LISTA ZADAŃ NR 9: Analiza korelacji i regresji

Zadanie 1

Zbadano zależność między dwiema cechami X i Y na podstawie 10-elementowej próby (np. X – wielkość pliku w MB, Y – czas przesyłania w sek.). Wyniki pomiarów:

- x_i: $3.5, 3.4, 2.1, 5.4, 1.1, 5.1, 6.9, 4.0, 4.5, 2.5$
- y_i: $1.6, 2.9, 1.5, 3.5, 0.6, 2.5, 7.1, 3.5, 2.1, 2.6$

Obliczyć współczynnik korelacji liniowej Pearsona r. Czy zależność jest silna?

Wskazówka: Należy stworzyć tabelkę pomocniczą z kolumnami $x_i^2, y_i^2, x_i y_i$ i obliczyć sumy.

Zadanie 2

W pewnym eksperymencie sieciowym zebrano dane dotyczące liczby błędów (X) i czasu reakcji systemu (Y). Ponieważ danych było dużo, obliczono gotowe sumy dla $n = 25$ pomiarów:

$$\sum x_i = 375, \quad \sum y_i = 175$$

$$\sum x_i^2 = 6125, \quad \sum y_i^2 = 1245, \quad \sum x_i y_i = 2615$$

Obliczyć współczynnik korelacji r oraz średnie $\bar{x}$ i $\bar{y}$.

Zadanie 3

Dla populacji, w której badane cechy (X, Y) mają dwuwymiarowy rozkład normalny (np. temperatura procesora a jego taktowanie), pobrano próbkę: $(3, 3), (5, 3), (6, 4), (5, 8), (7, 5), (8, 6), (8, 9), (5, 4), (6, 5)$... (pełne dane w zbiorze).

Wyznaczyć równanie prostej regresji liniowej $y = ax + b$ drugiego rodzaju (czyli regresji Y względem X).

Wskazówka: Wykorzystać metodę najmniejszych kwadratów.

Zadanie 4

Wykorzystując równanie prostej regresji wyznaczone w Zadaniu 3, oszacować przewidywaną wartość cechy Y (np. taktowanie), jeśli cecha X (temperatura) przyjmie wartość $x = 10$.

Jakie ryzyko niesie ze sobą prognozowanie dla x spoza zakresu danych z próby (ekstrapolacja)?

Zadanie 5

Dla danych z Zadania 2, gdzie współczynnik korelacji wyniósł $r \approx -0{,}1$ (do sprawdzenia w obliczeniach), zweryfikować na poziomie istotności $\alpha = 0{,}05$ hipotezę $H_0 : \rho = 0$ (brak korelacji w populacji) przeciwko hipotezie $H_1 : \rho \neq 0$.

Wskazówka: Zastosować statystykę t-Studenta:

$$t = \frac{r}{\sqrt{1 - r^2}} \sqrt{n - 2}$$

Zadanie 6

Dla wyznaczonej prostej regresji w Zadaniu 3 obliczyć wariancję resztkową s_r^2 (lub odchylenie standardowe reszt s_r). Interpretacja: Jak bardzo rzeczywiste punkty pomiarowe „rozrzucone" są wokół wyznaczonej prostej regresji?

Wzór: $s_r^2 = s_y^2(1 - r^2)$.

Zadanie 7

Dwóch ekspertów oceniało jakość interfejsu 10 aplikacji, przyznając im miejsca w rankingu od 1 do 10.

- Ekspert A: $1, 2, 3, 4, 5, 6, 7, 8, 9, 10$
- Ekspert B: $2, 1, 4, 3, 6, 5, 8, 7, 10, 9$

Obliczyć współczynnik korelacji rang Spearmana. Czy eksperci są zgodni w swoich ocenach?

Wskazówka: Stosujemy wzór na korelację rang, oparty na różnicach d_i między rangami.

Zadanie 8

Wyznaczono dwie proste regresji: y względem x oraz x względem y.

$$y = -0{,}6x + 2$$

$$x = -1{,}2y + 1$$

Obliczyć współczynnik korelacji r na podstawie współczynników kierunkowych tych prostych.

Wskazówka: Zachodzi związek $r^2 = a_{yx} \cdot a_{xy}$. Należy pamiętać o znaku współczynnika korelacji!

Zadanie 9

Przypuśćmy, że zależność między czasem wykonania algorytmu (Y) a wielkością danych (X) jest wykładnicza: $y = a \cdot e^{bx}$. W jaki sposób przekształcić te dane, aby można było zastosować znane wzory na regresję liniową i wyznaczyć parametry a i b?

Wskazówka: Zlogarytmować stronami równanie ($\ln y = \ln a + bx$). Wówczas nową zmienną zależną jest $Z = \ln Y$.

Zadanie 10

W systemie monitorowane są 3 parametry: X_1 (CPU), X_2 (RAM), X_3 (Disk IO). Obliczono korelacje parami: $r_{12} = 0.8$, $r_{13} = 0.1$, $r_{23} = 0.2$.

Które zmienne są ze sobą silnie skorelowane, a które są niemal niezależne? Co to oznacza dla administratora systemu (np. czy modernizacja RAMu wpłynie na zużycie CPU?)?