

  Zakładamy, że rozkład Y jest niezdegenerowany, czyli że $\mathrm{V}(Y) > 0$.

Stwierdzenie 1.3. *Własności Współczynnika Goodmana-Kruskala:*

1. $0 \leqslant \tau \leqslant 1$;
2. $\tau = 0 \Leftrightarrow X, Y$ niezależne.

Dowód. Oczywiście $\tau \leqslant 1$. Dla dowodu, że $\tau \geqslant 0$ zauważmy, że $\mathbb{E}\mathrm{V}(Y|X) = 1 - \sum_i p_{i\cdot} \sum_j p^2_{j|i}$. Wystarczy pokazać, że $\sum_j p^2_{\cdot j} \leqslant \sum_j \sum_i p_{i\cdot} p^2_{j|i}$. Z kolei wystarczy pokazać, że $p^2_{\cdot j} \leqslant \sum_i p_{i\cdot} p^2_{j|i}$. Lewa$= p^2_{\cdot j} = (\sum_i p_{ij})^2 = (\sum_i p_{j|i} p_{i\cdot})^2$, więc (1) wynika z nierówności Jensena.

Dla dowodu (2) zauważmy, że „=” w nierówności Jensena wyrazów $p_{j|i} = p_{\cdot j}$ $\forall i, j$ jest równoważna niezależności X, Y. □

- **Wspólna informacja** zawarta w X i Y:

$$\mathrm{M}(X, Y) = \sum_{j=1}^{l} \sum_{i=1}^{k} p_{ij} \ln \frac{p_{ij}}{p_{i\cdot} p_{\cdot j}}$$

Stwierdzenie 1.4. *Własności Wspólnej informacji:*

1. $\mathrm{M}(X, Y) \geqslant 0$;
2. $\mathrm{M}(X, Y) = 0 \Leftrightarrow X, Y$ niezależne.

Dowód. Wynika z własności odległości Kullbacka-Leiblera (stwierdzenie 1.2), bo $\mathrm{M}(X, Y) = H(p_{ij}||(p_{i\cdot} p_{\cdot j}))$ □

Uwaga 1.3. Korzystając z przybliżenia rozwinięciem w szereg Taylora logarytmu: $\ln(1 + x) \approx x - \frac{1}{2}x^2$, otrzymujemy:

$$\begin{aligned}
\mathrm{M}(X, Y) =& -\sum_j \sum_i p_{ij} \ln \left(\frac{p_{i\cdot} p_{\cdot j}}{p_{ij}} - 1 + 1 \right) \approx \\
=& -\left[\sum_{ji} \left(\frac{p_{i\cdot} p_{\cdot j}}{p_{ij}} - 1 \right) p_{ij} - \frac{1}{2} \sum_{ji} \left(\frac{p_{i\cdot} p_{\cdot j}}{p_{ij}} - 1 \right)^2 p_{ij} \right] = \\
=& \frac{1}{2} \sum_{ij} \frac{(p_{i\cdot} p_{\cdot j} - p_{ij})^2}{p_{ij}}.
\end{aligned}$$

Ostatnie wyrażenie oraz statystyka $\chi^2 = \sum_{ij} \frac{(p_{ij} - p_{i\cdot} p_{\cdot j})^2}{p_{i\cdot} p_{\cdot j}}$ dla testowania niezależności mają podobną interpretację, chociaż różnica w treści matematycznej jest zasadnicza. Być może o podobieństwie wyrażeń decydują własności błędu względnego: jeśli błąd względny oszacowania a za pomocą b jest nie większy od ε, to błąd względny oszacowania b za pomocą a jest nie większy niż $\frac{\varepsilon}{1-\varepsilon}$. Przy małym ε wyrażenia te są porównywalne.

Uwaga 1.4. Wspólna informacja dla rozkładu dwuwymiarowego normalnego, gdzie $(X,Y) \sim \mathcal{N}\left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}\right)$, jest równa:

$$\mathrm{M}(X,Y) = -\ln\sqrt{1-\rho^2},$$

gdzie $\rho = \mathrm{Cor}(X,Y)$.

- Jeżeli zamiast współczynnika Giniego we wzorze na współczynnik Goodmana-Kruskala użyjemy entropii, otrzymamy analogiczny **współczynnik $\mathrm{M}(Y|X)$**:
$$\mathrm{M}(Y|X) = \frac{\mathrm{H}(Y) - \mathbb{E}\mathrm{H}(Y|X)}{\mathrm{H}(Y)}.$$

Stwierdzenie 1.5.

$$\mathrm{M}(Y|X) = \frac{\mathrm{H}(Y) - \mathbb{E}\mathrm{H}(Y|X)}{\mathrm{H}(Y)} = \frac{\mathrm{M}(X,Y)}{\mathrm{H}(Y)}.$$

Dowód.

$$\begin{aligned}
\mathbb{E}\mathrm{H}(Y|X) &= -\sum_i p_{i\cdot} \sum_j p_{j|i} \ln p_{j|i} = -\sum_{ij} p_{ij} \ln p_{j|i} = \\
&= -\left[\sum_{ij} p_{ij} \ln p_{ij} - \sum_{ij} p_{ij} \ln p_{i\cdot}\right] = \\
&= \mathrm{H}(X,Y) + \sum_i \ln p_{i\cdot} \sum_j p_{ij} = \mathrm{H}(X,Y) + \sum_i \ln p_{i\cdot} \cdot p_{i\cdot} = \\
&= \mathrm{H}(X,Y) - \mathrm{H}(X),
\end{aligned}$$

gdzie $\mathrm{H}(X,Y) = \mathrm{H}(P_{(X,Y)})$.
Zatem $\mathrm{M}(Y|X) = \frac{\mathrm{H}(X)+\mathrm{H}(Y)-\mathrm{H}(X,Y)}{\mathrm{H}(Y)}$.

$$\begin{aligned}
\mathrm{M}(X,Y) &= \sum_i \sum_j p_{ij} \ln \frac{p_{ij}}{p_{i\cdot}p_{\cdot j}} = \\
&= \sum_i \sum_j p_{ij} \ln p_{ij} - \sum_i \sum_j p_{ij} \ln(p_{i\cdot}p_{\cdot j}) = \\
&= -\mathrm{H}(X,Y) - \sum_i \sum_j p_{ij} \ln p_{i\cdot} - \sum_j \sum_i p_{ij} \ln p_{\cdot j} = \\
&= -\mathrm{H}(X,Y) - \sum_i \ln(p_{i\cdot})p_{i\cdot} - \sum_j \ln(p_{\cdot j})p_{\cdot j} = \\
&= -\mathrm{H}(X,Y) + \mathrm{H}(X) + \mathrm{H}(Y).
\end{aligned}$$

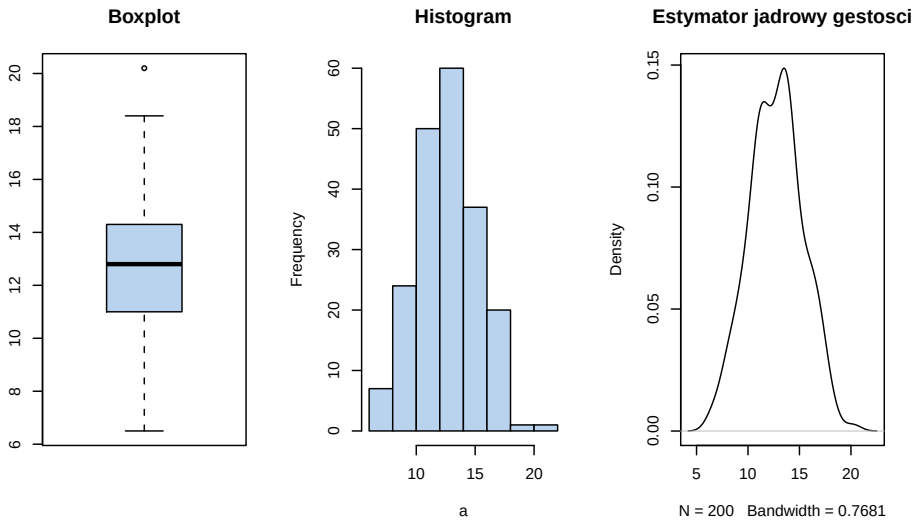
Zatem $\mathrm{M}(X,Y) = \mathrm{H}(X) + \mathrm{H}(Y) - \mathrm{H}(X,Y)$. Stąd $\mathrm{M}(Y|X) = \frac{\mathrm{M}(X,Y)}{\mathrm{H}(Y)}$. □

1.4 Przykłady w programie R

Obliczanie spółczynnika Goodmana-Kruskala: http://www.mimuw.edu.pl/~pokar/StatystykaII/EKSPLORACJA/tauGK.R

Rozdział 2

Podstawowe metody prezentacji danych



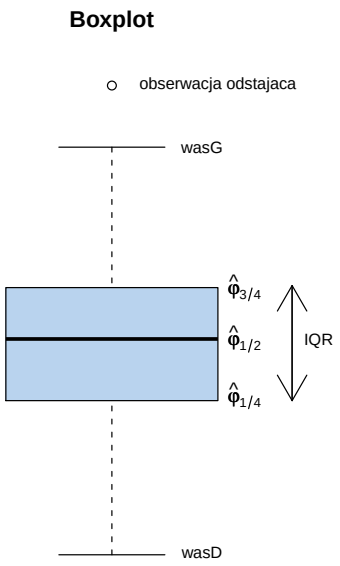
Rysunek 2.1: Przykładowe boxplot, histogram i estymator jądrowy gęstości dla zmiennej crabs[, 5] ze zbioru danych crabs (MASS).

2.1 Boxplot

Dla próby losowej $X = (X_1, \ldots, X_n)$ zajmiemy się reprezentacją graficzną danych. Zaczniemy od boxplotu.

- **Boxplot**. Przykładowy boxplot znajduje się na rysunku 2.1. Do jego narysowania potrzebne są następujące elementy:

1. kwartyle próbkowe $\hat{\varphi}_{\frac{1}{4}}(X), \hat{\varphi}_{\frac{1}{2}}(X), \hat{\varphi}_{\frac{3}{4}}(X)$;
2. rozstęp międzykwartylowy (wysokość pudełka) $\mathrm{IQR}(X) = \hat{\varphi}_{\frac{3}{4}}(X) - \hat{\varphi}_{\frac{1}{4}}(X)$;
3. wąs górny $\mathrm{wasG}(X) = (\hat{\varphi}_{\frac{3}{4}}(X) + 1,5 * \mathrm{IQR}(X)) \wedge \max(X)$, gdzie $\max(X)$ oznacza element maksymalny z próby;
4. wąs dolny $\mathrm{wasD}(X) = (\hat{\varphi}_{\frac{1}{4}}(X) - 1,5 * \mathrm{IQR(X)}) \vee \min(X)$, gdzie $\min(X)$ oznacza element minimalny z próby;
5. obserwacje odstające, które nie mieszczą się w przedziale $[\mathrm{wasD}(X), \mathrm{wasG}(X)]$ i nanosimy je oddzielnie w postaci punktów.



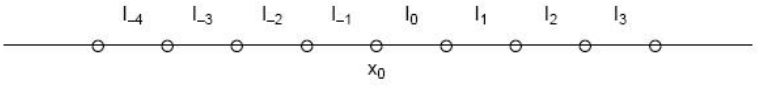
Rysunek 2.2: Przykładowy boxplot.

2.2 Estymacja gęstości

Załóżmy, że próba $X = X_1, \ldots, X_n$ pochodzi z rozkładu o gęstości f i jest iid (niezależna o tym samym rozkładzie), będziemy szukać estymatora dla gęstości f.

- **Histogram**. Przykładowy histogram znajduje się na rysunku 2.1. Wybieramy dowolne $x_0 \in \mathbb{R}$. Dla ustalonego $h > 0$, oznaczającego szerokość klasy, tworzymy odcinki:

$$I_m = [x_0 + mh, x_0 + (m+1)h), \quad m = \ldots, -2, -1, 0, 1, 2, \ldots;$$



wtedy $\forall x \in \mathbb{R} \quad \exists!\ m$, że $x \in I_m$, oznaczmy $I(x) = I_m$ jeśli $x \in I_m$.

Histogramem nazywamy funkcję $x \in \mathbb{R}$:

$$\hat{f}_n(x) = \frac{\#\{i : x_i \in I(x)\}}{nh} = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h}\mathbf{1}_{(x_i \in I(x))}.$$

Uwaga 2.1. Podczas rysowania histogramu ważną kwestią jest dobór odpowiedniej szerokości przedziału, h. Istnieje wiele konwencji wyboru, niektóre z nich to:

$$h_{opt} \approx cn^{-\frac{1}{3}}, \text{ gdzie}$$

1. Jeżeli $f, f' \in L^2(\mathbb{R})$, $c = \left(\frac{6}{\int [f'(x)]^2 dx}\right)^{\frac{1}{3}}$.
2. Jeśli f jest normalna, $c = 2 \cdot 3^{\frac{1}{3}} \pi^{\frac{1}{6}} \sigma \approx 3,186\sigma$.
3. Inny wybór to $c = 2,64 *$ IQR.

- **Estymator jądrowy gęstości**. Przykładowy estymator jądrowy gęstości znajduje się na rysunku 2.1.

$$\check{f}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\bar{h}} K\left(\frac{x - X_i}{\bar{h}}\right).$$

Dla budowy tego estymatora ważny jest dobór dwóch parametrów: szerokości pasma $\bar{h}$ oraz funkcji jądra K. Jądro jest gęstością dowolnego rozkładu, czyli jest dowolną funkcją określoną na $\mathbb{R}$ o własnościach $K \geqslant 0$, $\int K(x)dx = 1$. Jednym z wyborów może być jądro postaci:

$$K_e(t) = \begin{cases} \frac{3}{4\sqrt{5}}(1 - \frac{1}{5}t^2), & |t| \leqslant \sqrt{5}; \\ 0, & \text{wpp.} \end{cases}$$

Uwaga 2.2. 1. Jeśli $f \in C^2$ oraz $\int (f'')^2 < \infty$, to w klasie symetrycznych jąder $K \in L^2(\mathbb{R})$, asymptotycznie optymalne jest K_e. Ponadto:

$$\bar{h}_{opt} \approx c \cdot n^{-\frac{1}{5}} \text{ , gdzie}$$

$$c = \left[\int t^2 K(t)dt\right]^{-\frac{2}{5}} \left[\int K^2(t)dt\right]^{\frac{1}{5}} \left[\int (f''(x))^2 dx\right]^{-\frac{1}{5}}.$$

2. Jeśli f jest gęstością rozkładu normalnego, to $\int (f''(x))^2 dx = \frac{3}{8}\pi^{-\frac{1}{2}}\sigma^{-5} \approx 0,212\sigma^{-5}$.

3. Jeśli jądro K jest gęstością standardowego rozkładu normalnego oraz f jest rozkładem normalnym, to $\bar{h}_{norm} = 1,06\sigma n^{-\frac{1}{5}}$.

4. Jeśli jądro jest równe K_e oraz f jest rozkładem normalnym, to $\bar{h}_e = 1,05\sigma n^{-\frac{1}{5}}$

5. Domyślnie w programie R nastawiona jest metoda Silvermana wyboru parametru $\bar{h}$: $\bar{h} = 0,9 \min(\hat{\sigma}^2, \frac{\text{IQR}}{1,34}) n^{-\frac{1}{5}}$.

2.3 Przykłady w programie R

Estymator jądrowy gęstości:

- dla danych Pima jednowymiarowy: http://www.mimuw.edu.pl/~pokar/StatystykaII/EKSPLORACJA/density.R
- dla danych Pima dwuwymiarowy: http://www.mimuw.edu.pl/~pokar/StatystykaII/EKSPLORACJA/density2d.R
- nastawianie szerokosci pasma w estymatorze jadrowym gestosci http://www.mimuw.edu.pl/~pokar/StatystykaII/EKSPLORACJA/bandwidth.pp.R

Rozdział 3

Analiza Składowych Głównych

3.1 Analiza danych wielowymiarowych

Oznaczenia:

- X będzie oznaczał wektor losowy w przestrzeni $\mathbb{R}^p$: $X = \begin{pmatrix} X_1 \\ \dots \\ X_p \end{pmatrix}$.
- Przez t oznaczymy wektor liczb, $t \in \mathbb{R}^p$.
- C to macierz liczb: $C \in \mathbb{R}^{r \times p}$.
- Σ będzie oznaczać macierz kowariancji wektora losowego X, czyli:

$$\Sigma = (\mathbb{E}(X_i - \mathbb{E}X_i)(X_j - \mathbb{E}X_j))_{i,j=1}^{p}.$$

  Oznaczeń $\Sigma = \mathrm{Var}(X) = \mathrm{Cov}(X)$ będziemy używać zamiennie.

Stwierdzenie 3.1. *Proste własności wprowadzonych pojęć:*

1. $\mathbb{E}(t^T X) = t^T \mathbb{E}X$, $\mathbb{E}(X^T t) = (\mathbb{E}X)^T t$.
2. $\mathbb{E}(CX) = C\ \mathbb{E}X$, *gdzie:*

$$C = \begin{pmatrix} C_1^T \\ \dots \\ C_r^T \end{pmatrix}, \quad \mathbb{E}(CX) \overset{(1)}{=} \begin{pmatrix} C_1^T \\ \dots \\ C_r^T \end{pmatrix} \mathbb{E}X$$

   .

3. *Macierz kowariancji jest równa:*

$$\Sigma = \mathbb{E}\left[(X - \mathbb{E}X)(X - \mathbb{E}X)^T\right].$$

4. *Macierz kowariancji ma następującą własność:*

$$\begin{aligned} Var(CX) =& \mathbb{E}\left[(CX - \mathbb{E}CX)(CX - \mathbb{E}CX)^T\right] = \\ =& \mathbb{E}\left[C(X - \mathbb{E}X)(X - \mathbb{E}X)^T C^T\right] = \\ =& C\ \mathbb{E}\left[(X - \mathbb{E}X)(X - \mathbb{E}X)^T\right] C^T = \\ =& C(\ VarX)C^T. \end{aligned}$$

5. *Ponadto, macierz $Var(X)$ jest symetryczna i nieujemnie określona:*

   symetryczność wynika z symetryczności kowariancji dwóch zmiennych losowych;

   nieujemna określoność wynika z nieujemności wariancji dla zmiennej losowej. Dla C^T o wymiarach $1 \times p$:

$$0 \leqslant Var(\ \underbrace{C^T X}_{zm\ losowa}\ ) \overset{(4)}{=} C^T\, Var(X)C = C^T \Sigma C.$$

6. *Jeżeli $Var(X) = \sigma^2 I_p$, a macierz C jest ortonormlna o wymiarach $p \times p$ ($C^T C = CC^T = I_p$), to:*

$$Var(CX) = C\, Var(X)C^T = \sigma^2 CC^T = \sigma^2 I_p \text{ , czyli się nie zmienia.}$$

3.2 Redukcja Wymiaru danych

Wygodną postacią macierzy wariancji $\Sigma = \mathrm{Var}(X)$ jest postać diagonalna. Wtedy korelacje pomiędzy różnymi elementami wektora losowego są zerowe.

Problem 3.1. Jak przekształcić wektor losowy X żeby zdiagonalizować Σ?

Twierdzenie 3.1. ***Rozkład spektralny*** *macierzy symetrycznej A. Dla symetrycznej macierzy A o wymiarze $p \times p$ istnieją:*

- *ortonormalna (czyli $VV^T = I_p$) macierz kwadratowa V o wymiarze $p \times p$, oznaczmy $V = [v_1, \ldots, v_p]$;*
- *diagonalna macierz Λ o wyrazach na przekątnych $(\lambda_1, \ldots, \lambda_p)$, że*

$$Av_i = \lambda_i v_i \text{ , czyli}$$

v_i to wektory własne macierzy A, a λ_i to wartości własne, które dla macierzy symetrycznej są rzeczywiste. Wtedy:

$$\begin{aligned} A[v_1, \ldots, v_p] &= [\lambda_1 v_1, \ldots, \lambda_p v_p]; \\ AV &= V\Lambda; \\ A &= V\Lambda V^T; \\ \Lambda &= V^T A V. \end{aligned}$$

Ponieważ macierz kowariancji Σ wektora losowego X jest symetryczna, możemy zastosować do niej rozkład spektralny: $\Sigma = V\Lambda V^T$. Pomnóżmy wektor X przez macierz V^T: $V^T X$. Macierz kowariancji dla takiego wektora to:

$$\mathrm{Var}(V^T X) = V^T \mathrm{Var}(X) V = V^T \Sigma(X) V = \Lambda.$$

Ponieważ macierz Σ jest nieujemnie określona, wszystkie jej wartości własne są nieujemne: $\lambda_i \geqslant 0$. Uporządkujmy wartości własne λ_i i odpowiadające im wektory własne v_i tak, żeby $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_p \geqslant 0$. Oznaczmy dla tak ustawionych wektorów własnych:

$$Y = V^T X.$$

3.2.1 Analiza składowych głównych – wersja populacyjna

Definicja 3.1. Mamy wektor losowy $X \in \mathbb{R}^p$ oraz macierz kowariancji $\mathrm{Var}(X) = \Sigma = V\Lambda V^T$.

Składowymi głównymi (*principal components*) nazywamy elementy wektora $[Y_1, \ldots, Y_p]^T = Y = V^T X$.

Kierunkami głównymi (*rotations*) nazywamy kolumny macierzy $V = [v_1, \ldots, v_p]$.

Stwierdzenie 3.2. *Własności składowych głównych:*

- *wsółrzędne wektora Y są nieskorelowane;*
- *wariancje poszczególnych Y_i równe są λ_i;*
- *Y_i ustawione są od 1 do p w kolejności nierosnących wariancji;*
- *Y_i to kombinacje liniowe zmiennych losowych $X_1, \ldots, X_p$;*

Stwierdzenie 3.3. *Kierunki główne to unormowane wektory, w kierunku których obserwujemy największą wariancję danych, będące wzajemnie do siebie prostopadłe:*

1. *jeżeli $t_1 \in \mathbb{R}^p$, $t_1^T t_1 = 1 \Rightarrow$ $Var(t_1^T X)$ osiąga maksimum$= \lambda_1$ dla $t_1 = v_1$.*
2. *jeżeli $t_2 \in \mathbb{R}^p$, $t_2^T t_2 = 1$, $t_2^T t_1 = 0 \Rightarrow$ $Var(t_2^T X)$ osiąga maksimum$= \lambda_2$ dla $t_2 = v_2$.*

Dowód. $V = [v_1, \ldots, v_p]$ jest bazą ortonormalną przestrzeni $\mathbb{R}^p$.

1. Zapiszmy t_1 w tej bazie: $t_1 = \sum_{i=1}^p c_i v_i$, gdzie $c_i \in \mathbb{R}$ współczynniki. Z założeń wynika:

$$t_1^T t_1 = \sum_{i=1}^p c_i^2 = 1.$$

Zauważmy, że:

$\mathrm{Var}(t_1^T X) = t_1^T \mathrm{Var}(X) t_1 = (c_1 v_1^T + \ldots + c_p v_p^T) \cdot \Sigma \cdot (c_1 v_1 + \ldots + c_p v_p) =$

z własności wektorów własnych macierzy,

$$= (c_1 v_1^T + \ldots + c_p v_p^T) \cdot (\lambda_1 c_1 v_1 + \ldots + \lambda_p c_p v_p) = \sum_{i=1}^{p} \lambda_i c_i^2 \leqslant \lambda_1.$$

Jeżeli przyjmiemy $t_1 = v_1$, czyli $c_1 = 1$, $c_{\geqslant 2} = 0$, otrzymujemy kombinację liniową o maksymalnej wariancji równej λ_1.

2. Ponieważ $t_2^T v_1 = 0$, możemy zapisać:

$\mathrm{Var}(t_2^T X) = t_2^T \mathrm{Var}(X) t_2 = (c_2 v_2^T + \ldots + c_p v_p^T) \cdot \Sigma \cdot (c_2 v_2 + \ldots + c_p v_p) =$

$$= \sum_{i=2}^{p} \lambda_i c_i^2 \leqslant \lambda_2.$$

Analogicznie, $t_2 = v_2$.

□

Stwierdzenie 3.4. *Ponieważ V jest macierzą ortonormalną, możemy interpretować $V^T X$ jako współrzędne dla obróconego układu. Dla $p = 2$ obrócone osie byłyby wyznaczone przez v_1 i v_2, przy czym v_1 byłby kierunkiem, w którym mamy największą zmienność danych, a v_2 prostopadłym do niego (rysunek 3.1).*

Definicja 3.2. **Całkowity rozrzut danych** dla wektora losowego X to suma wariancji jego współrzędnych: $\sum_{i=1}^{p} \mathrm{Var}(X_i)$. Wariancje poszczególnych X_i można interpretować jako ilość informacji, jaką przechowuje dana zmienna: im większa wariancja, tym lepiej możemy różnicować obserwowane wielkości.

Uwaga 3.1. Ślady macierzy Σ i Λ równają się sobie, czyli całkowite rozrzuty danych dla X i Y są równe:

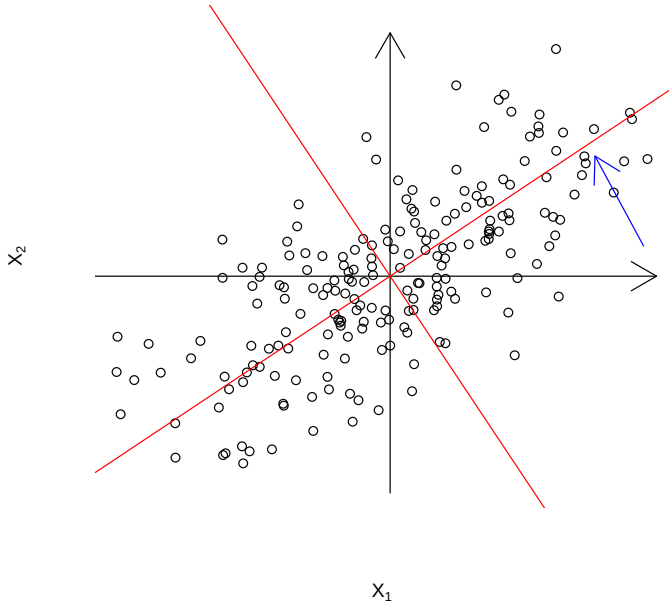
$$\sum_{i=1}^{p} \lambda_i = \mathrm{tr}(\Lambda) = \mathrm{tr}(V^T \Sigma V) = \mathrm{tr}(V^T V \Sigma) = \mathrm{tr}(\Sigma)$$

.

Istotnym parametrem diagnostycznym przy rozważaniu analizy składowych głównych jest:

$$\frac{\lambda_1 + \ldots + \lambda_k}{\lambda_1 + \ldots + \lambda_p},$$

czyli część całkowitego rozrzutu danych wyjaśniona przez k pierwszych składowych głównych. Na jego podstawie dokonuje się redukcji wymiaru danych: z p zmiennych zostaje utworzone k kombinacji liniowych tych zmiennych, które wyjaśniają np. 90% zmienności wyjściowych danych.



Rysunek 3.1: Kierunki główne wyznaczją obrócone osie układu współrzędnych.

3.2.2 Analiza składowych głównych – wersja próbkowa

Podejście próbkowe do analizy danych różni się od populacyjnego tym, że w podejściu populacyjnym do analizy brana jest zmienna losowa, a w podejściu próbkowym jej realizacje. Dlatego teraz zamiast wektora zmiennych losowych X będziemy rozpatrywać macierz jego n realizacji:

$$X = \begin{pmatrix} X_{11} & \dots & X_{1p} \\ X_{21} & \dots & X_{2p} \\ & \dots & \\ X_{n1} & \dots & X_{np} \end{pmatrix} = \begin{pmatrix} X_1^T \\ X_2^T \\ \dots \\ X_n^T \end{pmatrix}.$$

Do analizy potrzebna będzie macierz kowariancji próbkowej. Zdefiniujmy scentrowaną macierz X jako:

$$X_c = \begin{pmatrix} X_{11} - \overline{X}_{.1} & \dots & X_{1p} - \overline{X}_{.p} \\ X_{21} - \overline{X}_{.1} & \dots & X_{2p} - \overline{X}_{.p} \\ & \dots & \\ X_{n1} - \overline{X}_{.1} & \dots & X_{np} - \overline{X}_{.p} \end{pmatrix} = \begin{pmatrix} X_{c1}^T \\ X_{c2}^T \\ \dots \\ X_{cn}^T \end{pmatrix},$$

gdzie $\overline{X}_{.i} = \frac{1}{n} \sum_{j=1}^{n} X_{ji}$, $i = 1, \dots, p$.

Zauważmy, że **macierz kowariancji próbkowej** możemy wyrazić za pomocą macierzy:

$$S = \mathrm{var}(X) = \frac{1}{n-1} X_c^T X_c = \frac{1}{n-1} \sum_{i=1}^{n} X_{ci} X_{ci}^T,$$

która jest nieobciążonym estymatorem macierzy kowariancji:

$$\Sigma = \mathbb{E}[(X - \mathbb{E}X)(X - \mathbb{E}X)^T].$$

Macierz S jest symetryczna i nieujemnie określona. Znajdźmy składowe główne dla podejścia próbkowego tą samą metodą jak dla podejścia populacyjnego:

$$V^T S V = \Lambda = \frac{1}{n-1} V^T X_c^T X_c V = \frac{1}{n-1} (X_c V)^T X_c V = \frac{1}{n-1} Y^T Y.$$

Wniosek 3.1. ***Składowe główne*** *dla problemu próbkowego równe są wektorom* $[Y_1, \ldots, Y_p] = Y = X_c V$*, macierz kowariancji próbkowej dla* Y *jest równa* Λ*.*

3.2.3 Rozkład na wartości szczególne (Singular Value Decomposition)

Rozkład SVD posłuży nam do tańszej obliczeniowo konstrukcji składowych głównych w wersji próbkowej.

Twierdzenie 3.2. ***Rozkład na wartości szczególne*** *Dla dowolnej macierzy* $A \in \mathbb{R}^{m \times n}$ $m \geqslant n$, $\exists\, U$ *macierz ortonormalna* $\in \mathbb{R}^{m \times n}$ *oraz* V *macierz ortonormalna* $\in \mathbb{R}^{n \times n}$ *takie, że* $A = U \Sigma V^T$*, gdzie* $\Sigma \in \mathbb{R}^{m \times n}$ *jest macierzą diagonalną:*

$$\Sigma = \begin{pmatrix} \Sigma' & 0 \\ 0 & 0 \end{pmatrix} \quad , \quad \Sigma' = diag(\sigma_i) \ \in \mathbb{R}^{k \times k},$$

$$\sigma_1 \geqslant \ldots \geqslant \sigma_k > \sigma_{k+1} = \ldots = \sigma_n = 0,$$

gdzie k *jest rzędem macierzy* A*. Rozkład taki nazywamy szerokim rozkładem SVD, w odróżnieniu od wąskiego rozkładu SVD, w którym skracamy macierze do istotnych obliczeniowo:*

$$\begin{pmatrix} A \end{pmatrix}_{m \times n} = \begin{pmatrix} U \end{pmatrix}_{m \times m} \begin{pmatrix} \Sigma \end{pmatrix}_{m \times n} \begin{pmatrix} V \end{pmatrix}_{n \times n}^T =$$

$$= \left(\begin{pmatrix} U_1 \end{pmatrix}_{m \times k} \;\middle|\; \begin{pmatrix} U_2 \end{pmatrix}_{m \times (m-k)} \right) \begin{pmatrix} \Sigma' & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \begin{pmatrix} V_1 \end{pmatrix}_{n \times k}^T \\ \begin{pmatrix} V_2 \end{pmatrix}_{n \times (n-k)}^T \end{pmatrix} =$$

$$= U_1 \Sigma' V_1^T.$$

Dowód. Zauważmy, że macierz A^TA jest symetryczna i nieujemnie określona:

$$\forall t \in \mathbb{R}^n \quad t^T(A^TA)t = (At)^t(At) \geqslant 0.$$

Zatem, korzystając z rozkładu spektralnego dla A^TA otrzymujemy:

$$V^T(A^TA)V = \operatorname{diag}(\lambda_i) = \operatorname{diag}(\sigma_1^2, \ldots, \sigma_n^2), \tag{3.1}$$

gdzie założymy, że σ_i to nieujemne pierwiastki z λ_i:

$$\sigma_1 \geqslant \ldots \geqslant \sigma_k > \sigma_{k+1} = \ldots = \sigma_n = 0.$$

Zauważmy, że $V_1^TA^TAV_1$ jest podmacierzą $V^T(A^TA)V$ o niezerowych wyrazach na przekątnej:

$$V_1^TA^TAV_1 = \begin{pmatrix} \sigma_1^2 & & 0 \\ & \ldots & \\ 0 & & \sigma_k^2 \end{pmatrix}$$

Zdefiniujmy U_1 jako:

$$U_1 = AV_1 \operatorname{diag}(\sigma_1^{-1}, \ldots, \sigma_k^{-1}),$$

skąd otrzymujemy:

$$U_1 \operatorname{diag}(\sigma_1, \ldots, \sigma_k) = AV_1.$$

$$U_1^TU_1 = \begin{pmatrix} \sigma_1^{-1} & & 0 \\ & \ldots & \\ 0 & & \sigma_k^{-1} \end{pmatrix} \underbrace{V_1^TA^TAV_1}_{\operatorname{diag}(\sigma_1^2,\ldots,\sigma_k^2)} \begin{pmatrix} \sigma_1^{-1} & & 0 \\ & \ldots & \\ 0 & & \sigma_k^{-1} \end{pmatrix} = I_k.$$

Uzupełniamy dowolnie U_1 do ortonormalnej macierzy $n \times n$: $U = \begin{pmatrix} U_1 & | & U_2 \end{pmatrix}$. Wtedy:

$$U^TAV = \begin{pmatrix} U_1^T \\ U_2^T \end{pmatrix} A \begin{pmatrix} V_1 & | & V_2 \end{pmatrix} = \begin{pmatrix} U_1^TAV_1 & U_1^TAV_2 \\ U_2^TAV_1 & U_2^TAV_2 \end{pmatrix} =$$

ponieważ ze wzoru (3.1) wynika, że $\forall\ i$ takiego, że $v_i \in V_2$, $v_i^T(A^TA)v_i = (Av_i)^T(Av_i) = \sigma_i^2 = 0$, a norma euklidesowa wektora jest równa zero wtedy i tylko wtedy gdy wektor jest równy zero, otrzymujemy:

$$= \begin{pmatrix} \operatorname{diag}(\sigma_1^{-1}, \ldots, \sigma_k^{-1}) & \underbrace{V_1^TA^TAV_1}_{=\operatorname{diag}(\sigma_1^2,\ldots,\sigma_k^2)} & | & 0 \\ & & & \\ U_2^TU_1 \operatorname{diag}(\sigma_1, \ldots, \sigma_k) & & | & 0 \end{pmatrix} =$$

$$= \begin{pmatrix} \operatorname{diag}(\sigma_1, \ldots, \sigma_k) & | & 0_{k\times(n-k)} \\ 0_{(m-k)\times k} & | & 0_{(m-k)\times(n-k)} \end{pmatrix} = \Sigma.$$

Z równości $U^TAV = \Sigma$, ponieważ U i V są macierzami ortonormalnymi, wynika:

$$A = U\Sigma V^T = U_1\Sigma' V_1^T.$$

□

Stwierdzenie 3.5. *Wróćmy do analizy składowych głównych. Do scentrowanej macierzy danych X_c o wymiarze $n \times p$ użyjmy wąskiego rozkładu SVD i oznaczmy:*

$$X_c = U\Lambda V^T;$$

wtedy:

$$\mathrm{var}(X_c) = S = \frac{1}{n-1} X_c^T X_c = \frac{1}{n-1} V\Lambda^T \underbrace{U^T U}_{=I_p} \Lambda V^T =$$

$$= \frac{1}{n-1} V\Lambda^2 V^T = VDV^T.$$

Wniosek 3.2. *Zauważmy, że:*

1. *Składowe główne w wersji próbkowej przy użyciu rozkładu SVD:*

$$Y = X_c V = U\Lambda V^T V = U\Lambda = [\lambda_1 U_1, \ldots, \lambda_p U_p].$$

   Obliczanie składowych głównych z tego wzoru jest tańsze obliczeniowo.

2. *Widać związek pomiędzy rozkładem SVD dla $X = U\Sigma V^T$ oraz rozkładem spektralnym dla $X^T X = V\Sigma^2 V^T$.*

3. *Podobnie jest dla $XX^T = U\Sigma \underbrace{V^T V}_{=I} \Sigma U^T = U\Sigma^2 U^T$.*

3.2.4 Kolejna zaleta analizy składowych głównych

Wróćmy do analizy składowych głównych w wersji populacyjnej.

Stwierdzenie 3.6. *Przy założeniu, że wektor losowy $X \in \mathbb{R}^p$ jest scentrowany $\mathbb{E}X = 0$, możemy zapisać $Var(X) = \mathbb{E}(XX^T)$. Korzystając z rozkładu spektralnego, oznaczmy $Var(X) = \Sigma = V\Lambda V^T$. Wtedy:*

$$\forall\ k \leqslant p \quad \text{układ } v_1, \ldots, v_k \quad \text{minimalizuje} \quad \mathbb{E}||X - \sum_{i=1}^{k} (X^T a_i) a_i||^2$$

wśród wszystkich układów ortonormalnych $a_1, \ldots, a_k$.

Czyli w sensie minimalizacji błędu średniokwadratowego najlepszym k-wymiarowym przybliżeniem X jest rzut ortogonalny X na k pierwszych kierunków głównych.

Dowód.

$$\mathbb{E}(X - \sum_j (X^T a_j)a_j)^T (X - \sum_i (X^T a_i)a_i) =$$

$$= \mathbb{E}X^T X - \mathbb{E}\sum_i (X^T a_i)(a_i^T X) - \mathbb{E}\sum_i (X^T a_i)(X^T a_i) + \underbrace{\mathbb{E}(X^T a_j)a_j^T \sum_i (X^T a_i)a_i}_{a_1,\ldots,a_k \text{są ortonormalne}} =$$

$$= \mathbb{E}X^T X - \mathbb{E}\sum_i (X^T a_i)^2 - \mathbb{E}\sum_i (X^T a_i)^2 + \mathbb{E}\sum_j (X^T a_j)^2 =$$

$$= \mathbb{E}X^T X - \mathbb{E}\sum_{j=1}^{k} (X^T a_j)^2. \quad \text{To wyrażenie chcemy zminimalizować.}$$

Czyli maksymalizujemy po $a_1, \ldots, a_k$:

$$\mathbb{E}\sum_{j=1}^{k} (X^T a_j)^2 = \sum_{j=1}^{k} a_j^T \mathbb{E}(XX^T)a_j =$$

$$= \sum_{j=1}^{k} a_j^T [\sum_{i=1}^{p} \lambda_i v_i v_i^T] a_j = \sum_{j=1}^{k}\sum_{i=1}^{p} \lambda_i (v_i^T a_j)^2 = \clubsuit$$

Przyjrzyjmy się współczynnikom przy λ_i, są to kwadraty współczynników a_j w bazie ortonormalnej v_i, więc sumują się do jedynki:

$$\begin{array}{ccccccc}
(a_1^T v_1)^2 & + & (a_1^T v_2)^2 & + \ldots + & (a_1^T v_p)^2 & = 1 \\
+ & & + & & + & \\
(a_2^T v_1)^2 & + & (a_2^T v_2)^2 & + \ldots + & (a_2^T v_p)^2 & = 1 \\
\multicolumn{7}{c}{\ldots} \\
+ & & + & & + & \\
\underbrace{(a_k^T v_1)^2}_{=h_1} & + & \underbrace{(a_k^T v_2)^2}_{=h_2} & + \ldots + & \underbrace{(a_k^T v_p)^2}_{=h_p} & = 1 \\
\multicolumn{7}{c}{\ldots} \\
+ & & + & & + & \\
\underbrace{(a_p^T v_1)^2}_{=1} & + & \underbrace{(a_p^T v_2)^2}_{=1} & + \ldots + & \underbrace{(a_p^T v_p)^2}_{=1} & = 1
\end{array}$$

w każdej kolumnie można uzupełnić do bazy, wtedy suma = 1;

Czyli otrzymujemy:

$$\forall i = 1, \ldots, p \quad h_i \leqslant 1; \text{ jednocześnie } \sum_{i=1}^{k} h_i = k;$$

$$\clubsuit = \sum_{i=1}^{p} \lambda_i h_i \leqslant \lambda_1 + \lambda_2 + \ldots \lambda_k.$$

Jeśli podstawimy $a_1 = v_1, a_2 = v_2, \ldots, a_k = v_k$, otrzymujemy $h_1 = 1, \ldots, h_k = 1, h_{k+1} = 0, \ldots, h_p = 0$, dla których osiągane jest wyliczone maksimum. □

3.3 Przykłady w programie R

Analiza składowych głównych:

- dla danych Pima: http://www.mimuw.edu.pl/~pokar/StatystykaII/EKSPLORACJA/pca.R
- dla danych Iris i Kraby: http://www.mimuw.edu.pl/~pokar/StatystykaII/EKSPLORACJA/rzutDanych.R

Rozdział 4

Skalowanie wielowymiarowe

Skalowanie wielowymiarowe pozwala na redukcję wymiaru cech. Dla macierzy danych:

$$X = \begin{pmatrix} X_1^T \\ X_2^T \\ \dots \\ X_n^T \end{pmatrix}_{n \times p}$$

będziemy chcieli zrzutować „optymalnie" dane na $\mathbb{R}^k$, czyli zmniejszyć macierz X do $\hat{Z}$ o wymiarach $n \times k$, $k < p$.

Optymalność zdefiniujemy w kategoriach macierzy odległości lub podobieństwa dla n obiektów. Zadaniem będzie znalezienie optymalnej reprezentacji obiektów w $\mathbb{R}^k$.

Definicja 4.1. **Macierz odległości** to taka macierz, która spełnia własności:

$$D = (d_{ij})_{i,j=1}^n, \quad d_{ij} \geqslant 0, \quad d_{ij} = d_{ji}, \quad d_{ii} = 0.$$

Macierz podobieństwa jest macierzą konstruowaną w sposób przeciwstawny do macierzy odległości o własnościach:

$$C = (c_{ij})_{i,j=1}^n, \quad c_{ij} = c_{ji}, \quad c_{ii} \geqslant c_{ij} \ \forall i, j.$$

4.1 Metody skalowania danych

Dla macierzy danych X o wymiarach $n \times p$, zdefiniujmy D jako macierz odległości euklidesowych pomiędzy obiektami:

$$d_{ij}^2 = ||X_i - X_j||^2.$$

1. Classical multidimensional scaling:

$$(\widehat{Z}_1, \dots, \widehat{Z}_n) = \min_{Z_1,\dots,Z_n \in \mathbb{R}^k} \sum_{i \neq j} (d_{ij}^2 - ||Z_i - Z_j||^2);$$

2. Sammon scaling:

$$(\widehat{Z}_1, \ldots, \widehat{Z}_n) = \min_{Z_1,\ldots,Z_n \in \mathbb{R}^k} \frac{1}{\sum_{k\neq l} d_{kl}} \sum_{i\neq j} \frac{(d_{ij} - ||Z_i - Z_j||)^2}{d_{ij}} =$$

$$= \min_{Z_1,\ldots,Z_n \in \mathbb{R}^k} \sum_{i\neq j} \left(\frac{d_{ij} - ||Z_i - Z_j||}{d_{ij}} \right)^2 \frac{d_{ij}}{\sum_{k\neq l} d_{kl}};$$

3. Kruskal-Shepard scaling:

$$(\widehat{Z}_1, \ldots, \widehat{Z}_n) = \min_{Z_1,\ldots,Z_n \in \mathbb{R}^k} \sum_{i\neq j} (d_{ij} - ||Z_i - Z_j||)^2.$$

4.2 Własności

Niech L oznacza macierz ortogonalną $p \times p$, $L = (L_1, L_2)$, L_1 o wymiarze $p \times k$. Oznaczmy $\widehat{Z} = XL_1$, czyli rzut X na $\mathbb{R}^k$. Zdefiniujmy macierz odległości dla $\widehat{Z}$ jako $\widehat{D} = (\widehat{d}_{rs})$. Zauważmy, że:

$$d_{rs}^2 = ||X_r - X_s||^2 = ||L^T(X_r - X_s)||^2,$$

ponieważ mnożenie wektora przez macierz ortogonalną nie zmienia jego normy. Mamy więc:

$$d_{rs}^2 = \sum_{j=1}^{p} (X_{rj} - X_{sj})^2 = \sum_{j=1}^{p} [l_j^T (X_r - X_s)]^2 \geqslant \sum_{j=1}^{k} [l_j^T (X_r - X_s)]^2 = \widehat{d}_{rs}^2.$$

Stwierdzenie 4.1. *Rzut X na k pierwszych składowych głównych minimalizuje wyrażenie $\sum_{r\neq s}(d_{rs}^2 - \widehat{d}_{rs}^2)$ wśród wszystkich rzutów XL_1. Jest więc rozwiązaniem zadania* classical multidimensional scaling.

Dowód. Przyjrzyjmy się następującej macierzy:

$$\sum_{r,s=1}^{n} (X_r - X_s)(X_r - X_s)^T =$$

dla $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$,

$$= 2n \sum_{r=1}^{n} (X_r - \overline{X})(X_r - \overline{X})^T - 2 \sum_{r=1}^{n} (X_r - \overline{X}) \underbrace{\sum_{s=1}^{n} (X_s - \overline{X})^T}_{=0_p^T} =$$

$$= 2n \begin{pmatrix} \sum_{r=1}^{n} (X_{r1} - \overline{X}_{.1})^2 & \ldots & \sum_{r=1}^{n} (X_{r1} - \overline{X}_{.1})(X_{rp} - \overline{X}_{.p}) \\ \ldots & \ldots & \ldots \\ \sum_{r=1}^{n} (X_{rp} - \overline{X}_{.p})(X_{r1} - \overline{X}_{.1}) & \ldots & \sum_{r=1}^{n} (X_{rp} - \overline{X}_{.p})^2 \end{pmatrix} + 0 =$$

$$= 2nnS,$$

gdzie S jest macierzą kowariancji próbkowej (estymator obciążony).

Wróćmy do minimalizacji wyrażenia:

$$\sum_{r,s=1}^{n}(d_{rs}^2 - \widehat{d}_{rs}^2) = \underbrace{\sum_{r,s=1}^{n} d_{rs}^2}_{p \text{ współczynników}} - \underbrace{\sum_{r,s=1}^{n} \widehat{d}_{rs}^2}_{k \text{ współczynników}} =$$

$$= \underbrace{\sum_{r,s=1}^{n}\sum_{j=k+1}^{p}[l_j^T(X_r - X_s)]^2}_{\text{zostaje } p-k \text{ współczynników}}.$$

Ponieważ $\sum_{r,s=1}^{n} d_{rs}^2$ jest stałą, zadanie minimalizacji wyrażenia $\sum_{r,s=1}^{n}(d_{rs}^2 - \widehat{d}_{rs}^2)$ jest równoważne zadaniu maksymalizacji $\sum_{r,s=1}^{n} \widehat{d}_{rs}^2$. Maksymalizujemy po ortogonalnym układzie wektorów $l_1, \ldots, l_k$ wyrażenie:

$$\sum_{r,s=1}^{n}\sum_{j=1}^{k} l_j^T(X_r - X_s)(X_r - X_s)^T l_j =$$

$$= \sum_{j=1}^{k} l_j^T \left[\sum_{r,s=1}^{n}(X_r - X_s)(X_r - X_s)^T\right] l_j =$$

$$= 2n^2 \sum_{j=1}^{k} l_j^T S l_j =$$

korzystając z rozkładu spektralnego S,

$$= 2n^2 \sum_{j=1}^{k} l_j^T (\sum_{i=1}^{p} \lambda_i v_i v_i^T) l_j =$$

$$= 2n^2 \sum_{j=1}^{k}\sum_{i=1}^{p} \lambda_i (v_i^T l_j)^2.$$

Dalszy dowód przebiega analogicznie do dowodu stwierdzenia 3.6. Można zauważyć związek pomiędzy własnościami składowych głównych dla podejścia populacyjnego i próbkowego.

□

4.3 Przykłady w programie R

Skalowanie wielowymiarowe:

- porównanie skalowania wielowymiarowego i analizy składowych głównych dla danych Kraby: http://www.mimuw.edu.pl/~pokar/StatystykaII/EKSPLORACJA/mds.R
- porównanie skalowania wielowymiarowego i analizy składowych głównych dla danych Iris i Kraby: http://www.mimuw.edu.pl/~pokar/StatystykaII/EKSPLORACJA/rzutDanych.R

Rozdział 5

Test χ^2 i analiza odpowiedniości

Rozdział ten będzie dotyczył zmiennych jakościowych. Omawiane niżej testy χ^2 i niezależności opierają się na odległości χ^2 dla rozkładów dyskretnych (1.1). W obu przypadkach będziemy rozpatrywać odległości rozkładów zaobserwowanego od teoretycznego, dla których statystyka będzie asymptotycznie zbiegać do znanego rozkładu χ^2. Najpierw przyjrzyjmy się przypadkowi jednowymiarowemu.

5.1 Test χ^2 Pearsona

Dana jest zmienna losowa jakościowa o rozkładzie wielomianowym o wartościach $\{1, \ldots, k\}$ oraz prawdopodobieństwach $(p_1, \ldots, p_k)$. Podczas doświadczenia obserwujemy liczności, jakie przyjmowała zmienna w n niezależnych próbach: $n = \widehat{n}_1 + \widehat{n}_2 + \ldots + \widehat{n}_k$. Teoretyczne liczności będą wartościami oczekiwanymi dla rozkładu wielomianowego: $n_1 = np_1, \ldots, n_k = np_k$. Satystyka wyraża się wzorem:

$$Q = \sum_{i=1}^{k} \frac{(\widehat{n}_i - n_i)^2}{n_i} \xrightarrow[n\to\infty]{d} \chi^2(k-1);$$

przy spełnionej hipotezie zerowej H_0 : zmienna losowa pochodzi z rozkładu wielomianowego o parametrach $(p_1, \ldots p_k)$, zbiega ona według rozkładu do rozkładu χ^2 o $k-1$ stopniach swobody.

Omawianą statystykę można zapisać także jako:

$$Q = n \sum_{i=1}^{k} \frac{(\widehat{p}_i - p_i)^2}{p_i} = n \sum_{i=1}^{k} \left(\frac{\widehat{p}_i - p_i}{p_i} \right)^2 p_i,$$

gdzie $\widehat{p}_i = \frac{\widehat{n}_i}{n}$ to zaobserwowane prawdopodobieństwa (będące stymatorami największej wiarygodności parametrów p_i dla rozkładu wielomianowego).

5.2 Test niezależności

Będziemy rozpatrywać dwie zmienne losowe o rozkładzie dyskretnym:

$$X \text{ o wartościach } \{1, \ldots, k\}$$

$$Y \text{ o wartościach } \{1, \ldots, l\},$$

których rozkład łączny jest rozkładem wielomianowym o nieznanych parametrach p_{ij}, $i = 1, \ldots, k$, $j = 1, \ldots, l$. Znane są jedynie zaobserwowane liczności dla każdej pary (i, j) w postaci macierzy kontyngencji.

Definicja 5.1. **Macierz kontyngencji** to macierz N o wymiarach $k \times l$ zawierająca zaobserwowane liczności n_{ij}, $\sum_{i=1}^{k} \sum_{j=1}^{l} n_{ij} = n$, dla każdej z par wartości zmiennych losowych (X, Y):

$X \setminus Y$	1	...	l
1	n_{11}	...	n_{1l}
...	...	...	...
k	n_{k1}	...	n_{kl}

Hipotezę, którą będziemy testować to:

$$H_0: \text{ zmienne losowe } X \text{ i } Y \text{ są niezależne.}$$

W tym celu policzymy odległość rozkładu zaobserwowanego od teoretycznego rozumianego jako iloczyn rozkładów brzegowych. Macierz zaobserwowanych prawdopodobieństw możemy zapisać jako:

$$(\widehat{p}_{ij}) = \widehat{P} = \frac{N}{n},$$

gdzie N to macierz kontyngencji, a n to suma wszystkich elementów tej macierzy. Statystyka testowa wyraża się wzorem:

$$n \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{(\widehat{p}_{ij} - \widehat{p}_{i\cdot}\widehat{p}_{\cdot j})^2}{\widehat{p}_{i\cdot}\widehat{p}_{\cdot j}} \xrightarrow[n\to\infty]{d} \chi^2\left((k-1)(l-1)\right),$$

gdzie $\widehat{p}_{i\cdot} = \sum_{j=1}^{l} \widehat{p}_{ij}$ i $\widehat{p}_{\cdot j} = \sum_{i=1}^{k} \widehat{p}_{ij}$ to zaobserwowane rozkłady brzegowe dla X i Y.

5.3 Analiza odpowedniości (correspondence analysis)

Tak jak w poprzednim podrozdziale, dane mamy dwie zmienne losowe X i Y o rozkładzie dyskretnym. Analiza odpowiedniości to metoda prezentacji danych w przestrzeni o niewielkim wymiarze (zwykle równym 2, wtedy prezentację można przedstawić na płaszczyźnie), ilustrująca zależności pomiędzy danymi cechami X i Y.

Przykład 5.1. Rozpatrzmy następujący przykład: dla każdej osoby obserwujemy kolor oczu i włosów. Zmienna losowa X będzie oznaczać jeden z czterech kolorów oczu: brown, blue, hazel, green. Zmienna losowa Y będzie oznaczać jeden z czterech kolorów włosów: black, brown, red, blond. Tablica kontyngencji dla tego przykładu:

eyes \ hair	black	brown	red	blond
brown	68	119	26	7
blue	20	84	17	94
hazel	15	54	14	10
green	5	29	14	16

Analiza odpowiedniości pozwoli nam na przedstawienie graficzne zależności pomiędzy kolorami włosów i oczu. Na przykład, będziemy mogli zobaczyć, czy osoby o niebieskim kolorze oczu mają najczęsciej włosy koloru blond.

Do analizy korespondencji potrzebna nam będzie **macierz residuów Pearsona**, której konstrukcję omówimy przy pomocy komend programu R:

```
N=table(cbind(X,Y)) # macierz kontyngencji, gdzie X i Y to faktory
P=N/sum(N) # macierz zaobserwowanych prawdopodobieństw
Pi=apply(P,1,sum) # rozkład brzegowy dla X
Pj=apply(P,2,sum) # rozkład brzegowy dla Y
PP=Pi%*%t(Pj)
RP=sqrt(sum(N))*(P-PP)/sqrt(PP) # macierz residuów Pearsona
```

Uwaga 5.1. Zauważmy, że `sum(RP^2)` to statystyka testowa dla testu niezależności.

Przykład 5.1 Sama postać macierzy RP może nam wiele powiedzieć o zależności poszczególnych cech. Macierz residuów Pearsona dla przykładu kolory oczu i włosów znajduje się w tabeli 5.3.

Największe dodatnie wartości, a więc największą dodatnią zależność pomiędzy cechami mamy dla par (brown,black) i (blue,blond). Największą ujemne wartości, a więc największą ujemną zależność obserwujemy dla par (blue,black) i (brown,blond).

	BLACK	BROWN	RED	BLOND
Brown	4.40	1.23	-0.07	-5.85
Blue	-3.07	-1.95	-1.73	7.05
Hazel	-0.48	1.35	0.85	-2.23
Green	-1.95	-0.35	2.28	0.61

Tabela 5.1: Macierz residuów Pearsona dla przykładu kolory oczu i włosów.

Celem analizy odpowiedniości jest przedstawienie cech X i Y na płaszczyźnie, żeby widoczne były zależności między nimi. W tym celu zmniejszymy wymiar RP do 2, używając do tego analizy składowych głównych. Wiemy, że takie przybliżenie jest najlepsze w sensie błędu średniokwadratowego i opisuje możliwie najwięcej zmienności danych.

$$RP = d_1 u_1 v_1^T + d_2 u_2 v_2^T + \ldots + d_l u_l v_l^T \approx$$
$$\approx (\sqrt{d_1} u_1)(\sqrt{d_1} v_1)^T + (\sqrt{d_2} u_2)(\sqrt{d_2} v_2)^T.$$

Cechy X i Y przedstawiamy jako punkty:

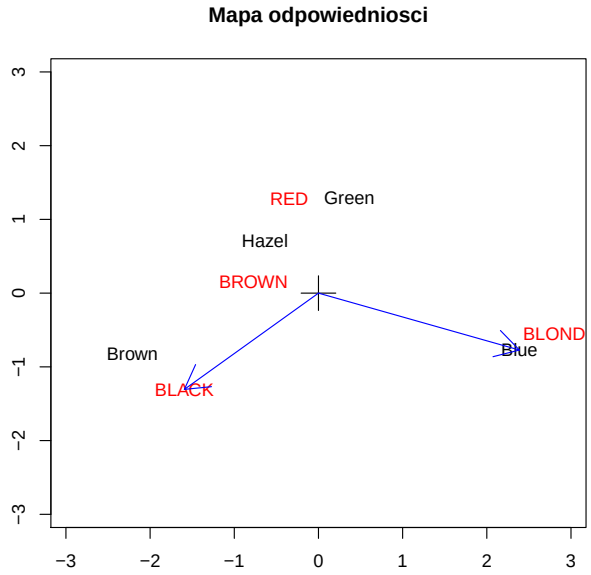
$$\widetilde{X} = [\sqrt{d_1} u_1, \sqrt{d_2} u_2]$$
$$\widetilde{Y} = [\sqrt{d_1} v_1, \sqrt{d_2} v_2].$$

Pamiętamy z rozkładu SVD, że $(RP)V = UD$, $U^T(RP) = DV^T$, U rozpina przestrzeń kolumn macierzy RP, a V przestrzeń wierszy. Scentrowane punkty (od kolumn macierzy $\widetilde{X}$ i $\widetilde{Y}$ odejmujemy średnie w kolumnach tak żeby środek danych był w (0,0)) nanosimy na wykres (rysunek 5.1).

Przykład 5.1 Dla interpretacji mapy odpowiedniości, potraktujmy wiersze macierzy $\widetilde{X}$ oraz $\widetilde{Y}$ jako współrzędne wektorów, zaczepionych w punkcie (0,0). Wiersze macierzy $\widetilde{X}$ odpowiadają kolorom oczu, wiersze macierzy $\widetilde{Y}$ kolorom włosów. Na rysunku zaznaczone zostały dla przykładu wektory odpowiadające cechom blue oraz black. Zauważmy, że iloczyn skalarny dwóch wektorów, i-tego z macierzy $\widetilde{X}$ i j-tego z macierzy $\widetilde{Y}$ równy jest przybliżeniu macierzy residuów Pearsona. Oznaczmy:

$$a = \widetilde{X}[i,] = (\sqrt{d_1} u_1[i], \sqrt{d_2} u_2[i]);$$
$$b = \widetilde{Y}[i,] = (\sqrt{d_1} v_1[i], \sqrt{d_2} v_2[i]);$$
$$RP \approx a^T b = a \circ b = |a||b| \cos(\vartheta),$$

gdzie θ oznacza kąt pomiędzy wektorami. Interpretacja dla wektorów blue i black może być następująca: ponieważ długości obu wektorów są duże oraz $\cos(\vartheta)$ jest ujemne o wartości bezwzględnej w przybliżeniu $\frac{1}{2}$, zależność pomiędzy cechami jest silnie ujemna. Na tej samej zasadzie możemy zaobserwować silną zależność pomiędzy włosami blond i oczami blue oraz włosami black i oczami brown.



Rysunek 5.1: Analiza odpowiedniości dla przykładu kolory oczu i włosów.

5.4 Przykłady w programie R

Analiza odpowiedniości oraz konstrukcja testu χ^2 dla danych kolor oczu i włosów:

- http://www.mimuw.edu.pl/~pokar/StatystykaII/EKSPLORACJA/CorrespondenceAnaly
  corresp.r
- http://www.mimuw.edu.pl/~pokar/StatystykaII/EKSPLORACJA/CorrespondenceAnaly
  corresp0.r

Rozdział 6

Klasteryzacja

Klasteryzacja jest, podobnie jak analiza składowych głównych, metodą redukcji wymiaru danych. W tym przypadku jednak redukcja będzie się odbywać w pionie a zamiast odcinania części danych, będziemy je grupować. Nowym wymiarem danych będzie liczba grup. Dla macierzy $X_{n\times p}$ będziemy szukać optymalnego podziału na K części, czyli szukać podziału C na K grup:

$$\{1,\dots,n\} = C_1 \cup \dots \cup C_K,$$

parami rozłącznych o licznościach odpowiednio $n_1,\dots,n_K$. Będziemy używać oznaczenia X^k na podmacierz X o indeksach z C_k, $k = 1,\dots,K$.

6.1 Klasteryzacja K-średnich

Klasteryzacji K-średnich używamy, gdy znamy ilość grup K, na ile chcemy podzielić dane. Zdefiniujmy następujące macierze:

Macierz wariancji całkowitej:

$$T = \mathrm{var}(X) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(X_i - \overline{X})^T,$$

nie zależy od podziału C, var oznacza próbkową macierz kowariancji.

Zmienność całkowita to ślad macierzy T: $\mathrm{tr}(T)$.

Macierz wariancji wewnątrzgrupowej:

$$W_C = \sum_{k=1}^{K}\frac{n_k}{n}\mathrm{var}(X^k) = \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in C_k}(X_i - \overline{X^k})(X_i - \overline{X^k})^T,$$

zależy od podziału C.

Zmienność wewnątrzgrupowa to ślad macierzy W_C: $\mathrm{tr}(W_C)$.

Macierz wariancji międzygrupowej:

$$B_C = \sum_{k=1}^{K} \frac{n_k}{n} (\overline{X^k} - \overline{X})(\overline{X^k} - \overline{X})^T,$$

zależy od podziału C.

Zmienność międzygrupowa to ślad macierzy B_C: $\mathrm{tr}(B_C)$.

$\overline{X^k}$ oznacza p-wymiarowy wektor średnich kolumnowych dla macierzy X^k, a $\overline{X}$ p-wymiarowy wektor średnich kolumnowych dla całej macierzy X. $\overline{X^k}$ nazywane są centroidami, redukcja wymiaru polega na zastępowaniu grup danych przez ich centroidy.

Stwierdzenie 6.1.

$$T = W_C + B_C \quad \forall \textit{ podziału } C.$$

Dowód.

$$T = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})(X_i - \overline{X})^T =$$

$$= \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in C_k} (X_i - \overline{X^k} + \overline{X^k} - \overline{X})(X_i - \overline{X^k} + \overline{X^k} - \overline{X})^T =$$

$$= \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in C_k} (X_i - \overline{X^k})(X_i - \overline{X^k})^T + \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in C_k} (\overline{X^k} - \overline{X})(\overline{X^k} - \overline{X})^T +$$

$$+ \underbrace{\frac{1}{n} \sum_{k=1}^{K} \sum_{i \in C_k} (X_i - \overline{X^k})(\overline{X^k} - \overline{X})^T}_{=0} + \underbrace{\frac{1}{n} \sum_{k=1}^{K} \sum_{i \in C_k} (\overline{X^k} - \overline{X})(X_i - \overline{X^k})^T}_{=0} =$$

$$= \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in C_k} (X_i - \overline{X^k})(X_i - \overline{X^k})^T + \sum_{k=1}^{K} \frac{n_k}{n} (\overline{X^k} - \overline{X})(\overline{X^k} - \overline{X})^T =$$

$$= W_C + B_C.$$

□

Wniosek 6.1.

$$tr(T) = tr(W_C) + tr(B_C).$$

Czyli

zmienność całkowita = zmienność wewnątrzgrupowa + zmienność międzygrupowa.

Ideą klasteryzacji K-średnich jest minimalizacja po podziałach zmienności wewnątrzgrupowej, co jest jednoznaczne z maksymalizcją zmienności międzygrupowej:

$$C_{opt} = \min_C \operatorname{tr}(W_C).$$

$$\operatorname{tr}(W_C) = \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in C_k}\sum_{j=1}^{p}(X_{ij} - \overline{X_j^k})^2 = \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in C_k}||X_i - \overline{X^k}||^2.$$

Idea zachłannego algorytmu K-średnich (zależnego od wybranego podziału startowego C), z którego można korzystać np. w programie R wygląda następująco:

Algorytm K-średnich

Wielokrotnie powtarzamy przy różnym podziale startowym C:
repeat
 for (k **in** $1 \dots K$)
  $m_k = \operatorname{argmin}_m \sum_{i\in C^k} ||X_i - m||^2 = \frac{1}{n_k}\sum_{i\in C^k} X_i$
 for (i **in** $1 \dots n$)
  $i \in C^k \Leftrightarrow k = \operatorname{argmin}_l ||X_i - m_l||^2$
 until warunek stopu

Przykładowy wynik algorytmu klasteryzacji K-średnich znajduje się na rysunku 6.1.

6.2 Klasteryzacja K-medoidów

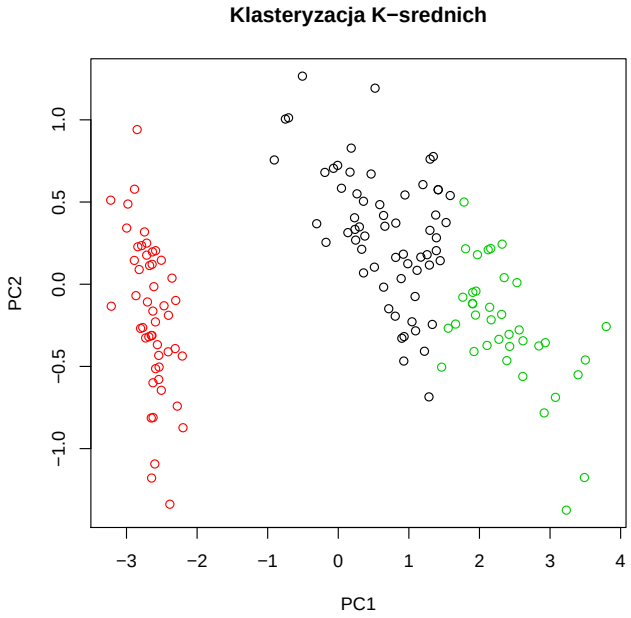
Klasteryzacja K-medoidów jest podobna do klasteryzacji K-średnich, z tą różnicą, że zamiast średnich arytmetycznych w algorytmie będziemy używać median. Dzięki takiemu sformułowaniu, możemy go używać przy dowolnej macierzy odległości między obiektami $D = (d_{ij})_{i,j=1}^n$.

Algorytm K-medoidów

Wielokrotnie powtarzamy przy różnym podziale startowym C:
repeat
 for (k **in** $1 \dots K$)
  $m_k = \operatorname{argmin}_m \sum_{i\in C^k} d_{im}$ # m jako mediana, należy do zbioru obserwacji
 for (i **in** $1 \dots n$)
  $i \in C^k \Leftrightarrow k = \operatorname{argmin}_l d_{im_l}$
 until warunek stopu

6.3 Klasteryzacja hierarchiczna

Używając klasteryzacji hierarchicznej nie zakładamy z góry ilości klastrów, na jakie chcemy podzielić dane. Wychodzimy od sytuacji, gdy mamy n kla-



Rysunek 6.1: Klasteryzacja K-średnich dla danych Iris, $K = 3$. Żeby można było przedstawić wyniki na płaszczyźnie, został zmniejszony wymiar danych poprzez analizę składowych głównych.

strów, czyli każda obserwacja jest oddzielną grupą. W każdym kroku algorytmu łączymy 2 klastry, czyli zmniejszamy ich liczbę o jeden i tak aż do połączenia wszystkich obserwacji w jedną grupę. Wybór ilości klastrów opieramy na wykresie separowalności, która obliczana jest dla każdego kroku algorytmu.

W klasteryzacji hierarchicznej możemy używać różnych metod aglomeracji danych. Dla macierzy odległości $D = (d_{ij})_{i,j=1}^{n}$ odległość dwóch klastrów G i H od siebie przy założeniach

$$G, H \subseteq \{1, \ldots, n\} \quad , \quad G \cap H = \emptyset,$$

możemy zdefiniować jako:

1. **Single linkage**
$$d_{G,H} = \min_{i \in G, j \in H} d_{ij}.$$

2. **Average linkage**
$$d_{G,H} = \frac{1}{|G||H|} \sum_{ij} d_{ij},$$
gdzie $|\cdot|$ oznacza liczność zbioru.

3. **Complete linkage**

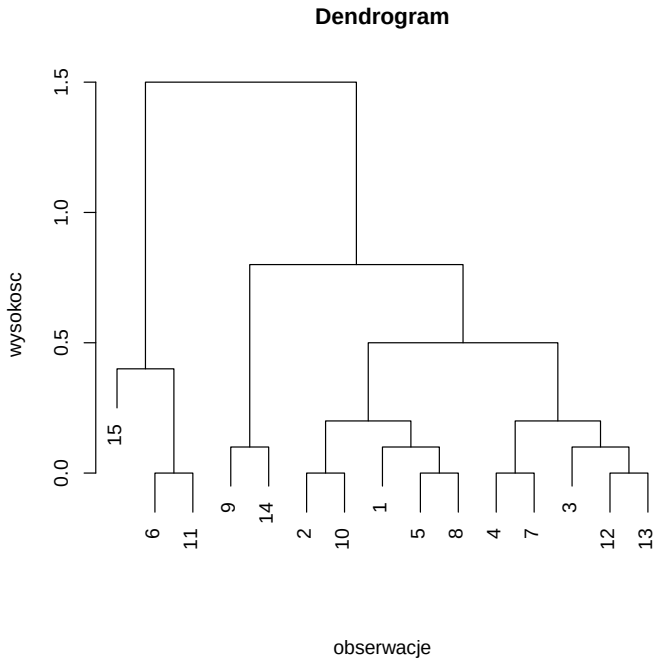
$$d_{G,H} = \max_{i \in G, j \in H} d_{ij}.$$

Ideę algorytmu klasteryzacji hierarchicznej możemy zapisać jako:

Algorytm klasteryzacji hierarchicznej

$C = \{1\}, \{2\}, \ldots, \{n\}$
for (l **in** 1:(n−1))
 połącz najbliższe dwa klastry:
  $(i_*, j_*) =$ argmin$_{i,j:i<j} d_{ij}$
  klastry i_* oraz j_* zastąp przez 0
 odnów macierz odległości $d_{0,k} = \min(d_{i_*k}, d_{j_*k})$

Definicja 6.1. **Dendrogram** jest metodą ilustracji wyników klasteryzacji hierarchicznej. Możemy obserwować od dołu dendrogramu (rysunek 6.2) jak kolejne klastry się łączą i dla jakiej wysokości (odległości klastrów) to zachodzi.



Rysunek 6.2: Przykładowy dendrogram dla klasteryzacji hierarchicznej.

Definicja 6.2. Oznaczmy h_k jako minimalną wysokość, na której obserwujemy podział na k części. Na przykład, na obrazku 6.2 dla $k = 5$ $h_5 \approx 0,2$.

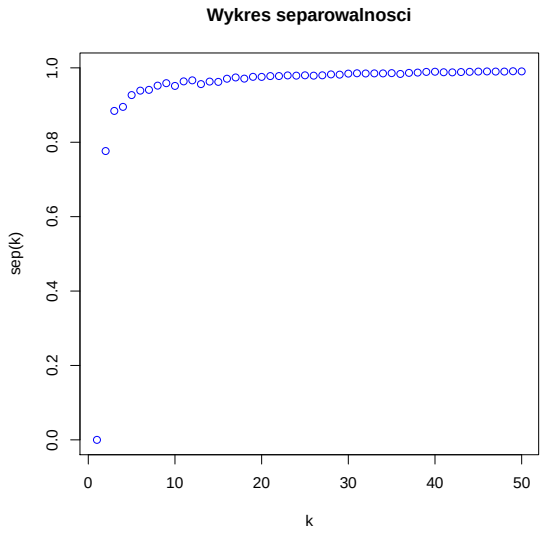
Separowalność dla klasteryzacji hierarchicznej definiujemy jako:

$$\text{sep}(k) = 1 - \frac{h_k}{h_1}.$$

Z definicji separowalności możemy wywnioskować następujące własności:

- separowalność przyjmuje wartości z przedziału $[0, 1]$;
- jest niemalejącą funkcją liczby klastrów.

Przykładowy wykres separowalności znajduje się na rysunku 6.3. Na podstawie tego wykresu podejmuje się decyzję dotyczącą optymalnej ilości klastrów. Szukamy takiego k, żeby $sep(k) - sep(k-1)$ było duże w stosunku do $sep(k+1) - sep(k)$. Chcemy znaleźć taką niewielką liczbę klastrów, żeby zysk mierzony separowalnością przy łączeniu klastrów w danym kroku był duży, a dalsze sklejanie grup nie dawało już takich korzyści. Graficznie sprowadza się to do szukania „kolanka" funkcji separowlaności. Jednym ze sposobów jest szukanie punktu na wykresie najbliższego punktowi $(0, 1)$. Przykładowo, na rysunku 6.3 optymalnym wyborem jest $k = 3$ ($k = 5$ też jest dobrym wyborem, chociaż dążymy do tego aby jak najbardziej zredukować wymiar danych, czyli wybrać jak najmniejsze k).



Rysunek 6.3: Przykładowy wykres separowalności dla danych Iris.

Definicja 6.3. Można zdefiniować także separowalność dla klasteryzacji K-średnich. Oznaczmy:

$$t = \text{tr}(T);$$

$$w_k = \min_C \operatorname{tr}(W_C), \quad k \text{ jest liczbą klastrów};$$
$$t = w_k + w_b.$$

Ponieważ wiemy, że:

$$1 = \frac{w_k}{t} + \frac{b_k}{t}, \quad w_1 = t, \quad w_n = 0,$$

możemy zdefiniować separowalność jako:

$$\operatorname{sep}(k) = 1 - \frac{w_k}{t}.$$

Stwierdzenie 6.2. *Separowalność dla klasteryzacji K-średnich jest niemalejącą funkcją k, liczby klastrów. Funkcja w_k jest więc nierosnąca ze względu na liczbę klastrów.*

Dowód. Jako praca domowa. □

6.4 Przykłady w programie R

Klasteryzacja:

- k-średnich i hierarchiczna na danych Kraby, kobiety Pima i Irysy: http://www.mimuw.edu.pl/~pokar/StatystykaII/EKSPLORACJA/ileKlastrow.R
- wybór liczby klastrów na podstawie wykresu separowalności i sylwetki dla algorytmów k-średnich i k-medoidów: http://www.mimuw.edu.pl/~pokar/StatystykaII/EKSPLORACJA/sep_syl.r
- k-średnich i hierarchiczna zobrazowane przy pomocy analizy składowych głównych: http://www.mimuw.edu.pl/~pokar/StatystykaII/EKSPLORACJA/pca.R
- k-średnich i hierarchiczna oraz PCA i skalowanie wielowymiarowe dla danych Iris i Kraby: http://www.mimuw.edu.pl/~pokar/StatystykaII/EKSPLORACJA/rzutDanych.R

Rozdział 7

Klasyfikacja

Zadanie klasyfikacji polega na konstrukcji funkcji (klasyfikatora), która na podstawie zaobserwowanych cech będzie przydzielała obserwację do którejś z wcześniej zdefiniowanych grup. Do estymacji funkcji potrzebne są obserwacje, które już zostały sklasyfikowane, będziemy je nazywać próbą uczącą:

$$(y_i, X_i) \quad i \in \{1, \ldots, n\} \text{ niezależne obserwacje, } X_i \in \mathbb{R}^p \quad y_i \in \{1, \ldots, K\}.$$

Dane X_i oznaczają zaobserwowane cechy, y_i grupę, do której obserwacja została zaklasyfikowana.

C^j oznaczaja zbiór tych indeksów i, że $y_i = j$; $C^1, \ldots, C^K$ są rozłączne o licznościach odpowiednio $n_1, \ldots, n_K$.

Funkcja wiarygodności dla opisanych danych wyraża się wzorem:

$$L(\pi_1, \ldots, \pi_K, \theta_1, \ldots, \theta_K) = \prod_{i=1}^{n} g(y_i, X_i) = \prod_{i=1}^{n} \pi_{y_i} f_{\theta_{y_i}}(X_i)$$

$$\text{pod warunkiem: } \sum_{k=1}^{K} \pi_k = 1.$$

Logwiarygodność to logarytm funkcji wiarygodności:

$$l = \log L;$$

$$l = \sum_{i=1}^{n} \log \pi_{y_i} + \sum_{i=1}^{n} \log f_{\theta_{y_i}}(X_i).$$

Zadanie klasyfikacji dzielimy na dwa kroki:

1. Estymujemy parametry $\pi_1, \ldots, \pi_K$ oraz $\theta_1, \ldots, \theta_K$ na podstawie zaobserwowanych par (y_i, X_i) przy użyciu metody największej wiarygodności. Parametry π_k możemy interpretować jako prawdopodobieństwa

przynależności do danej grupy danych, a θ_k jako parametry rozkładu w danej grupie (na przykład dla wielowymiarowego rozkładu normalnego, byłyby to średnia μ i macierz kowariancji Σ).

2. Obserwujemy nowe cechy X_{n+1} i przyporządkowujemy im $\widehat{y}_{i+1}$ na podstawie zbudowanego przez nas klasyfikatora. Będziemy go także nazywać regułą decyzyjną.

Maksymalizujemy funkcję wiarygodności pod warunkiem $\sum_{k=1}^{K} \pi_k = 1$ przy użyciu metody mnożników Lagrange'a:

$$\max_{\pi_1,\ldots,\pi_K,\theta_1,\ldots,\theta_K} F(\pi_1,\ldots,\pi_K,\theta_1,\ldots,\theta_K) =$$

$$\max_{\pi_1,\ldots,\pi_K,\theta_1,\ldots,\theta_K} n_1 \log \pi_1 + \ldots + n_K \log \pi_K - \lambda(\sum_{k=1}^{K} \pi_k - 1) +$$

$$+ \sum_{i \in C^1} \log f_{\theta_1}(X_i) + \ldots \sum_{i \in C^K} \log f_{\theta_K}(X_i).$$

Liczymy estymatory $\widehat{\pi}_1, \ldots, \widehat{\pi}_K$:

$$\frac{\partial F}{\partial \pi_k} = \frac{n_k}{\pi_k} - \lambda = 0 \quad \forall k = 1 \ldots, K; \tag{7.1}$$

$$\frac{\partial F}{\partial \lambda} = \sum_{k=1}^{K} \pi_k - 1 = 0; \tag{7.2}$$

Z równań 7.1 otrzymujemy:

$$\forall k = 1 \ldots, K \quad \frac{n_k}{\lambda} = \pi_k.$$

Sumujemy po k korzystając z równania 7.2 :

$$\frac{n}{\lambda} = 1 \quad \Rightarrow \quad \lambda = n \quad \Rightarrow \quad \widehat{\pi}_k = \frac{n_k}{n}.$$

Estymację parametrów θ_k odłożymy do dalszej części wykładu.

7.1 Optymalna reguła decyzyjna

Zobaczmy teraz, jak można zdefiniować optymalny klasyfikator w zależności od funkcji straty karzącej za błędne sklasyfikowanie danych.

Definicja 7.1. **Funkcja straty** to funkcja przyporządkowująca nieujemną wielkość kary poprzez porównanie prawdy (założymy chwilowo, że ją znamy) do podjętej decyzji (wyliczonego estymatora):

$$L : \underbrace{K}_{prawda} \times \underbrace{K}_{decyzja} \to \underbrace{\mathbb{R}^+}_{kara}.$$

Przykładową funkcją kary dla ciągłego y jest $L(y, \widehat{y}) = (y - \widehat{y})^2$.

Mając wyestymowaną regułę decyzyjną sensownym jest rozpatrywanie średniej straty dla naszego klasyfikatora:

Definicja 7.2. **Ryzyko reguły decyzyjnej** dla $d : \mathbb{R}^p \to K$:

$$\text{Ryzyko} = \text{średnia strata reguły decyzyjnej } d =$$

$$= \mathrm{R}(d) = \sum_{y \in K} \int_{\mathbb{R}^p} L(y, d(X)) g(y, X) dX =$$

gdzie $g(y, X)$ jest gęstością łącznego rozkładu danych. Z twierdzenia o prawdopodobieństwie całkowitym:

$$= \sum_{k=1}^{K} \left[\int_{\mathbb{R}^p} L(k, d(X)) f(X|k) dX \right] \pi_k.$$

Definicja 7.3. **Optymalna reguła decyzyjna** d_* to taka reguła decyzyjna, że

$$\forall d \quad \mathrm{R}(d_*) \leqslant \mathrm{R}(d).$$

Definicja 7.4. **Reguła bayesowska** $d_B(X)$ to reguła decyzyjna, która lokalnie dla danego X spełnia warunek:

$$d_B(X) = \operatorname{argmin}_{1 \leqslant l \leqslant K} \mathbb{E}_{y|X} L(y, l) =$$

$$= \operatorname{argmin}_l \sum_{k=1}^{K} L(k, l) p(k|X) =$$

ze wzoru Bayesa:

$$= \operatorname{argmin}_l \left[\sum_{k=1}^{K} L(k, l) \frac{\pi_k f(X|k)}{\sum_{s=1}^{K} \pi_s f(X|s)} \right] =$$

$$= \operatorname{argmin}_l \left[\sum_{k=1}^{K} L(k, l) \pi_k f(X|k) \right].$$

Stwierdzenie 7.1.

$$R(d_B) = R(d_*).$$

Reguła bayesowska jest optymalną regułą decyzyjną.

Dowód. Dla dowolnej reguły decyzyjnej d zachodzi:

$$\mathrm{R}(d) = \sum_{k=1}^{K} \left[\int_{\mathbb{R}^p} L(k, d(X)) f(X|k) dX \right] \pi_k =$$

$$= \int_{\mathbb{R}^p} \left[\sum_{k=1}^{K} L(k, d(X)) \pi_k f(X|k) \right] dX \geqslant$$

$$\geqslant \int_{\mathbb{R}^p} \left[\min_{1 \leqslant l \leqslant k} \sum_{k=1}^{K} L(k,l) \pi_k f(X|k) \right] dX =$$

$$= \int_{\mathbb{R}^p} \left[\sum_{k=1}^{K} L(k, d_B(X)) \pi_k f(X|k) \right] dX = \mathrm{R}(d_B).$$

□

7.2 Wielowymiarowy rozkład normalny

W dalszej części wykładu będziemy zakładać, że f_{θ_i} mają rozkłady normalne. Dlatego przyjrzyjmy się bliżej własnościom wielowymiarowego rozkładu normalnego i estymacji jego parametrów metodą największej wiarygodności.

Definicja 7.5. Wektor losowy $X = (x_1, \ldots, x_p)$ ma **rozkład wielowymiarowy normalny** w $\mathbb{R}^p$ jeśli $\forall\ u \in \mathbb{R}^p$ $u^T X$ ma rozkład normalny w $\mathbb{R}$. Oznaczmy ten rozkład poprzez $\mathcal{N}(\mu, \Sigma)$, gdzie $\mu = \mathbb{E}x$, $\Sigma = \mathrm{Var}(X)$.

Twierdzenie 7.1. *Jeżeli X ma rokład normalny w $\mathbb{R}^p$, to $\forall\ a \in \mathbb{R}^k$ i macierzy A wymiaru $k \times p$, $AX + a$ ma rozkład normalny w $\mathbb{R}^k$.*

Dowód.

$$\forall\ u \in \mathbb{R}^k \quad u^T(AX + a) = (u^T a)X + u^T a.$$

□

Wniosek 7.1. *Rozkłady brzegowe wielowymiarowego rozkładu normalnego są normalne w odpowiednich podprzestrzeniach $\mathbb{R}^p$.*

Twierdzenie 7.2. *Fcja charakterystyczna zmiennej losowej X o rozkładzie normalnym w $\mathbb{R}^p$ jest postaci:*

$$\varphi_X(t) = e^{it^T \mu - \frac{1}{2} t^T \Sigma t}. \tag{7.3}$$

Także na odwrót: jeżeli Σ jest symetryczną macierzą dodatnio określoną o wymiarach $p \times p$, to φ_X określona w równaniu 7.3 jest funkcją charakterystyczną wektora losowego o rozkładzie normalnym w $\mathbb{R}^p$.

Wniosek 7.2. *Dowolna macierz symetryczna dodatnio określona o wymiarach $p \times p$ jest macierzą kowariancji wektora losowego o rokładzie normalnym w $\mathbb{R}^p$.*

Twierdzenie 7.3. *Gęstość wielowymiarowego rozkładu normalnego $\mathcal{N}(\mu, \Sigma)$:*

$$f(X) = \frac{1}{(2\pi)^{\frac{p}{2}} \underbrace{(det(\Sigma))^{\frac{1}{2}}}_{=|\Sigma|}} \exp\left[-\frac{1}{2} (X - \mu)^T \Sigma^{-1} (X - \mu) \right].$$

Twierdzenie 7.4. *Jeżeli X ma rozkład normalny w $\mathbb{R}^p$: $\mathcal{N}(\mu, \Sigma)$, to współrzędne wektora X są niezależne $\Leftrightarrow$ Σ jest diagonalna. Dla rozkładu normalnego brak korelacji oznacza niezależność.*

Twierdzenie 7.5. *Jeżeli $X \sim \mathcal{N}(\mu, \sigma^2\mathbb{I})$, C jest macierzą ortonormalną o wymiarach $p \times p$, to:*

$$CX \sim \mathcal{N}(C\mu, C\sigma^2\mathbb{I}C^T) = \mathcal{N}(C\mu, \sigma^2 \underbrace{CC^T}_{=\mathbb{I}}) = \mathcal{N}(C\mu, \sigma^2\mathbb{I}).$$

7.2.1 Estymatory największej wiarygodności dla rozkładu normalnego $\mathcal{N}(\mu, \Sigma)$

Niech $X_1, \ldots, X_n$ będą niezależnymi wektorami losowymi z p-wymiarowego rozkładu $\mathcal{N}(\mu, \Sigma)$. Znajdziemy estymatory dla parametrów μ i Σ. Łączna funkcja wiarygodności dla n wektorów losowych:

$$L(\mu, \Sigma) = \prod_{i=1}^{n} \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(X_i - \mu)^T \Sigma^{-1} (X_i - \mu)\right].$$

Najpierw szukamy estymatora $\widehat{\mu}$; w tym celu opuszczamy wszystkie wyrazy nie zależeżące od μ, które by się wyzerowały po policzeniu pochodnej. Dla prostoty obliczeń maksymalizujemy podwojoną logwiarygodność:

$$2\log(L(\mu)) = 2l(\mu) = n\mu^T\Sigma^{-1} - 2n\mu^T\Sigma^{-1}\overline{X},$$

gdzie $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.
Przypomnijmy fakt:

Lemat 7.1. *Oznaczmy: a, b – wektory tej samej długości p, A macierz o wymiarach $p \times p$.*

$$\frac{\partial(a^Tb)}{\partial a} = \frac{\partial(b^Ta)}{\partial a} = b.$$

$$\frac{\partial(b^TAb)}{\partial b} = \underbrace{(A + A^T)b}_{=2Ab \text{ jeśli } A \text{ symetryczna}}.$$

Skorzystajmy z lematu 7.1 żeby obliczyć pochodną logwiarygodności:

$$\frac{1}{n}\frac{\partial(2l(\mu))}{\partial\mu} = 2\Sigma^{-1}\mu - 2\Sigma^{-1}\overline{X} = 0,$$

stąd

$$\widehat{\mu} = \overline{X},$$

czyli estymatorem największej wiarygodności dla średniej rozkładu normalnego jest średnia arytmetyczna obserwacji.

Ponieważ optymalne $\widehat{\mu}$ nie zależy od Σ, przy obliczaniu $\widehat{\Sigma}$ możemy wstawić $\overline{X}$ za μ. Maksymalizujemy po Σ wyrażenie:

$$L(\overline{X}, \Sigma) \propto |\Sigma|^{-\frac{n}{2}} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}(X_i - \overline{X})^T\Sigma^{-1}(X_i - \overline{X}).\right]$$

Symbol $\propto$ oznacza proporcjonalność, możemy opuścić wszystkie stałe, które nie wpływają na wynik optymalizacji.
Ponieważ $(X_i - \overline{X})^T\Sigma^{-1}(X_i - \overline{X})$ jest liczbą, a tr(liczba) = liczba, oraz $\mathrm{tr}(AB) = \mathrm{tr}(BA)$, otrzymujemy:

$$L(\overline{X}, \Sigma) \propto |\Sigma|^{-\frac{n}{2}} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}\mathrm{tr}\{(X_i - \overline{X})(X_i - \overline{X})^T\Sigma^{-1}\}\right] =$$

ślad macierzy jest funkcją liniową argumentu, więc zachodzi:

$$= |\Sigma|^{-\frac{n}{2}} \exp\left[-\frac{1}{2}\mathrm{tr}\{\underbrace{(X_i - \overline{X})(X_i - \overline{X})^T}_{=S}\Sigma^{-1}\}\right] =$$

pomnóżmy i podzielmy przez $|S|^{\frac{n}{2}}$

$$= |S|^{-\frac{n}{2}}|S\Sigma^{-1}|^{\frac{n}{2}} \exp\left[-\frac{1}{2}\mathrm{tr}\{S\Sigma^{-1}\}\right].$$

Ponieważ $|S|^{-\frac{n}{2}}$ nie zależy od Σ, możemy to wyrażenie opuścić. Podstawmy $B = S\Sigma^{-1}$:

$$L(\overline{X}, B) \propto |B|^{\frac{n}{2}} \exp\left[-\frac{1}{2}\mathrm{tr}B\right].$$

Lemat 7.2. *Dla macierzy kwadratowej A o wymiarach $p \times p$ zachodzi:*

$$det(A) = \prod_{i=1}^{p}\lambda_i,$$

$$tr(A) = \sum_{i=1}^{p}\lambda_i,$$

gdzie λ_i to wartości własne macierzy.

Korzystając z lematu 7.2:

$$|B|^{\frac{n}{2}} \exp\left[-\frac{1}{2}\mathrm{tr}B\right] = \prod_{j=1}^{p}\lambda_j^{\frac{n}{2}}e^{-\frac{1}{2}\lambda_j}$$

Zmaksymalizujmy to wyrażenie po każdej wartości własnej λ_j, co sprowadza się do maksymalizacji po λ funkcji:

$$F(\lambda) = \lambda^{\frac{n}{2}}e^{-\frac{1}{2}\lambda};$$

$$\log F(\lambda) = \frac{n}{2}\log\lambda - \frac{1}{2}\lambda;$$
$$\frac{\partial(\log F(\lambda))}{\partial\lambda} = \frac{n}{2\lambda} - \frac{1}{2} = 0;$$

skąd $\widehat{\lambda} = \widehat{\lambda_1} = \ldots = \widehat{\lambda_p} = n$.
Macierzą o wszystkich wartościach własnych równych n jest $n\mathbb{I}$:

$$B = S\Sigma^{-1} = n\mathbb{I},$$

skąd:

$$\widehat{\Sigma} = \frac{1}{n}S = \frac{1}{n}(X_i - \overline{X})(X_i - \overline{X})^T,$$

czyli estymatorem największej wiarygodności dla macierzy kowariancji rozkładu normalnego jest obciążony estymator próbkowy macierzy kowariancji.

7.3 Klasyfikacja w modelu normalnym

Zrobimy dwa założenia dotyczące rozważanego wcześniej klasyfikatora:

1. Funkcja straty jest postaci:

$$L(k,l) = \mathbf{1}_{k\neq l}.$$

2. W każdej z grup dane pochodzą z rozkładu normalnego, czyli f_{θ_k} to gęstość rozkładu normalnego, $\theta_k = (\mu_k, \Sigma_k)$.

Dal zadanej funkcji straty optymalna (bayesowska) reguła decyzyjna będzie miała postać:

$$d_B(X) = \operatorname{argmin}_l \left[\sum_{k=1}^{K} L(k,l)\pi_k f(X|k)\right] = \operatorname{argmin}_l \left[\sum_{k=1}^{K} \mathbf{1}_{k\neq l}\pi_k f(X|k)\right] =$$

$$= \operatorname{argmin}_l \left[\underbrace{\sum_{k=1}^{K} \pi_k f(X|k)}_{\text{nie zależy od wyboru } l} - \pi_l f(X|l)\right] = \operatorname{argmax}_l \left[\pi_l f(X|l)\right].$$

Znamy już postać szukanego klasyfikatora, potrzebujemy jeszcze estymatorów dla występujących w nim parametrów. Wiemy jak wyglądają estymatory $\widehat{\pi}_k$:

$$\widehat{\pi}_k = \frac{n_k}{n}, \quad n_k = \sum_{i=1}^{n} \mathbf{1}(y_i = k).$$

Estymatory największej wiarygodności dla parametrów θ_k przy założeniu normalności rozkładów w grupach są postaci:

$$\widehat{\mu_k} = \frac{1}{n_k} \sum_{i \in C^k} X_i = \frac{\sum_{i=1}^n X_i \mathbf{1}(y_i = k)}{\sum_{i=1}^n \mathbf{1}(y_i = k)};$$

$$\widehat{\Sigma_k} = \frac{1}{n_k} \sum_{i \in C^k} (X_i - \overline{X^k})(X_i - \overline{X^k})^T,$$

gdzie $\overline{X^k}$ oznacza wektor średnich obserwacji dla $X_i \in C^k$.

Dla X niezależnego od próby uczącej: $(y_1, X_1), \ldots, (y_n, X_n)$ estymator reguły decyzyjnej ma postać:

$$\widehat{d}(X) = \text{argmax}_{1 \leqslant l \leqslant K} \widehat{\pi}_l \left[f_{\widehat{\mu}_l, \widehat{\Sigma}_l}(X) \right].$$

7.3.1 Kwadratowa (qda) i liniowa (lda) funkcja klasyfikacyjna

W zależności od założeń dotyczących parametrów, możemy otrzymać klasyfikator będący różną funkcją swojego argumentu X: albo kwadratową albo liniową.

Kwadratowa funkcja klasyfikacyjna (qda) nie wymaga dodatkowych założeń o parametrach:

$$d(X) = \text{argmax}_l \left[\pi_l f_{\mu_l, \Sigma_l} \right] =$$

$$= \text{argmax}_l \left[\frac{\pi_l}{(2\pi)^{\frac{p}{2}} |\Sigma_l|^{\frac{1}{2}}} \exp \left\{ -\frac{1}{2} (X - \mu_l)^T \Sigma_l^{-1} (X - \mu_l) \right\} \right] =$$

po opuszczeniu wyrażeń niezależnych od l i zlogarytmowaniu:

$$= \text{argmax}_l \left[\log(\pi_l) - \frac{1}{2} \log(|\Sigma_l|) - \frac{1}{2} (X - \mu_l)^T \Sigma_l^{-1} (X - \mu_l) \right],$$

czyli kwadratowa funkcja argumentu X.

Liniowa funkcja klasyfikacyjna (lda) wymaga założenia:

$$\Sigma_1 = \ldots = \Sigma_K = \Sigma.$$

Dzięki niemu mamy podwójny zysk obliczeniowy: o $K-1$ parametrów mniej do wyestymowania i liniową funkcję optymalizowaną:

$$d(X) = \text{argmax}_l \left[\log(\pi_l) - \frac{1}{2} \log(|\Sigma_l|) - \frac{1}{2} (X - \mu_l)^T \Sigma_l^{-1} (X - \mu_l) \right] =$$

ponieważ $\log(|\Sigma|)$ oraz $X^T \Sigma^{-1} X$ nie zależy od l,

$$= \text{argmax}_l \left[\log(\pi_l) + X^T \Sigma^{-1} \mu_l - \frac{1}{2} \mu_l^T \Sigma^{-1} \mu_l \right].$$

7.4 Metody porównywania klasyfikatorów

Chcemy znaleźć taką metodę porónywania, żeby każdą obserwację spośród $(y_1, X_1), \ldots, (y_n, X_n)$ wykorzystać do uczenia i testu, ale tak żeby testować tylko na tych obserwacjach, które nie były brane pod uwagę przy uczeniu klasyfikatorów.

Kroswalidacja m-krotna (walidacja krzyżowa) polega na podziale danych na m części (popularnymi wyborami są $m = 5$, $m = 10$): $m - 1$ będzie tworzyć próbę uczącą, ostatnia będzie próbą testową. Estymujemy klasyfikatory na próbie uczącej, porównujemy metody na próbie testowej. Powtarzamy procedurę m razy tak, żeby każda z części była próbą testową. Dokładniej:

1. Permutujemy obserwacje. Jeżeli dane mają jakąś strukturę, na przykład można je podzielić na klasy, permutujemy obserwacje w klasach.
2. Dzielimy próbę na m części tak, żeby w każdej z grup było po tyle samo obserwacji z każdej klasy.
3. Uczymy klasyfikatory na próbie uczącej - estymujemy parametry.
4. Porównujemy metody na próbie testowej (np. poprzez estymację prawdopodobieństwa poprawnej predykcji)

Definicja 7.6. **Prawdopodobieństwo poprawnej predykcji** to dla danego klasyfikatora $\mathbb{P}(d(X) = y)$. Np. jeżeli funkcja straty wyraża się wzorem $L(k, l) = \mathbf{1}_{k \neq l}$, możemy estymować prawdopodobieństwo poprawnej predykcji dla konkretnej próby treningowej i testowej następująco:

$$\widehat{ppp}_i = \frac{\sum_{i \in \text{próba testowa}} \mathbf{1}(d(X_i) = y_i)}{\sum_{i \in \text{próba testowa}} 1},$$

gdzie d jest klasyfikatorem wyestymowanym na podstawie próby uczącej. Uśrednione $\widehat{ppp}$ jest dobrą metodą porównywania klasyfikatorów:

$$\widehat{ppp} = \frac{\sum_{i=1}^{m} \widehat{ppp}_i}{m}.$$

7.5 Przykłady w programie R

Klasyfikacja:

- kwadratowa funkcja klasyfikacyjna dla danych Iris, kroswalidacja: `http://www.mimuw.edu.pl/~pokar/StatystykaII/PREDYKCJA/qda.R`

- liniowa funkcja klasyfikacyjna: http://www.mimuw.edu.pl/~pokar/StatystykaII/PREDYKCJA/lda.R
- kwadratowa funkcja klasyfikacyjna oraz sieci neuronowe: http://www.mimuw.edu.pl/~pokar/StatystykaII/PREDYKCJA/CrossValKlasCrabs.R
- kwadratowa funkcja klasyfikacyjna oraz sieci neuronowe, kroswalidacja: http://www.mimuw.edu.pl/~pokar/StatystykaII/PREDYKCJA/crossValKlas.R
- porównanie różnych funkcji klasyfikacyjnych: http://www.mimuw.edu.pl/~pokar/StatystykaII/PREDYKCJA/zmDyskrym.R

Rozdział 8

Modele liniowe

8.1 Predykcja cechy ciągłej

Będziemy obserwować ciągłą zmienną objaśnianą y_i oraz zmienne objaśniające $x_i = (x_{i1}, \ldots, x_{ip})$, $i = 1, \ldots, n$. Na ich podstawie będziemy chcieli znaleźć funkcję zależącą od x, która będzie najlepiej przybliżać cechę y. Ograniczymy się przy tym tylko do zależności liniowej. Na podstawie znalezionej funkcji, dla nowo zaobserwoawanych x_{n+1} będziemy mogli znaleźć predykcję $\widehat{y}_{n+1}$. Jeżeli rozpatrzymy jednowymiarowy x_i ($p = 1$), szukanie funkcji liniowej najlepiej przybliżającej dane obrazuje rysunek 8.1.

Rysunek 8.1: Regresja liniowa jako dopasowanie prostej do danych.

Dane są postaci:

$$\underbrace{\begin{pmatrix} y_1 \\ \dots \\ y_n \end{pmatrix}}_{\text{losowe,obserwowane}} = \underbrace{\begin{pmatrix} x_{11} & \dots & x_{1p} \\ \dots & \dots & \dots \\ x_{n1} & \dots & x_{np} \end{pmatrix}}_{\text{deterministyczne,obserwowane}} \cdot \underbrace{\begin{pmatrix} \beta_1 \\ \dots \\ \beta_p \end{pmatrix}}_{\text{szukane}} + \underbrace{\begin{pmatrix} \varepsilon_1 \\ \dots \\ \varepsilon_n \end{pmatrix}}_{\text{losowe,nieobserwowane}} .$$

W zapisie macierzowym:

$$y = X\beta + \varepsilon,$$

gdzie:

- wektor y będziemy nazywać zmienną objaśnianą;
- macierz X macierzą planu;
- β estymowanymi parametrami;
- ε to wektor efektów losowych (wektor realizacji zmiennej losowej).

Dla tak sformułowanego problemu przyjmiemy następujące założenia:

1. $\mathbb{E}(\varepsilon) = 0$;
2. $\mathrm{Var}(\varepsilon) = \sigma^2 I_n$;
3. rząd macierzy X jest pełny: $\mathrm{rank}(X) = p$.

Dla tak sformułowanych danych, problem szukania estymatora parametru β będziemy nazywać **problemem liniowym**.

Twierdzenie 8.1. ***Rozkład QR macierzy***

Szeroki rozkład QR: Każdą rzeczywistą macierz A wymiaru $n \times m$ $(m \leqslant n)$ można zapisać jako iloczyn maierzy ortogonalnej Q wymiaru $n \times n$ $(Q^T Q = I_n)$ oraz górnotrójkątnej macierzy R wymiaru $n \times m$:

$$A = QR = \begin{pmatrix} & Q & \end{pmatrix} \begin{pmatrix} \circ & \circ \\ & \circ \end{pmatrix}.$$

Wąski rozkład QR: Ponieważ $(n-m)$ dolnych wierzy macierzy R jest zerowa, można skrócić zapis do

$$A = QR = Q \begin{pmatrix} R \\ 0 \end{pmatrix} = \begin{pmatrix} Q_1 & Q_2 \end{pmatrix} \cdot \begin{pmatrix} R_1 \\ 0 \end{pmatrix} = Q_1 R_1 =,$$

$$= \begin{pmatrix} & Q_1 & \end{pmatrix} \begin{pmatrix} \circ & \circ \\ & \circ \end{pmatrix}.$$

gdzie Q_1 jest macierzą wymiaru $n \times m$ o ortogonalnych kolumnach a R_1 jest macierzą górnotrójkątną wymiaru $m \times m$.

Dowód. Wąski rozkład QR jest zapisem macierzowym ortogonalizacji Gramma-Schmidta układu wektorów będących kolumnami macierzy A. Szeroki rozkład otrzymujemy dopełniając macierz Q_1 do bazy przestrzeni $\mathbb{R}^n$. □

Problem 8.1. Przy założeniu modelu liniowego, będziemy chcieli wyestymować nieznane parametry: β i σ^2.

8.2 Metoda najmniejszych kwadratów (mnk)

Zauważmy, że

$$\mathbb{E}(y) = \mathbb{E}(X\beta + \varepsilon) = X\beta + \mathbb{E}(\varepsilon) = X\beta.$$

Estymator najmnieszych kwadratów parametru β to taka jego wartość, dla której odległości euklidesowe przybliżanych danych od prostej je przybliżających jest najmniejsza:

$$\widehat{\beta} = \min_{\beta} ||y - X\beta||^2.$$

Twierdzenie 8.2. *Estymator najmniejszych kwadratów wyraża się wzorem*

$$\widehat{\beta} = R_1^{-1} Q_1^T y,$$

gdzie R_1 i Q_1 pochodzą z wąskiego rozkładu QR macierzy planu X.

Dowód. Skorzystajmy z szerokiego rozkładu QR macierzy X: $X = QR$. Ponieważ mnożenie wektora przez macierz ortogonalną nie zmienia jego normy, możemy zapisać:

$$||y - X\beta||^2 = ||Q^T(y - QR\beta)||^2 =$$
$$= ||Q_1^T(y - Q_1R_1\beta)||^2 + ||Q_2^T(y - Q_1R_1\beta)||^2 =$$
$$= ||Q_1^T y - R_1\beta||^2 + ||Q_2^T y||^2.$$

Wyrażenie to osiąga minimum ze względu na parametr β, jeżeli wyzerujemy pierwszy składnik sumy:

$$Q_1^T y = R_1\widehat{\beta}.$$

Ponieważ macierz R_1 jest kwadratowa i pełnego rzędu ($\text{rank}(X) = p$), możemy ją odwrócić:

$$\widehat{\beta} = R_1^{-1} Q_1^T y.$$

□

Wniosek 8.1. *Zauważmy, że:*

1. *Predykcja dla y jest równa $\widehat{y} = X\widehat{\beta}$;*

2.

$$||y - \widehat{y}||^2 = ||Q_2^T y||^2. \tag{8.1}$$

Przyjrzyjmy się własnościom metody najmniejszych kwadratów (zostaną one udowodnione w dalszej części wykładu):

1. $\widehat{y} = X\widehat{\beta}$ jest rzutem ortogonalnym y na przestrzeń rozpiętą przez kolumny macierzy planu X.

2. Nieobciążonym estymatorem parametru σ^2 jest $\widehat{\sigma}^2 = \frac{||y-\widehat{y}||^2}{n-p}$.

3. **Twierdzenie Gaussa-Markowa**: estymator $\widehat{\beta}$ jest liniowym, nieobciążonym estymatorem o najmniejszej wariancji parametru β (BLUE-Best Linear Unbiased Estimator).

4. Przy założeniu $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$, zachodzi **twierdzenie Fishera**:

   - $\widehat{\beta} \sim \mathcal{N}(\beta, \sigma^2(X^T X)^{-1})$;
   - $\frac{(n-p)\widehat{\sigma}^2}{\sigma^2} \sim \chi^2(n-p)$;
   - $\widehat{\beta}$ i $\widehat{\sigma}^2$ są niezależne.

8.3 Inne wyprowadzenie estymatora najmniejszych kwadratów

Wyprowadzimy estymator mnk jako rozwiązanie zadania BLUE – liniowy, nieobciążony estymator o najmniejszej wariancji. Rozumowanie będzie jednocześnie dowodem twierdzenia Gaussa-Markowa.

Twierdzenie 8.3. *Dla problemu liniowego estymator postaci $\widehat{\beta} = (X^T X)^{-1} X^T y$ jest liniowym, nieobciążonym estymatorem o najmniejszej wariancji parametru β.*

Dowód. Żeby łatwiej mówić o nieobciążoności, czy minimalnej wariancji, zredukujemy wymiar problemu do jednowymiarowego zakładając, że własności będą zachodzić dla wszystkich możliwych kombinacji liniowych zmiennej objaśnianej: Dla danego wektora l, konstruujemy kombinację liniową $\mu = l^T\beta$ i szukamy dla niej estymatora zależącego liniowo od y:

$$\widehat{\mu} = c^T y$$

przy założeniu nieobciążoności:

$$\mathbb{E}(\widehat{\mu}) = \mu = l^T\beta.$$

Jednocześnie wiemy, że:

$$\mathbb{E}(\widehat{\mu}) = c^T \mathbb{E}(y) = c^T X \beta.$$

Stąd:

$$c^T X = l^T.$$

Będziemy minimalizować wariancję estymatora $\widehat{\mu}$:

$$\mathrm{Var}(\widehat{\mu}) = \mathrm{Var}(c^T y) = c^T \mathrm{Var}(y) c = \sigma^2 c^T c.$$

Zadanie optymalizacyjne wygląda następująco:

$$\min_c \left[\sigma^2 c^T c\right] \text{ pod warunkiem } c^T X = l^T.$$

Skorzystajmy z metody mnożników Lagrange'a:

$$\min_c \left[F(c, \lambda) = \sigma^2 c^T c + (c^T X - l^T) \underbrace{\lambda}_{\text{wektor}} \right].$$

Szukamy estymatora wektora c, spełniającego dwa równania:

$$\frac{\partial F}{\partial c} = 2\sigma^2 c + X\lambda = 0; \tag{8.2}$$

$$\frac{\partial F}{\partial \lambda} = X^T c - l = 0. \tag{8.3}$$

Z równania 8.2 otrzymujemy: $c = -\frac{1}{2\sigma^2} X\lambda$, wstawiamy do równania 8.3:

$$-\frac{1}{2\sigma^2} X^T X = l,$$

skąd:

$$\lambda = -2\sigma^2 (X^T X)^{-1} l.$$

Macierz X jest pełnego rzędu, więc macierz $(X^T X)$ jest odwracalna. Wstawiając λ do wzoru na c, otrzymujemy:

$$c = X(X^T X)^{-1} l \text{ , inaczej: } c^T = l^T (X^T X)^{-1} X^T.$$

Estymator $\widehat{\mu}$ jest więc postaci:

$$\widehat{\mu} = c^T y = l^T (X^T X)^{-1} X^T y;$$

podstawiając za l kolejne wektory bazy kanonicznej $e_i = (0, \ldots, 0, 1, 0, \ldots, 0)$, znajdujemy kolejne estymatory kombinacji liniowych $l^T \beta = e_i^T \beta = \widehat{\beta}_i$, co łącznie możemy zapisać jako:

$$\widehat{\beta} = (X^T X)^{-1} X^T y.$$

□

Stwierdzenie 8.1. *Liniowy, nieobciążony estymator o najmniejszej wariancji parametru β w modelu liniowym jest równy estymatorowi najmniejszych kwadratów.*

Dowód.

$$\hat{\beta} = (X^T X)^{-1} X^T y =$$

korzystając z wąskiego rozkładu QR: $X = Q_1 R_1$,

$$= (R_1^T \underbrace{Q_1^T Q_1}_{=I_p} R_1)^{-1} R_1^T Q_1^T y =$$

$$= R_1^{-1} \underbrace{(R_1^T)^{-1} R_1^T}_{=I_p} Q_1^T y = R_1^{-1} Q_1^T y.$$

□

8.4 Estymatory metody największej wiarygodności prametrów modelu liniowego

Estymatory największej wiarygodności to takie wartości parametrów, których prawdopodobieństwo zaobserwowania danych jest największe. Żeby skorzystać z tej metody estymacji, potrzebna jest funkcja wiarygodności, niezbędne więc będzie założenie na temat rozkładu danych:

Zamiast zakładać:

$$\mathbb{E}\varepsilon = 0 \quad , \quad \mathrm{Var}(\varepsilon) = \sigma^2 I_n$$

założymy:

$$\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n),$$

skąd mamy $y \sim \mathcal{N}(X\beta, \sigma^2 I_n)$ oraz funkcję wiarygodności:

$$f_{\beta,\sigma^2}(y_1, \ldots, y_n) = f_{\beta,\sigma^2}(y) = \frac{1}{(2\pi)^{\frac{n}{2}} |\sigma^2 I_n|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(y - X\beta)\sigma^{-2} I_n (y - X\beta)\right] =$$

$$= \frac{1}{(2\pi)^{\frac{n}{2}} (\sigma^2)^{\frac{n}{2}}} \exp\left[-\frac{1}{2\sigma^2}||y - X\beta||^2\right].$$

Funkcję wiarygodności będziemy chcieli zmaksymalizować po parametrach β i σ^2. Ponieważ logarytm jest funkcją rosnącą, jest to równoważne z maksymalizacją logarytmu funkcji wiarygodności:

$$F_{\beta,\sigma^2}(y) = \log f_{\beta,\sigma^2}(y) = C - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}||y - X\beta||^2,$$

gdzie C jest stałą niezależną od szukanych parametrów. Zadanie maksymalizacji logwiarygodności F_{β,σ^2} jest rónoważne minimalizacji $-2F_{\beta,\sigma^2}$:

$$-2F_{\beta,\sigma^2}(y) = C' + n\log(\sigma^2) + \frac{1}{\sigma^2}||y - X\beta||^2.$$

Część sumy zależąca od parametru β to $\frac{1}{\sigma^2}||y - X\beta||^2$. Wartością parametru β minimalizującą to wyrażenie jest:

$$\widehat{\beta}_{mnw} = (X^TX)^{-1}X^Ty = R_1^{-1}Q_1^Ty,$$

co udowodniliśmy już w twierdzeniu 8.2.

Ponieważ $\widehat{\beta}$ nie zależy od parametru σ^2, mogę wstawić estymator do funkcji wiarygodności przy szukaniu optymalnego parametru σ^2. Oznaczmy także $\tau = \sigma^2$ żeby nie mylił się nam kwadrat przy parametrze:

$$-2F_{\beta,\sigma^2}(y) = C' + n\log(\tau) + \frac{1}{\tau}||y - X\widehat{\beta}||^2.$$

$$\frac{\partial(-2F_{\beta,\sigma^2})}{\partial\tau} = \frac{n}{\tau} - \frac{1}{\tau^2}||y - X\widehat{\beta}||^2 = 0,$$

skąd otrymujemy:

$$n\tau = ||y - X\widehat{\beta}||^2;$$

$$\widehat{\tau} = \widehat{\sigma}^2 = \frac{||y - X\widehat{\beta}||^2}{n}.$$

Wniosek 8.2. *Przy założeniu rozkładu normalnego:*

1. *estymatory parametru β dla metody największej wiarygodności i metody najmniejszych kwadratów są równe:*

$$\widehat{\beta}_{mnk} = \widehat{\beta}_{mnw} = (X^TX)^{-1}X^Ty;$$

2. *estymatory parametru σ^2 dla metody największej wiarygodności i metody najmniejszych kwadratów są równe z dokładnością do stałej:*

$$\widehat{\sigma}^2_{mnk} = \frac{||y - X\widehat{\beta}||^2}{n-p}, \quad \widehat{\sigma}^2_{mnw} = \frac{||y - X\widehat{\beta}||^2}{n}.$$

8.5 Kolejne własności estymatorów mnk

8.5.1 Wartość oczekiwana i wariancja estymatora $\widehat{\beta}$

- Wartość oczekiwana:

$$\mathbb{E}\widehat{\beta} = \mathbb{E}(X^TX)^{-1}X^Ty = (X^TX)^{-1}X^T\mathbb{E}y = (X^TX)^{-1}X^TX\beta = \beta. \tag{8.4}$$

  Estymator jest niebciążony.

- Macierz wariancji:

$$\mathrm{Var}(\widehat{\beta}) = \mathbb{E}(\widehat{\beta}-\beta)(\widehat{\beta}-\beta)^T =$$

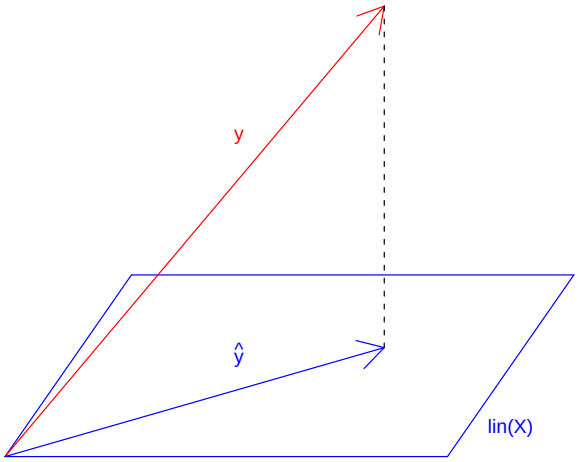
$$= \mathbb{E}((X^TX)^{-1}X^Ty - (X^TX)^{-1}X^T\mathbb{E}y)((X^TX)^{-1}X^Ty - (X^TX)^{-1}X^T\mathbb{E}y)^T =$$

$$= (X^TX)^{-1}X^T\mathbb{E}\left[(y-\mathbb{E}y)(y-\mathbb{E}y)^T\right]((X^TX)^{-1}X^T)^T =$$

$$= (X^TX)^{-1}X^T\sigma^2 I_n((X^TX)^{-1}X^T)^T = \sigma^2(X^TX)^{-1}\underbrace{(X^TX)(X^TX)^{-1}}_{=I_p} =$$

$$= \sigma^2(X^TX)^{-1}. \tag{8.5}$$

8.5.2 Dopasowanie $\widehat{y}$ jako rzut ortogoanlny y na przestrzeń rozpiętą przez kolumny macierzy X



Rysunek 8.2: Dopasowanie $\widehat{y}$ jako rzut ortogonalny y na $\mathrm{lin}(X)$.

Prypomnijmy

$$\widehat{\beta} = (X^TX)^{-1}X^Ty \quad , \quad \widehat{y} = X\widehat{\beta}.$$

Definicja 8.1. **Macierzą daszkową** H nazwiemy taką macierz, że:

$$\widehat{y} = X(X^TX)^{-1}X^Ty = Hy.$$

Stąd:

$$H = X(X^TX)^{-1}X^T.$$

Stwierdzenie 8.2. *Zauważmy, że $\widehat{y}$ jest nieobciążonym estymatorem y:*

$$\mathbb{E}\widehat{y} = X(X^TX)^{-1}X^TX\beta = X\beta.$$

Stwierdzenie 8.3. *Własności macierzy daszkowej H:*

1. *$HX = X$:*
$$X\underbrace{(X^TX)^{-1}X^TX}_{=I_p} = X;$$
2. *macierz H jest idempotentna, czyli $HH = H$:*
$$HH = X\underbrace{(X^TX)^{-1}X^TX}_{=I_p}(X^TX)^{-1}X^T = H;$$
3. *symetryczna, czyli $H^T = H$:*
$$H^T = (X(X^TX)^{-1}X^T)^T = X(X^TX)^{-1}X^T = H.$$
4. *korzystając z wąskiego rozkładu QR macierzy X, $H = Q_1Q_1^T$:*
$$H = Q_1R_1(R_1^T\underbrace{Q_1^TQ_1}_{=I_p}R_1)^{-1}R_1^T\underbrace{Q_1^TQ_1}_{=I_p}\underbrace{R_1R_1^{-1}}_{=I_p}\underbrace{(R_1^T)^{-1}R_1^T}_{=I_p}Q_1^T =$$
$$= Q_1\underbrace{R_1R_1^{-1}}_{=I_p}\underbrace{(R_1^T)^{-1}R_1^T}_{=I_p}Q_1^T = Q_1Q_1^T.$$
5. *korzystając z szerokiego rozkładu QR macierzy X, możemy przyjrzeć się rozkładowi spektralnemu macierzy daszkowej:*
$$H = QR(R^T\underbrace{Q^TQ}_{=I_n}R)^{-1}R^TQ^T =$$
ponieważ $R^TR = R_1^TR_1$,
$$= QR(R_1^TR_1)^{-1}R^TQ^T = QRR_1^{-1}(R_1^T)^{-1}R^TQ^T =$$
$$= \underbrace{Q\begin{pmatrix}1 & & & \\ & \cdots & & \\ & & 1 & \\ & & & 0\end{pmatrix}Q^T}_{\text{rozkład spektralny macierzy } H} =$$
dla $Q = [q_1, \ldots, q_p, \ldots, q_n]$:
$$= Q_1Q_1^T = \sum_{i=1}^{p} 1 \cdot q_iq_i^T + \sum_{i=p+1}^{n} 0 \cdot q_iq_i^T$$

Wniosek 8.3. *1. Macierz daszkowa H jest macierzą rzutu ortogonalnego na przestrzeń rozpiętą przez kolumny macierzy X.*

2. Jeżeli $\widehat{y} = X\widehat{\beta}$ minimalizuje wyrażenie $||y - \widehat{y}||^2$, to jest rzutem ortogonalnym y na $lin(X)$.

8.5.3 Nieobciążony estymator parametru σ^2

Stwierdzenie 8.4. *Macierz $(I_n - H)$ jest macierzą rzutu ortogonalnego na przestrzeń prostopadłą do przestrzeni rozpiętej przez kolumny macierzy X, jest więc w szczególności symetryczna i idempotentna.*

Dowód.

$$I_n - H = I_n - Q \begin{pmatrix} I_p & \\ & 0 \end{pmatrix} Q^T =$$

$$= QQ^T - Q \begin{pmatrix} I_p & \\ & 0 \end{pmatrix} Q^T = Q \begin{pmatrix} 0 & \\ & I_{n-p} \end{pmatrix} Q^T.$$

□

Wniosek 8.4. *Ponieważ ślad macierzy równy jest sumie jego wartości własnych, ślady macierzy daszkowej H i macierzy $(I - H)$ to:*

$$tr(H) = p;$$

$$tr(I - H) = n - p.$$

Stwierdzenie 8.5. *Twierdzenie Pitgorasa w postaci macierzowej:*

$$Var(y) = Var(\widehat{y}) + Var(y - \widehat{y}).$$

Dowód.

$$\sigma^2 I_n = \mathrm{Var}(Hy) + \mathrm{Var}((I - H)y);$$

$$\sigma^2 I_n = H\mathrm{Var}(y)H^T + (I - H)\mathrm{Var}(y)(I - H)^T;$$

ponieważ macierze H i $I - H$ są symetryczne i idempotentne, zachodzi:

$$\sigma^2 I_n = \sigma^2 H + \sigma^2 (I_n - H). \tag{8.6}$$

□

Stwierdzenie 8.6. *Niebciążonym estymatorem parametru σ^2 w modelu liniowym jest:*

$$\widehat{\sigma}^2 = \frac{||y - \widehat{y}||^2}{n - p}.$$

Dowód. Ponieważ $\widehat{y}$ jest nieobciążonym estymatorem y, możemy zapisać:

$$\mathbb{E}||y-\widehat{y}||^2 = \sum_{i=1}^{n}\left(\mathrm{Var}(y_i-\widehat{y}_i)\right) = \mathrm{tr}\left(\mathrm{Var}(y-\widehat{y})\right) =$$

$$= \mathrm{tr}\Big(\sigma^2(I-H)\Big) = \sigma^2(n-p).$$

Stąd:

$$\frac{\mathbb{E}||y-\widehat{y}||^2}{n-p} = \sigma^2.$$

□

8.5.4 Model z większą liczbą parametrów nie musi być lepiej dopasowany dla nowych danych

- Błąd predykcji y za pomocą $\widehat{y}$ na tej samej próbie, korzystając ze wzoru 8.6, można zapisać w postaci:

  $$\mathbb{E}(y_i-\widehat{y}_i)^2 = \mathrm{Var}(y_i-\widehat{y}_i) = \sigma^2(1-h_{ii}),$$

  gdzie h_{ii} jest elementem macierzy daszkowej: $H=(h_{ij})_{i,j=1}^{n}$.

  Definicja 8.2. Elementy przekątnej macierzy daszkowej H: h_{ii} będziemy nazywać **ładunkami obserwsacji i-tej** i oznaczać h_i.

- Dla nowych obserwacji mamy:

  Zakładamy niezależność nowych obserwacji zmiennej objaśnianej i p-wymiarowego wektora zmiennych objaśniających: (y_{n+1}, x_{n+1}^T) od (y, X). Będziemy estymować parametry używając danych treningowych (y, X), a obliczać błąd dla nowych danych testowych:

  $$y_{n+1} = x_{n+1}^T\beta + \varepsilon_{n+1};$$

  $$\widehat{\beta} = (X^TX)^{-1}X^Ty;$$

  $$\widehat{y}_{n+1} = x_{n+1}^T\widehat{\beta}.$$

  Błąd predykcji jest równy:

  $$\mathbb{E}(y_{n+1}-\widehat{y}_{n+1})^2 = \mathrm{Var}(y_{n+1}-\widehat{y}_{n+1}) \underbrace{=}_{\text{z niezależności}} \mathrm{Var}(y_{n+1})+\mathrm{Var}(\widehat{y}_{n+1}) =$$

  $$= \sigma^2 + \mathrm{Var}(x_{n+1}^T\widehat{\beta}) = \sigma^2 + x_{n+1}^T\mathrm{Var}(\widehat{\beta})x_{n+1} =$$
  $$= \sigma^2 + x_{n+1}^T\sigma^2(X^TX)^{-1}x_{n+1} = \sigma^2(1+h_{n+1}),$$

  gdzie $h_{n+1} = x_{n+1}^T(X^TX)^{-1}x_{n+1}$, analogicznie do ładunków obserwacji dla $i=1,\ldots,n$: $h_i = x_i^T(X^TX)^{-1}x_i$.

- Porównanie obu błędów predykcji dla tej samej macierzy planu:

  Dane treningowe, dla których będziemy estymować parametr $\widehat{\beta}$ to (y, X) gdzie $y = (y_1, \ldots, y_n)$. Dane testowe, dla których będziemy liczyć błąd predykcji to w pierwszym przypadku ten sam zbiór (y, X), a w drugim (y^{te}, X) gdzie $y = (y_{n+1}, \ldots, y_{2n})$ są nowymi obserwacjami, a macierz planu X pozostaje niezmieniona. Porównajmy uśrednione oba błędy predykcji:

$$\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2\right) = \frac{1}{n}\mathbb{E}||y - \widehat{y}||^2 = \frac{(n-p)\sigma^2}{n};$$

$$\mathbb{E}\left(\frac{1}{n}\sum_{i=n+1}^{2n}(y_i - \widehat{y}_i)^2\right) = \frac{1}{n}\sum_{i=n+1}^{2n}\sigma^2(1+h_i) = \frac{(n+p)\sigma^2}{n},$$

  gdzie korzystamy z równości $h_{n+i} = h_i$, co zachodzi dzięki użyciu tej samej macierzy planu X w zbiorze treningowym i testowym oraz własnści macierzy daszkowej $\sum_{i=1}^{n} h_i = p$.

  Wniosek 8.5. *Na podstawie obliczonych błędów predykcji możemy wywnioskować:*

  1. *Większy model nie zawsze oznacza lepsze dopasowanie.*
  2. *Różnica pomiędzy błędami predykcji wynosi:*

$$\frac{1}{n}\mathbb{E}||y - \widehat{y}||^2 - \frac{1}{n}\mathbb{E}||y^{tr} - \widehat{y}||^2 = \frac{2p\sigma^2}{n}$$

8.5.5 Kroswalidacja leave-one-out

Estymację błędu predykcji można oprzeć na kroswalidacji leave-one-out. Dla każdej obserwacji będziemy estymować model za pomocą wszystkich obserwacji oprócz niej samej i obliczać błąd predykcji na nowych danych dla tej pominiętej obserwacji. W ten sposób dostaniemy n błędów predykcji, które następnie uśrednimy.

Niech $X_{(i)}$, $i = 1, \ldots, n$ oznacza macierz X z usuniętą i-tą obserwacją (i-tym wierszem), $y_{(i)}$ wektor obserwacji z usuniętą i-tą obserwacją. Estymator $\widehat{\beta}_{(i)}$ będzie oznaczać estymator mnk na podstawie danych $(y_{(i)}, X_{(i)})$:

$$\widehat{\beta}_{(i)} = (X_{(i)}^T X_{(i)})^{-1} X_{(i)}^T y_{(i)};$$

Predykcja dla pominiętej obserwacji wyraża się wzorem:

$$\widehat{y}_{(i)} = x_i^T \widehat{\beta}_{(i)},$$

gdzie tak jak przy liczeniu błędu predykcji na nowych danych, (y_i, x_i^T) jest niezależne od $(y_{(i)}, X_{(i)})$.

Korzystając z tego, że $\mathbb{E}(\widehat{y}_{(i)}) = \mathbb{E}(y_i) = x_i^T\beta$, otrzymujemy:

$$\mathbb{E}(\widehat{y}_{(i)} - y_i)^2 = \mathrm{Var}(\widehat{y}_{(i)} - y_i) = \sigma^2(1 + x_i^T(X_{(i)}^T X_{(i)})^{-1}x_i) = \frac{\sigma^2}{1 - h_{ii}},$$

gdzie h_{ii} to i-ty wyraz na przekątnej macierzy daszkowej dla pełnej macierzy X: $H = X(X^TX)^{-1}X^T$. Fakt ostatniej równości w powyższym wzorze przyjmiemy bez dowodu.

Wniosek 8.6. *Estymator błędu predykcji przy użyciu kroswalidacji leave-one-out można uprościć do wzoru:*

$$\frac{||y - \widehat{y}_{()}||^2}{n} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_{(i)})^2 = \sum_{i=1}^{n}\frac{\sigma^2}{1 - h_{ii}}. \tag{8.7}$$

8.6 Model liniowy przy założeniu normalności

Zamiast w modelu liniowym zakładać:

$$\mathbb{E}\varepsilon = 0 \quad , \quad \mathrm{Var}(\varepsilon) = \sigma^2 I_n,$$

założymy:

$$\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n).$$

Dzięki takiemu sformułowaniu zadania, będziemy mogli znaleźć rozkłady estymatorów $\widehat{\beta}$ i $\widehat{\sigma}^2$, co umożliwi wnioskowanie statystyczne na ich temat, na przykład kondtrukcję przedziałów ufności. Udowodnimy:

Twierdzenie 8.4 (**Fishera**). *Przy założeniu $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$, estymatory modelu liniowego spełniają:*

1. $\widehat{\beta} \sim \mathcal{N}(\beta, \sigma^2(X^TX)^{-1})$;
2. $\widehat{\beta}$ *i* $\widehat{\sigma}^2 = \frac{||y - X\widehat{\beta}||^2}{n-p}$ *są niezależne;*
3. $\frac{(n-p)\widehat{\sigma}^2}{\sigma^2} \sim \chi^2(n-p)$;

Dowód. Ponieważ $y = X\beta + \varepsilon$, mamy:

$$y \sim \mathcal{N}(X\beta, \sigma^2 I_n). \tag{8.8}$$

Wiemy, że nieobciążonymi estymatorami parametrów modelu liniowego są:

$$\widehat{\beta} = (X^TX)^{-1}X^Ty \quad , \quad \widehat{\sigma}^2 = \frac{||y - X\widehat{\beta}||^2}{n-p}.$$

1. Korzystając z twierdzenia 7.1 oraz wzoru 8.8, możemy wywnioskować, że $\widehat{\beta}$ ma rozkład normalny. Wartość oczekiwana estymatora (8.4) i wariancja (8.5) wyznaczają jednoznacznie rozkład.

2. Przypomnijmy wzory 8.2 i 8.1:

$$\widehat{\beta} = R_1^{-1} Q_1^T y \quad , \quad ||y - X\widehat{\beta}||^2 = ||Q_2 y||^2.$$

Wiemy z twierdzenia 7.1, że Qy ma rozkład normalny. Policzmy macierz wariancji tego wektora losowego:

$$\mathrm{Var}(Q^T y) = Q^T \mathrm{Var}(y) Q = \sigma^2 I_n. \tag{8.9}$$

Wiemy zatem, że wektory losowe $Q_1^T y$ i $Q_2^T y$ są nieskorelowane. Dla rozkładu normalnego brak korelacji jest równoważny niezależności.

3. Rozkład χ^2 z $n-p$ stopniami swobody to suma $n-p$ kwadratów niezależnych zmiennych losowych o rozkładzie standardowym normalnym. Udowodnimy, że $\frac{\widehat{\sigma}^2(n-p)}{\sigma^2} = \frac{||y-X\widehat{\beta}||^2}{\sigma^2} = \frac{||Q_2^T y||^2}{\sigma^2}$ ma rozkład $\chi^2(n-p)$.

Z rozkładu QR macierzy X znamy wymiary macierzy Q_2, długość wektora $Q_2^T y$ to $n-p$. Oznaczmy:

$$||Q_2^T y||^2 = (q_{p+1}^T y)^2 + \ldots + (q_n^T y)^2 = z_1^2 + \ldots + z_{n-p}^2.$$

Udowodnimy, że z_i, $i = 1, \ldots, n-p$ są niezależne i mają rozkład $\mathcal{N}(0, \sigma^2)$.

Współrzędne wektora $Q_2^T y$ są niezależnymi zmiennymi losowymi o rozkładzie normalnym. Normalność wynika z twierdzenia 8.2, niezależność z braku korelacji (8.9). Ze wzoru 8.9 widzimy także, że wariancje z_i są równe σ^2.

Współrzędne wektora $Q_2^T y$ mają wartość oczekiwaną równą zero:

$$\mathbb{E}(Q_2^T y) = Q_2^T \mathbb{E}(y) = Q_2^T X\beta =$$

z wąskiego rozkładu QR macierzy X,

$$= \underbrace{Q_2^T Q_1}_{=0} R_1 = 0$$

z ortogonalności kolumn macierzy Q.

Otrzymujemy więc:

$$\left(\frac{z_1}{\sigma}\right)^2 + \ldots + \left(\frac{z_{n-p}}{\sigma}\right)^2 = \frac{||Q_2^T y||^2}{\sigma^2} \sim \chi^2(n-p),$$

gdzie $\frac{z_i}{\sigma}$ są niezależnymi zmiennymi losowymi o rozkładzie $\mathcal{N}(0,1)$.

□

8.7 Test ilorazu wiarygodności (Likelihood Ratio Test) hipotez liniowych

Hipotezy liniowe przy założeniach modelu liniowego można ogólnie sformułować jako:

$$\begin{cases} y = X\beta + \varepsilon \\ A\beta = 0 \end{cases}$$

gdzie macierz X jest wymiaru $n \times p$, a macierz A wymiaru $(p-q) \times p$.

Przykład 8.1. Jeżeli wektor współczynników jest postaci:

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix}$$

i chcemy nałożyć ograniczenie liniowe na parametry: $\beta_2 = \beta_3$, to można go zapisać postaci:

$$\underbrace{\begin{pmatrix} 0 & 1 & -1 & 0 \end{pmatrix}}_{A} \beta = 0.$$

8.7.1 LRT ogólnie

Ogólnie test ilorazu wiarygodności dotyczący parametru θ rozkładu zmiennej losowej X można zapisać jako:

$$H_0 : \ X \sim f_\theta(x) \ , \theta \in \Theta_0;$$

$$H_1 : \ X \sim f_\theta(x) \ , \theta \in \Theta_1 = \Theta \setminus \Theta_0;$$

gdzie $f_\theta(x)$ oznacza gęstość rozkładu zmiennej X zależącą od parametru θ.
Statystyka testowa wyraża się wzorem:

$$\lambda(x) = \frac{\sup_{\theta \in \Theta_1} f_\theta(x)}{\sup_{\theta \in \Theta_0} f_\theta(x)} = \frac{f_{\widehat{\theta}(x)}(x)}{f_{\dot{\theta}(x)}(x)},$$

gdzie:

$\widehat{\theta}$ to estymator największej wiarygodności dla $\theta \in \Theta_1$;

$\dot{\theta}$ to estymator największej wiarygodności dla $\theta \in \Theta_0$.

Uwaga 8.1. Jeżeli $f_\theta(x) = \prod_{i=1}^n f_\theta(x_i)$ to:

$$\log \lambda(x) = \sum_{i=1}^n \log f_{\widehat{\theta}(x)}(x) - \sum_{i=1}^n \log f_{\dot{\theta}(x)}(x).$$

8.7.2 Modele zagnieżdżone

Z modelem zagnieżdżonym mamy do czynienia gdy $\Theta_0 \subseteq \Theta$.
Rozpatrzmy następujący problem:

$$\Theta \subseteq \mathbb{R}^p, \quad h : \mathbb{R}^p \to \mathbb{R}^{p-q}, \quad \Theta_0 = \{\theta : h(\theta) = 0\}.$$

Dla hipotez liniowych mamy:

$$h(\theta) = A\theta = 0,$$

wtedy typowo $\Theta_0 \subseteq \mathbb{R}^q$, skąd możemy zapisać:

$$\sup_{\theta \in \Theta_1} f_\theta(x) = \sup_{\theta \in \Theta} f_\theta(x).$$

Dzięki takiemu zapisowi upraszcza się wzór na statystykę testową LRT:

$$\lambda(x) = \frac{\sup_{\theta \in \Theta} f_\theta(x)}{\sup_{\theta \in \Theta_0} f_\theta(x)}.$$

Twierdzenie 8.5 (**Asymptotyczny rozkład LRT**)**.** *Przy założeniach: $\Theta \subseteq \mathbb{R}^p$ otwarty, $f_\theta(x)$ regularna rodzina gęstości, $h : \mathbb{R}^p \to \mathbb{R}^{p-q}$ funkcja gładka, $\Theta_0 = \{\theta : h(\theta) = 0\}$:*

$$\forall \theta \in \Theta_0 \quad \forall t \quad \mathbb{P}_\theta \left(2 \log \lambda(x) \leqslant t\right) \xrightarrow[n \to \infty]{} F_{\chi^2(p-q)}(t),$$

gdzie $F_{\chi^2(p-q)}$ oznacza dystrybuantę rozkładu χ^2 o $p-q$ stopniach swobody.

8.7.3 LRT w modelu liniowym

Wracamy teraz do modelu linowego i zakładamy normalność rozkładu ε:

$$\begin{cases} y = X\beta + \varepsilon \\ A\beta = 0 \end{cases}$$

gdzie X ma wymiary $n \times p$, A wymiary $(p-q) \times p$;

$$\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n).$$

Dla tak sformułowanego zadania wiemy, że rozkład danych y jest normalny i wyraża się wzorem:

$$f_\theta(y) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left[-\frac{||y - X\beta||^2}{2\sigma^2}\right],$$

gdzie $\theta = (\beta_1, \ldots, \beta_p, \sigma^2)$.

Stwierdzenie 8.7. *Statystyka testowa testu ilorazu wiarygodności dla* $H_0 : A\beta = 0$ *jest równa:*

$$\lambda(y) = \left[\frac{\dot{\sigma}^2}{\widehat{\sigma}^2}\right]^{\frac{n}{2}} = \left[\frac{||y - X\dot{\beta}||^2}{||y - X\widehat{\beta}||^2}\right]^{\frac{n}{2}}, \tag{8.10}$$

gdzie:

$\widehat{\theta} = (\widehat{\beta}, \widehat{\sigma}^2)$ *estymatory największej wiarygodności bez ograniczeń (na zbiorze* Θ*);*

$\dot{\theta} = (\dot{\beta}, \dot{\sigma}^2)$ *estymatory największej wiarygodności z ograniczeniami* $(A\beta = 0$*, na zbiorze* $\Theta_0)$*.*

Dowód.

$$\lambda(y) = \frac{f_{\widehat{\theta}(y)}(y)}{f_{\dot{\theta}(y)}(y)} = \frac{\frac{1}{(\sqrt{2\pi})^n}\frac{1}{(\widehat{\sigma}^2)^{\frac{n}{2}}}\exp\left[-\frac{1}{2}\frac{||y-X\widehat{\beta}||^2}{\widehat{\sigma}^2}\right]}{\frac{1}{(\sqrt{2\pi})^n}\frac{1}{(\dot{\sigma}^2)^{\frac{n}{2}}}\exp\left[-\frac{1}{2}\frac{||y-X\dot{\beta}||^2}{\dot{\sigma}^2}\right]} = \clubsuit$$

korzystając z postać estymatora największej wiarygodności dla parametru σ^2 w modelu liniowym, możemy zapisać:

$$\widehat{\sigma}^2 = \frac{||y - X\widehat{\beta}||^2}{n} \quad , \quad \dot{\sigma}^2 = \frac{||y - X\dot{\beta}||^2}{n};$$

pdstawiając otrzymujemy:

$$\clubsuit = \frac{\frac{1}{(\widehat{\sigma}^2)^{\frac{n}{2}}}\exp\left[-\frac{1}{2}\frac{\widehat{\sigma}^2}{n\widehat{\sigma}^2}\right]}{\frac{1}{(\dot{\sigma}^2)^{\frac{n}{2}}}\exp\left[-\frac{1}{2}\frac{\dot{\sigma}^2}{n\dot{\sigma}^2}\right]} = \left[\frac{\dot{\sigma}^2}{\widehat{\sigma}^2}\right]^{\frac{n}{2}}.$$

□

Stwierdzenie 8.8. *Statystyka testowa*

$$\lambda(y) = \left[\frac{\dot{\sigma}^2}{\widehat{\sigma}^2}\right]^{\frac{n}{2}}$$

jest równoważna statystyce:

$$F = \frac{(R_0 - R)/(p - q)}{R/(n - p)},$$

gdzie $R_0 = ||y - X\dot{\beta}||^2$*,* $R = ||y - X\widehat{\beta}||^2$*.*

Dowód. Ze wzoru 8.10 widzimy, że:

$$\lambda(y) = \left(\frac{R_0}{R}\right)^{\frac{n}{2}}.$$

Statystyka λ jako iloraz norm dwóch wektorów, jest nieujemna, a dla $p < n$ dodatnia z dodatniości R i R_0. Istnieje rosnące przekształcenie λ w F dla $\lambda > 0$, więc statystyki są rónoważne. □

Twierdzenie 8.6. *Statystyka F przy $p \ll n$ ma rozkład $\mathcal{F}$-Snedecora:*

$$F = \frac{(R_0 - R)/(p-q)}{R/(n-p)} \sim \mathcal{F}(p-q, n-p).$$

Dowód. Zmieńmy oznaczenia dotyczące macierzy planu. Macierz $X = (x_1, \ldots, x_p)$ gdzie x_i będą oznaczać kolumny macierzy, zwane **predyktorami**. Możemy wtedy zapisac:

$$y = X\beta + \varepsilon = x_1\beta_1 + \ldots + x_p\beta_p + \varepsilon.$$

Wiemy, że:

$$X\widehat{\beta} \in \mathcal{L} = \{X\beta = x_1\beta_1 + \ldots + x_p\beta_p,\ \beta \in \mathbb{R}^p\};$$
$$X\dot{\beta} \in \mathcal{L}_0 = \{X\beta : A\beta = 0\};$$

gdzie

$$\mathcal{L}_0 \subseteq \mathcal{L} \subseteq \mathbb{R}^n,$$

przestrzenie $\mathcal{L}_0$ i $\mathcal{L}_0$ są przestrzeniami liniowymi o wymiarach:

$$\underbrace{\dim(\mathcal{L}_0)}_{=q} < \underbrace{\dim(\mathcal{L})}_{=p} \ll n.$$

Ortogonalizujemy bazę przestrzeni $\mathcal{L}_0$, uzupełniamy do bazy $\mathcal{L}$, a następnie do bazy $\mathbb{R}^n$. Oznaczmy:

$$V = \underbrace{(\underbrace{v_1, \ldots, v_q}_{\text{baza } \mathcal{L}_0}, v_{q+1}, \ldots, v_p, v_{p+1}, \ldots, v_n)}_{\text{baza } \mathcal{L}}.$$

oraz:

$$Z = V^T y, \quad \widehat{Z} = V^T X\widehat{\beta}, \quad \dot{Z} = V^T X\dot{\beta}.$$

Zauważmy, że wektory te są postaci:

$$Z = \begin{pmatrix} z_1 \\ \ldots \\ z_q \\ \ldots \\ z_p \\ \ldots \\ \ldots \\ z_n \end{pmatrix} \quad \widehat{Z} = \begin{pmatrix} \widehat{z}_1 \\ \ldots \\ \widehat{z}_q \\ \ldots \\ \widehat{z}_p \\ 0 \\ \ldots \\ 0 \end{pmatrix} \quad \dot{Z} = \begin{pmatrix} \dot{z}_1 \\ \ldots \\ \dot{z}_q \\ 0 \\ \ldots \\ \ldots \\ \ldots \\ 0 \end{pmatrix}.$$

Możemy wtedy zapisać:

$$R = ||y - X\widehat{\beta}||^2 =$$

ponieważ mnożenie wektora przez macierz ortogonalną nie zmienia jego normy,

$$= ||V^T(y - X\widehat{\beta})||^2 = ||Z - \widehat{Z}||^2;$$

$$R = ||y - X\dot{\beta}||^2 = ||V^T(y - X\dot{\beta})||^2 = ||Z - \dot{Z}||^2.$$

Najlepszymi dopasowaniami $\dot{Z}$ do Z oraz $\widehat{Z}$ do Z minimalizującymi błędy średniokwadratowe R_0 i R są:

$$\dot{Z} = \begin{pmatrix} \dot{z}_1 = z_1 \\ \dots \\ \dot{z}_q = z_q \\ 0 \\ \dots \\ \dots \\ 0 \end{pmatrix} \qquad \widehat{Z} = \begin{pmatrix} \widehat{z}_1 = z_1 \\ \dots \\ \dots \\ \widehat{z}_p = z_p \\ 0 \\ \dots \\ 0 \end{pmatrix}$$

Stąd:

$$R_0 = z_{q+1}^2 + \ldots + z_p^2 + \ldots + z_n^2;$$

$$R = z_{p+1}^2 + \ldots + z_n^2.$$

Ponieważ założyliśmy rozkład normalny dla ε, możemy zapisać:

$$y \sim \mathcal{N}(X\beta, \sigma^2 I_n),$$

a także, ponieważ V jest macierzą ortogonalną:

$$Z = V^T y \sim \mathcal{N}(V^T X\beta, \sigma^2 I_n). \tag{8.11}$$

Współrzędne wektora Z: $z_1, \ldots, z_n$ mają więc rozkłady normalne i są niezależne (bo nieskorelowane). Co więcej, przy założeniu hipotezy zerowej, $X\beta \in \mathcal{L}_0$, czyli jest postaci:

$$X\beta = \begin{pmatrix} w_1 \\ \dots \\ w_q \\ 0 \\ \dots \\ 0 \end{pmatrix}$$

w bazie V. Ze wzoru 8.11, $\mathbb{E}(Z) = V^T X\beta$, czyli:

$$\mathbb{E}(z_{q+1}) = \ldots = \mathbb{E}(z_n) = 0.$$

Widzimy teraz, że wyrażenie:

$$\frac{R_0 - R}{\sigma^2} = \left(\frac{z_{q+1}}{\sigma}\right)^2 + \ldots + \left(\frac{z_p}{\sigma}\right)^2$$

ma rozkład $\chi^($p - q), a wyrażenie:

$$\frac{R}{\sigma^2} = \left(\frac{z_{p+1}}{\sigma}\right)^2 + \ldots + \left(\frac{z_n}{\sigma}\right)^2$$

rozkład $\chi^($n - p) oraz oba wyrażenia są od siebie ziezależne. Wróćmy do postaci statystyki F:

$$F = \frac{(R_0 - R)/(p-q)}{R/(n-p)} = \frac{(z_{q+1}^2 + \ldots + z_p^2)/(p-q)}{(z_{p+1}^2 + \ldots + z_n^2)/(n-p)}$$

ma więc rozkład $\mathcal{F}(p-q, n-p)$.

□

Uwaga 8.2. Zauważmy ciekawą własność bazującą na dowodzie twierdzenia: dla $H_0 : \beta_1 = \ldots = \beta_p = 0$ przy modelu postaci:

$$y = \begin{pmatrix} 1 & & \\ 1 & \star & \\ \ldots & & \\ 1 & & \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \ldots \\ \beta_p \end{pmatrix} + \varepsilon,$$

możemy zapisać:

$$||y - \underbrace{\overline{y}1_n}_{=X\dot{\beta}}||^2 = ||y - \underbrace{\widehat{y}}_{=X\widehat{\beta}}||^2 + ||\widehat{y} - \overline{y}1_n||^2,$$

gdzie $\overline{y}$ jest średnią arytmetyczną z obserwacji w wektorze y.

Wniosek 8.7. *Testowanie hipotez o istotności współczynników (testowanie hipotez, czy kolejne grupy β_i są równe zeru) służy wyborowi modelu (podzbioru zmiennych objaśniających $x_1, \ldots, x_p$).*

8.8 Popularne kryteria wyboru modelu – kryteria informacyjne

W poprzednim rozdziale zostało opisane testowanie hipotez o istotności współczynników jako sposób wyboru modelu. Wybór predyktorów można także oprzeć na minimalizacji estymatora błędu predykcji wyliczonego na podstawie kroswalidacji leave-one-out (8.7). Opiszemy teraz jeszcze inną

metodę wyboru zmiennych objaśniających bazującą na tak zwanych **kryteriach informacyjnych** postaci:

$$\text{kryterium} \;=\; -2\cdot\text{logwiarygodność} \;+\; \text{kara za złożoność modelu},$$

które obliczane są dla każdego modelu (dla każdego podzbioru p predyktorów) i wybierany jest ten minimalnej wartości kryterium. Dwa popularne kryteria informacyjne:

1. **Akaike Information Criterion** (AIC):

$$AIC = \;-2\cdot\text{loglik}\; + 2\cdot p,$$

   gdzie p to liczba zmiennych objaśniających w modelu.

2. **Bayes Information Criterion** (BIC):

$$BIC = \;-2\cdot\text{loglik}\; + \log(n)\cdot p,$$

   gdzie n to liczba obserwacji w modelu.

Przy założeniach modelu liniowego i normalności rozkładu $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$, kryteria przyjmują łatwiejszą postać:

1. Przy znanym σ^2:

$$AIC = \frac{||y-\widehat{y}||^2}{\sigma^2} + 2\cdot p;$$

$$BIC = \frac{||y-\widehat{y}||^2}{\sigma^2} + \log(n)\cdot p;$$

2. Przy nieznanym σ^2:

$$AIC = n\log\left(\frac{||y-\widehat{y}||^2}{n}\right) + 2\cdot p;$$

$$BIC = n\log\left(\frac{||y-\widehat{y}||^2}{n}\right) + \log(n)\cdot p;$$

8.9 Model logistyczny – przykład uogólnionego modelu liniowego

Modelu logistycznego używa się do objaśniania zmiennej binarnej, czyli przyjmującej wartości ze zbioru $\{0,1\}$. Poprzednio zakładaliśmy:

$$y_i \sim \mathcal{N}(x_i^T\beta, \sigma^2), \quad y_i \text{ niezależne } i = 1,\ldots,n,$$

gdzie wektor x_i oznacza wiersz macerzy planu.
Teraz będziemy zakładać rozkład:

$$y_i \sim \underbrace{\mathcal{B}(1, p(x_i))}_{\text{rozkład Bernoulliego}} = \mathcal{B}\left(1, \frac{\exp(x_i^T\beta)}{1+\exp(x_i^T\beta)}\right), \quad y_i \text{ niezależne } i = 1, \ldots, n,$$

gdzie postać funkcji $p(x_i)$ można tłumaczyć tym, że prawdopodobieństwo powinno przyjmować wartości z przedziału $[0, 1]$. Parametry modelu (β) estymuje się metodą największej wiarygodności, gdzie funkcja wiarygodności jest równa:

$$L(\beta) = \prod_{i=1}^{n} (p(x_i))^{y_i} (1 - p(x_i))^{1-y_i} = \prod_{i=1}^{n} \frac{\exp(y_i x_i^T\beta)}{1+\exp(x_i^T\beta)}.$$

Logarytm funkcji wiarygodności maksymalizuje się numerycznie aby otrzymać estymatory $\widehat{\beta}$. Predykcję w modelu można oprzeć na klasyfikatorze:

$$t = x_{n+1}^T\widehat{\beta},$$

gdzie x_{n+1} jest wektorem nowych obserwacji. Przewidywany na podstawie modelu $\widehat{y}_{n+1}$ to wtedy:

$$\widehat{y}_{n+1} = \begin{cases} 1, & t \geqslant 0 \text{ (wtedy } p(x_{n+1}) \geqslant \frac{1}{2}); \\ 0, & t < 0 \text{ (wtedy } p(x_{n+1}) < \frac{1}{2}). \end{cases}$$

8.10 Przykłady w programie R

Model liniowy:

- regresja liniowa z diagnostyką dla danych Bodyfat: http://www.mimuw.edu.pl/~pokar/StatystykaII/DANE/bodyfat.R
- regresja liniowa z diagnostyką dla danych Samochody: http://www.mimuw.edu.pl/~pokar/StatystykaII/PREDYKCJA/samochodyNowe.R
- regresja liniowa dla danych Iris: http://www.mimuw.edu.pl/~pokar/StatystykaII/PREDYKCJA/lm.R
- regresja liniowa dla danych Samochody: http://www.mimuw.edu.pl/~pokar/StatystykaII/PREDYKCJA/samochody.R
- porównanie metody najmniejszych kwadratów i sieci neuronowych: http://www.mimuw.edu.pl/~pokar/StatystykaII/PREDYKCJA/crossValRegr.R

Logit (model logistyczny):

- estymacja parametrów: http://www.mimuw.edu.pl/~pokar/StatystykaII/PREDYKCJA/logit.R
- estymacja parametrów i rysowanie wyników: http://www.mimuw.edu.pl/~pokar/StatystykaII/PREDYKCJA/Orings.R

Rozdział 9

Własności estymatorów MNK

9.1 Warunkowa wartość oczekiwana jako rzut ortogonalny

Definicja 9.1. Przypomnienie

Warunkowa wartość oczekiwana:

Niech Y będzie całkowalną zmienną losową w przestrzeni probabilistycznej $(\Omega, \mathcal{F}, \mathbb{P})$, $\mathcal{M}$ σ-ciałem takim, że $\mathcal{M} \in \mathcal{F}$. Warunkową wartością oczekiwaną Y pod warunkiem $\mathcal{M}$ nazywamy zmienną losową $\mathbb{E}(Y|\mathcal{M})$, że:

1. $\mathbb{E}(Y|\mathcal{M})$ jest $\mathcal{M}$-mierzalna,
2. $\forall A \in \mathcal{M} \quad \int_A Y dP = \int_A \mathbb{E}(Y|\mathcal{M}) dP$.

Załóżmy, że dla przestrzeni probabilistycznej $(\Omega, \mathcal{F}, \mathbb{P})$ określone zostały zmienne losowe całkowalne z kwadratem: $X, Y \in L^2(\Omega, \mathcal{F}, \mathbb{P})$. Zdefiniujmy $\widehat{Y} = \mathbb{E}(Y|X)$ $(= \mathbb{E}(Y|\sigma(X)))$.

Stwierdzenie 9.1. *$\widehat{Y}$ jest rzutem ortogonalnym Y na $L^2(\Omega, \sigma(X), \mathbb{P}_X)$, gdzie $\sigma(X)$ jest σ-ciałem generowanym przez X, a $\mathbb{P}_X$ to miara prawdopodobieństwa warunkowego pod warunkiem zmiennej losowej X.*

Dowód. Załóżmy, że Z jest rzutem ortogonalnym Y na $\sigma(X)$. Wtedy $\forall\ A \in \sigma(X)$:

$$\mathbb{E}\mathbf{1}(A)Y = \mathbb{E}\mathbf{1}(A)Z + \mathbb{E}\ \underbrace{\mathbf{1}(A)}_{\in\sigma(X)}\ \underbrace{(Y - Z)}_{\perp\sigma(X)} = \mathbb{E}\mathbf{1}(A)Z + 0$$

Z definicji warunkowej wartości oczekiwanej $\widehat{Y} = Z$ p.n.

□

9.2 Twierdzenie Pitagorasa

Niech X, Y oznaczają zmienne losowe, $X, Y \in L^2(\Omega, \mathcal{F}, \mathbb{P})$, $X \in \mathcal{L}_0 = L^2(\Omega, \sigma(X), \mathbb{P}_X)$. Zdefiniujmy iloczyn skalarny jako $< X, Y >= \mathbb{E}_p(XY)$.

Twierdzenie 9.1. *Pitagorasa*

$$||Y - X||^2 = ||Y - \widehat{Y}||^2 + ||\widehat{Y} - X||^2.$$

Dowód.

$$||Y-X||^2 = ||Y-\widehat{Y}+\widehat{Y}-X||^2 = ||Y-\widehat{Y}||^2+||\widehat{Y}-X||^2+2 < \underbrace{Y-\widehat{Y}}_{\perp\sigma(X)}, \underbrace{\widehat{Y}-X}_{\in\sigma(X)} >=$$

$$= ||Y - \widehat{Y}||^2 + ||\widehat{Y} - X||^2 + 0.$$

□

Wniosek 9.1. *Rzut ortogonalny jest więc najlepszym przybliżeniem Y w klasie $f \in \mathcal{L}_0$ w sensie:*

$$||Y - \widehat{Y}||^2 = \min_{f \in \mathcal{L}_0} ||Y - f||^2$$

Stwierdzenie 9.2. *Jeśli $f \in \mathcal{L}_0$, to* $< Y, f >=< \widehat{Y}, f >$.

Dowód.

$$< Y, f >=< \underbrace{Y - \widehat{Y}}_{\perp\sigma(X)}, \underbrace{f}_{\in\sigma(X)} > + < \widehat{Y}, f >= 0 + < \widehat{Y}, f > .$$

□

Stwierdzenie 9.3. *Niech $\widehat{\widehat{Y}}$ oznacza rzut $\widehat{Y}$ na $\mathcal{L}_1 \subseteq \mathcal{L}_0$. $\widehat{\widehat{Y}}$ jest rzutem Y na przestrzeń $\mathcal{L}_1$. Rzut rzutu jest rzutem.*

Dowód.

$$\forall f \in \mathcal{L}_1 \quad < Y - \widehat{\widehat{Y}}, f >=< Y, f > - < \widehat{\widehat{Y}}, f >=$$

korzystając ze stwierdzenia 9.2,

$$=< \widehat{Y}, f > - < \widehat{\widehat{Y}}, f >=< \widehat{Y} - \widehat{\widehat{Y}}, f >= 0$$

z założenia. □

Stwierdzenie 9.4. *Oznaczmy $\overline{Y}$ jako rzut ortogonalny Y na lin$\{1\}$, $\widehat{Y}$ rzut Y na $\mathcal{L} = lin(X, 1)$. Wtedy:*

$$< Y - \overline{Y}, \widehat{Y} - \overline{Y} >= ||\widehat{Y} - \overline{Y}||^2.$$

Dowód. Ponieważ $\widehat{Y}-\overline{Y} \in \mathcal{L}_0$, równość wynika łatwo ze stwierdzenia 9.2. □

Definicja 9.2. Przypomnienie
Korelacja:

$$\mathrm{Cor}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma(X)\sigma(Y)} = \frac{\mathbb{E}(X-\mathbb{E}X)(Y-\mathbb{E}Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}.$$

Współczynnik dopasowania R^2 to część zmienności Y wyjaśnionej przez zmienność $\widehat{Y}$:

$$\mathrm{R}^2 = \frac{\mathrm{Var}(\widehat{Y})}{\mathrm{Var}(Y)}.$$

Błąd średniokwadratowy między X i Y:

$$\mathbb{E}(X-Y)^2.$$

Twierdzenie 9.2.

$$\mathrm{Cor}^2(Y,\widehat{Y}) = \mathrm{R}^2 = \frac{||\widehat{Y}-\overline{Y}||^2}{||Y-\overline{Y}||^2} = 1 - \frac{||\widehat{Y}-Y||^2}{||Y-\overline{Y}||^2}.$$

Dowód.

$$\mathrm{Cor}^2(Y,\widehat{Y}) = \frac{< Y-\overline{Y}, \widehat{Y}-\overline{Y} >^2}{< Y-\overline{Y}, Y-\overline{Y} >< \widehat{Y}-\overline{Y}, \widehat{Y}-\overline{Y} >} =$$

Korzystając z 9.4:

$$= \frac{||\widehat{Y}-\overline{Y}||^2}{||Y-\overline{Y}||^2} =$$

Korzystając z 9.2 dla $X = \overline{Y}$:

$$= 1 - \frac{||\widehat{Y}-Y||^2}{||Y-\overline{Y}||^2}.$$

□

Twierdzenie 9.3. *Pitagrorasa dla korelacji*

$$\forall f \in \sigma(X) \quad \mathrm{Cor}^2(Y,f) = \mathrm{Cor}^2(Y,\widehat{Y})\mathrm{Cor}^2(\widehat{Y},f).$$

Dowód. Załóżmy, że $Y = Y-\overline{Y}$, $\widehat{Y} = \widehat{Y}-\overline{Y}$, $f = f-\overline{f}$ (centrujemy zmienne losowe).

$$\mathrm{Cor}(Y,f) = \frac{< Y,f >}{< Y,Y >^{\frac{1}{2}} < f,f >^{\frac{1}{2}}} =$$

Korzystając z 9.2 i 9.4, mamy:

$$= \frac{< Y, \widehat{Y} >}{< Y, Y >^{\frac{1}{2}} < \widehat{Y}, \widehat{Y} >^{\frac{1}{2}}} \cdot \frac{< \widehat{Y}, f >}{< \widehat{Y}, \widehat{Y} >^{\frac{1}{2}} < f, f >^{\frac{1}{2}}} =$$

$$= \mathrm{Cor}^2(Y, \widehat{Y}) \mathrm{Cor}^2(\widehat{Y}, f).$$

□

Wniosek 9.2. *Największą korelację ze wszystkich $f \in \sigma(X)$, Y ma ze swoim rzutem ortogonalnym na przestrzeń rozpiętą przez X:*

$$\max_{f \in \sigma(X)} Cor^2(Y, f) = Cor^2(Y, \widehat{Y}).$$

Wniosek 9.3. *Patrząc na wnioski 9.1 i 9.2 oraz twierdzenie 9.2 zauważmy, że zachodzi zależność:*

Minimalizacja błędu średniokwadratowego $\Leftrightarrow$ Maksymalizacja kwadratu korelacji. Dla równoważnych problemów optymalnym jest rzut ortogonalny.

Obrazuje tą zależność także kolejne stwierdzenie:

Stwierdzenie 9.5. *Niech X, Y będą wystandaryzowanymi zmiennymi losowymi ($\mathbb{E}X = 0$, $\mathbb{E}Y = 0$, $\sigma(X) = 1$, $\sigma(Y) = 1$). Wtedy:*

$$\mathbb{E}(X - Y)^2 = 2 - 2\mathrm{Cor}(X, Y).$$

Dowód.

$$\mathbb{E}(X - Y)^2 = \underbrace{\mathbb{E}X^2}_{=1} + \underbrace{\mathbb{E}Y^2}_{=1} - 2\mathbb{E}XY = 2 - 2\mathrm{Cor}(X, Y).$$

□

Z twierdzenia 9.1 wynika jeszcze bardzo ważna zależność znana z rachunku prawdopodobieństwa:

Wniosek 9.4.

$$Var(Y) = Var(\mathbb{E}(Y|X)) + \mathbb{E}(Var(Y|X)),$$

gdzie $Var(Y|X) = \mathbb{E}((Y - \mathbb{Y})^2|X)$.

Dowód. W twierdzeniu 9.1 za X podstawmy $\mathbb{Y}$. □