

1 Wykład 11

Analiza wariancji (ANOVA)

Mamy próbę statystyczną postaci $X = (x_{ij})$ dla $i = 1, 2, \dots, k$, $j = 1, 2, \dots, n_i$. Model statystyczny jest postaci

$$x_{ij} = \mu_i + \varepsilon_{ij}$$

gdzie ε_{ij} jest próbą prostą z rozkładu normalnego $\mathcal{N}(0, \sigma)$

Dane są podzielone na k grup. W grupie o numerze i wektor $X_i^T = (x_{i1}, x_{i2}, \dots, x_{in_i})$ jest próbą prostą z rozkładu normalnego $\mathcal{N}(\mu_i, \sigma)$. Dane w poszczególnych grupach różnią się wartościami średnimi, natomiast mają ten sam rozrzut. Parametry $\mu_1, \mu_2, \dots, \mu_k$ oraz σ są nieznane. Stawiamy hipotezę

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_k = \mu$$

przeciwko hipotezie

$$H_1 : \text{nieprawda, że } (\mu_1 = \mu_2 = \dots = \mu_k)$$

Takie hipotezy stawia się w praktyce, gdy chcemy porównać różne sposoby działania, skuteczność różnych leków itp.

Odpowiedni test znajdziemy korzystając z ogólnego testu ilorazu wiarygodności:

$$\begin{aligned} L_X(H_0, H_1) &= \frac{\sup \left\{ \prod_{i=1}^k l(x_i, \mu_i, \sigma) : \mu_1, \mu_2, \dots, \mu_k, \sigma \right\}}{\sup \left\{ \prod_{i=1}^k l(x_i, \mu, \sigma) : \mu, \sigma \right\}} = \\ &= \left(\frac{s_0}{s_1} \right)^{N/2} \end{aligned}$$

gdzie

$$\begin{aligned} s_0 &= \sum_{ij} (x_{ij} - \bar{x}_{..})^2, \quad s_1 = \sum_{ij} (x_{ij} - \bar{x}_{i.})^2, \quad N = \sum_{i=1}^k n_i, \\ \bar{x}_{i.} &= \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}, \quad \bar{x}_{..} = \frac{1}{N} \sum_{ij} x_{ij} = \frac{1}{N} \sum_{i=1}^k n_i \bar{x}_{i.} \end{aligned}$$

Estymator $\bar{x}_{..}$ nazywany jest **średnią ogólną** i jest estymatorem μ gdy hipoteza zerowa jest prawdziwa. Estymator $\bar{x}_{i.}$ nazywany jest **średnią grupową** i jest estymatorem μ_i gdy prawdziwa jest hipoteza konkurencyjna.

Z postaci ilorazu wiarygodności wynika, że będziemy odrzucać hipotezę zerową, gdy iloraz $\frac{s_0}{s_1}$ będzie duży. Łatwo po prostych przekształceniach pokazać, że

$$\begin{aligned} s_0 &= \sum_{ij} (x_{ij} - \bar{x}_{..})^2 = \sum_{ij} (x_{ij} - \bar{x}_{i.} + \bar{x}_{i.} - \bar{x}_{..})^2 = \\ &= \sum_{ij} (x_{ij} - \bar{x}_{i.})^2 + 2 \sum_i (\bar{x}_{i.} - \bar{x}_{..}) \sum_j (x_{ij} - \bar{x}_{i.}) + \sum_i n_i (\bar{x}_{i.} - \bar{x}_{..})^2 = \\ &= \sum_{ij} (x_{ij} - \bar{x}_{i.})^2 + \sum_i n_i (\bar{x}_{i.} - \bar{x}_{..})^2 \stackrel{\text{def}}{=} s_1 + s_2 \end{aligned}$$

Stąd

$$\frac{s_0}{s_1} = \frac{s_1 + s_2}{s_1} = 1 + \frac{s_2}{s_1}$$

a więc iloraz wiarygodności jest rosnącą funkcją $\frac{s_2}{s_1}$. Licznik w tym ułamku mierzy, jak daleko są od siebie centra grup, mianownik - rozrzut wewnątrz każdej grupy. Dlatego s_1 nazywamy **sumą kwadratów wewnątrz grup**, s_2 **sumą kwadratów między grupami**. Można sobie wyobrazić grupy jako "grube" punkty. Chcąc stwierdzić, czy te "grube" punkty są od siebie odległe nie wystarczy stwierdzić, że ich środki są różne. Trzeba również stwierdzić, czy punkty nie "zachodzą na siebie". Te dwa warunki są tym bardziej spełnione im większą wartość ma $\frac{s_2}{s_1}$. Oczywiście iloraz wiarygodności jest również rosnącą funkcją

$$\frac{s_2/(k-1)}{s_1/(N-k)}$$

Nazwa testu analizy wariancji bierze się stąd, że zbiór krytyczny jest oparty o ten iloraz, w którym licznik jest estymatorem **wariancji międzygrupowej** a mianownik estymatorem **wariancji wewnątrzgrupowej**.

Tradycyjnie wyniki analizy wariancji zapisuje się w tablicy

zmiennosc	st. swobody	suma kwadratów	wariancja	F	wartość krytyczna
między grupami	$k - 1$	s_2	$v_2 = s_2 / (k - 1)$	$\frac{v_2}{v_1}$	
wewnątrz grup	$N - k$	s_1	$v_1 = s_1 / (N - k)$		
całkowita	$N - 1$	s_0			

Zmienna losowa $\frac{v_2}{v_1}$ ma rozkład $F_{k-1, N-k}$ Snedecora z $k - 1, N - k$ stopniami swobody.

Przykład 1.1 (kontynuacja przykładu z wodorostami). Po dołączeniu obserwacji z rejonu 3 macierz danych będzie miała postać:

R1	94	90	95	73	77	95	78	98	95	90	79	88	109	78
R2	89	75	65	82	73	87	76	72	83	58	89	78		
R3	60	53	55	62	74	71	53	75	68	83	67	71	67	51

Z tablicy tej wyliczamy wszystkie potrzebne dane

$$\bar{x}_{1.} = 88.5, \bar{x}_{2.} = 77.25, \bar{x}_{3.} = 65, \bar{x}_{..} = 76.9$$

Tablica analizy wariancji na poziomie istotności 0.05^1 :

zmiennosc	st. swobody	suma kwadratów	wariancja	F	wartość krytyczna
między grupami	2	3867.8	1933.9	20.06	< 3.316
wewnątrz grup	37	3567.8	96.4		
całkowita	39	7435.6			

¹w tablicach nie ma wszystkich układów stopni swobody - bierzemy wartości, leżące najbliżej

Z tablicy tej wynika, że możemy na poziomie istotności 0.05, odrzucić hipotezę o równości średnich w tych grupach². ■

Modele regresji liniowej

O regresji mówiliśmy gdy estymowaliśmy parametry metodą najmniejszych kwadratów. Przez regresję liniową rozumiemy

Definicja 1.1 Model regresji liniowej jest to funkcja losowa postaci

$$Y(x) = \alpha + \beta x + \varepsilon$$

gdzie α, β, x są liczbami, ε jest zmienną losową o wartości oczekiwanej równej 0.

Próba statystyczna w modelu regresji liniowej opisana jest przez:

- wektor argumentów : $X^T = (x_1, x_2, \dots, x_n)$
- wektor wartości: $Y^T = (y(x_1), y(x_2), \dots, y(x_n)) \stackrel{\text{def}}{=} (y_1, y_2, \dots, y_n)$

Zakładamy, że zachodzi model regresji liniowej: $Y(x_i) = \alpha + \beta x_i + \varepsilon_i$ gdzie $\varepsilon^T = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)$ jest próbą prostą z rozkładu o wartości oczekiwanej 0.

W naszych rozważaniach ograniczymy się do przypadku, że ε_i mają rozkłady $\mathcal{N}(0, \sigma)$. Parametry α, β, σ są nieznane.

Dla uproszczenia rachunków założymy, że $\bar{x} = 0$. Gdyby tak nie było to wystarczy zdefiniować nowe parametry $x_i^* = x_i - \bar{x}$. Jak łatwo zauważyć, $\bar{x}^* = 0$ oraz $\alpha + \beta x_i = \alpha^* + \beta^* x_i^* \Leftrightarrow \beta = \beta^*, \alpha = \alpha^* - \beta^* \bar{x}$ co oznacza, że gdy znamy parametry "z gwiazdką" to zawsze możemy obliczyć wyjściowe parametry.

Twierdzenie 1.1 Estymatory największej wiarygodności dla α, β są wyrażone wzorami

$$\hat{\alpha} = \bar{y}, \quad \hat{\beta} = \frac{\langle x, y \rangle}{\|x\|^2}$$

oraz

$$\begin{aligned} \langle x, y \rangle &= \sum_{i=1}^n x_i y_i \text{ (iloczyn skalarny),} \\ \|x\|^2 &= \sum_{i=1}^n x_i^2 \text{ (kwadrat długości wektora)} \end{aligned}$$

Estymatory te są estymatorami najmniejszych kwadratów.

Dowód. Funkcja wiarygodności wektora $Y^T = (y_1, y_2, \dots, y_n)$ będzie postaci³

$$\begin{aligned} l(y, \alpha, \beta, \sigma) &\stackrel{\text{def}}{=} \prod_{i=1}^n \frac{1}{\sigma} \phi\left(\frac{y_i - \alpha - \beta x_i}{\sigma}\right) = \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(\frac{-1}{2\sigma^2} \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2\right) \end{aligned}$$

²można ją odrzucić również na poziomie 0.01 gdyż wartość krytyczna jest mniejsza od 5.390. Poziom krytyczny, który można obliczyć za pomocą programu statystycznego, wynosi 0.000001 czyli jest niezwykle mały.

³i-ta obserwacja y_i ma rozkład $\mathcal{N}(\alpha + \beta x_i, \sigma)$

a więc jest malejącą funkcją $\sum_{i=1}^n (y_i - \alpha - \beta x_i)^2$ co oznacza, że estymatory największej wiarygodności są jednocześnie estymatorami najmniejszych kwadratów. Korzystając z wyników estymacji najmniejszych kwadratów (wykład 3, propozycja 3.1) otrzymamy tezę. ■

Warto zauważyć, że estymatory $\hat{\alpha}$ i $\hat{\beta}$ są liniowymi kombinacjami obserwacji y_i . Wyobraźmy sobie teraz, że chcemy znaleźć estymatory parametru β wychodząc z innych założeń:

- estymatorów są liniowe $\hat{\beta}(c, y) = \langle c, y \rangle = \sum_{i=1}^n c_i y_i$
- nieobciążone: $E\hat{\beta}(c, y) = \beta$
- minimalizujące średni błąd kwadratowy, w tym przypadku $\text{Var}(\hat{\beta}(c, y))$ ma być minimalna

Podobne warunki nałożymy na estymatory parametru α .

Twierdzenie 1.2 *Estymatory regresji są najlepszymi liniowymi, nieobciążonymi estymatorami parametrów (BLUE⁴)*

Dowód. Z nieobciążoności

$$\begin{aligned}\beta &= E\hat{\beta}(c, y) = E \sum_{i=1}^n c_i y_i = \sum_{i=1}^n c_i E y_i = \sum_{i=1}^n c_i (\alpha + \beta x_i) = \\ &= \alpha \sum_{i=1}^n c_i + \beta \sum_{i=1}^n c_i x_i\end{aligned}$$

a stąd

$$\sum_{i=1}^n c_i = 0, \quad \sum_{i=1}^n c_i x_i = 1$$

Z kolei

$$\begin{aligned}\text{Var}(\hat{\beta}(c, y)) &= \text{Var}\left(\sum_{i=1}^n c_i y_i\right) = \sum_{i=1}^n c_i^2 \text{Var}(y_i) = \\ &= \sigma^2 \sum_{i=1}^n c_i^2\end{aligned}$$

Należy więc znaleźć minimum wyrażenia

$$\sum_{i=1}^n c_i^2$$

przy ograniczeniach

$$\sum_{i=1}^n c_i = 0, \quad \sum_{i=1}^n c_i x_i = 1$$

Stosując metodę mnożników Lagrange'a otrzymamy znane nam estymatory najmniejszych kwadratów (i również największej wiarygodności) dla parametrów regresji. Podobne rozumowanie można zastosować do estymatorów parametru α . ■

⁴Best Linear Unbiased Estimator