Analiza korelacji i regresji

KORELACJA – zależność liniowa

Obserwujemy parę cech ilościowych (X,Y). Doświadczenie jest tak pomyślane, aby obserwowane pary cech X i Y (tzn *i*–ta para x_i i y_i dla różnych *i*) mogły być zależne. Najczęściej są to cechy tego samego (*i*–tego) obiektu.

Współczynnik korelacji *r* mierzy **siłę** zależności **liniowej** między dwiema cechami (tzn. wzrost wartości jednej cechy o 1 powoduje wzrost drugiej o *a*, niezależnie od obecnej wartości cechy).

Własności współczynnika korelacji *r*:

1. jest liczbą niemianowaną,
2. $r \in <-1,1>$,
3. jeśli $r>\mathbf{0}$, to większym wartościom jednej cechy odpowiadają (średnio) większe

wartości drugiej cechy (zależność rosnąca, cechy zachowują się zgodnie).

4. jeśli $r<0$, to większym wartościom jednej cechy odpowiadają (średnio) mniejsze wartości drugiej cechy i odwrotnie (zależność malejąca, cechy zachowują się przeciwnie).
5. jeśli $r=0$, to bez względu na wartości przyjmowane przez jedną z cech, **średnie** wartości drugiej cechy są takie same – zmienne są **nieskorelowane**.
   def: cechy są **nieskorelowane** jeśli nie ma **zależności liniowej** między nimi. (niezależne –> nieskorelowane)
6. współczynnik korelacji jest **miernikiem liniowej zależności** między cechami X i Y. Im $|r|$ jest bliższe 1, tym bardziej „liniowa” jest zależność między cechami.
7. jeżeli (X,Y) ma dwuwymiarowy rozkład normalny, to $r=0$ jest równoważne niezależności cech X, Y.

Oceny siły związku (pamiętając o odpowiedniej liczebności próby):

$\lvert r \rvert$	siła związku korelacyjnego
0.0	brak
? - 0.4	słaba
0.4 - 0.7	średnia
0.7 - 0.9	silna
0.9 - 1.0	bardzo silna

Niech (X1,Y1), (X2,Y2), ... (Xn,Yn) będzie próbą, **współczynnik korelacji** z próby (próbkowy) ***r*** definiujemy:

$$r = \frac{\operatorname{cov}(X,Y)}{\sqrt{\operatorname{var} X \cdot \operatorname{var} Y}}$$

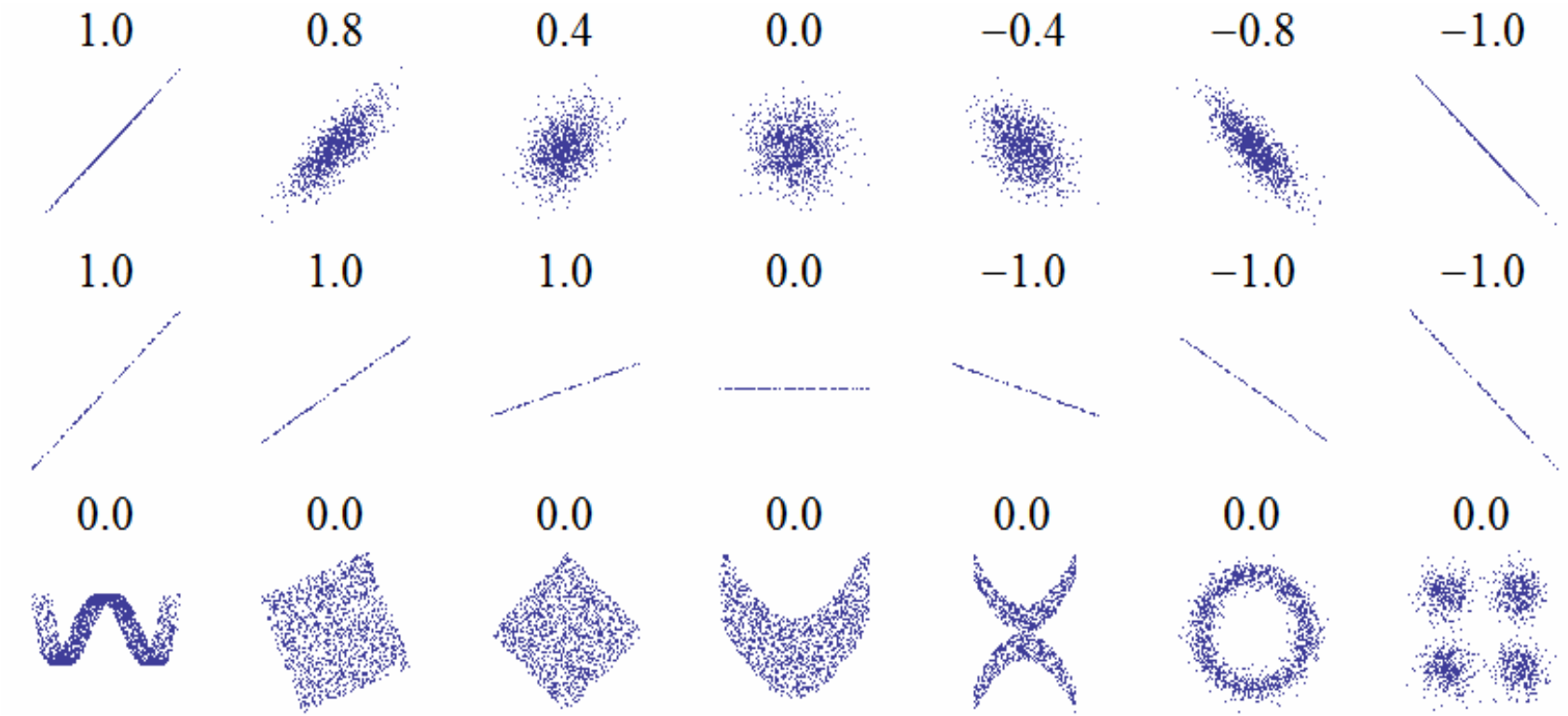
współczynnik kowariancji **cov(X,Y)**

$$\operatorname{cov}(X,Y) = \sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}$$

natomiast warians: var(X)=cov(XX)

Współczynnik r, obliczony z próby, jest estymatorem współczynnika populacyjnego r.

Przykładowe wykresy danych (x, y) i odpowiadające im wartości **współczynnika korelacji liniowej Pearsona**:



H_0: Cechy X i Y są niezależne H_0: $r = 0$

$$r_{emp} = \frac{\operatorname{cov}(X,Y)}{\sqrt{\operatorname{var} X \cdot \operatorname{var} Y}}$$

Jeśli $|r_{emp}| > r(\alpha, n-2)$, to hipotezę H_0 odrzucamy.

Testowanie istotności współczynnika korelacji liniowej Persona

$$H_0: \rho = 0$$

Do weryfikacji tej hipotezy służy statystyka:

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

gdzie: r jest próbkową wartością współczynnika korelacji Pearsona, n liczebnością próby. Jeśli $|t| \geq t_{\alpha,\ n-2}$ to H_0

odrzucamy; w przypadku relacji $|t| < t_{\alpha, n-2}$ nie mamy podstaw do odrzucenia H_0.

Często bada się korelacje między wieloma cechami (tzn. analizuje się współczynnik korelacji **r** wraz z jego realnym poziomem istotności $\mathbf{P_{value}}$ dla każdej pary cech). Ponieważ wykonuje się wiele porównań (podobnie jak w przypadku testu porównań wielokrotnych dla średnich po wykonaniu analizy wariancji ANOVA) to łączny poziom istotności jest o wiele wyższy niż dla pojedynczego porównania.

Przykład: badamy korelację między 20 cechami przyjmując poziom istotności 5%. Oznacza to 190 różnych par cech. Nawet jeśli są one nieskorelowane, to średnio 9,5 par uznamy za istotnie skorelowane (EX rozkładu Bernoulliego = np=190*5%). Prawdopodobieństwo, że wszystkie pary uznamy (słusznie) za nieskorelowane wynosi zaledwie 0,00006 (0,006%)! Na

poziomie istotności 1% średnio wykażemy skorelowanie 1,9 par a brak korelacji wszystkich par z P=15%.

Przykład: Badamy, czy istnieje zależność między wzrostem i wagą dorosłych osób. Losowo wybranym 20 osobom zmierzono wzrost i wagę (pierwsza osoba miała 180cm i 80kg, druga 170cm i 78kg itd). Policzono współczynnik korelacji prostej Pearsona **r**=0,85 i jego rzeczywisty poziom istotności $\mathbf{P_{value}}$=0,007. Na poziomie istotności 5% odrzucono hipotezę zerową o braku korelacji tych cech (braku zależności liniowej między wartościami cech) na rzecz hipotezy alternatywnej, że korelacja zachodzi. Współczynnik r=0,85 pokazuje, że zależność ta jest silna a cechy zachowują się zgodnie (osoby wyższe to zazwyczaj osoby o większej wadze). Analiza **nie określa**, o ile cięższa będzie średnio osoba wyższa o np 1cm czy o ile wyższa będzie średnio osoba cięższa o 1kg.

Badano współzależność liniową 3 cech soi: średniej masy ziarna w strąku, liczby nasion w strąku i liczby strąków. Wyniki przedstawiają pomiary **15** strąków, każdy z innej **rośliny**.

liczba strąków	średnia liczba nasion	średnia masa nasion
21	18	8
25	22	6
21	16	9
36	23	6
31	22	8
23	20	6
28	19	7
19	17	8
28	21	6
16	16	7
31	20	5
26	17	7
23	17	7
33	22	5
25	20	8

	l_nasion	l_strąków	m_nasiona
l_nasion		0,8133 0,0002	-0,5644 0,0284
l_strąków	**0,8133** 0,0002		-0,5518 0,0329
m_nasiona	**-0,5644** 0,0284	**-0,5518** 0,0329	

Liczba nasion i liczba strąków są **współzależne** liniowo i zachowują się **zgodnie** (duże wartości jednej cechy średnio odpowiadają dużym wartościom drugiej zmiennej). Obie te cechy są liniowo współzależne z masą nasiona.

Cecha masa nasiona zachowuje się przeciwnie do tamtych (duża masa wiązała się średnio z małą liczbą nasion i małą liczbą strąków, i odwrotnie, mała masa wiązała się średnio z dużą liczbą nasion i dużą liczbą strąków).

REGRESJA (prosta) Ilościowy opis zależności

Ilościowy opis zależności Y od X:

$E(Y|X=x)=f(x)$

Funkcja f(x) to funkcja regresji

Przy założeniu, że jest to dwuwymiarowy rozkład normalny $f(x) = \alpha + \beta x$

Regresja liniowa: zakładamy, że $r \neq 0$

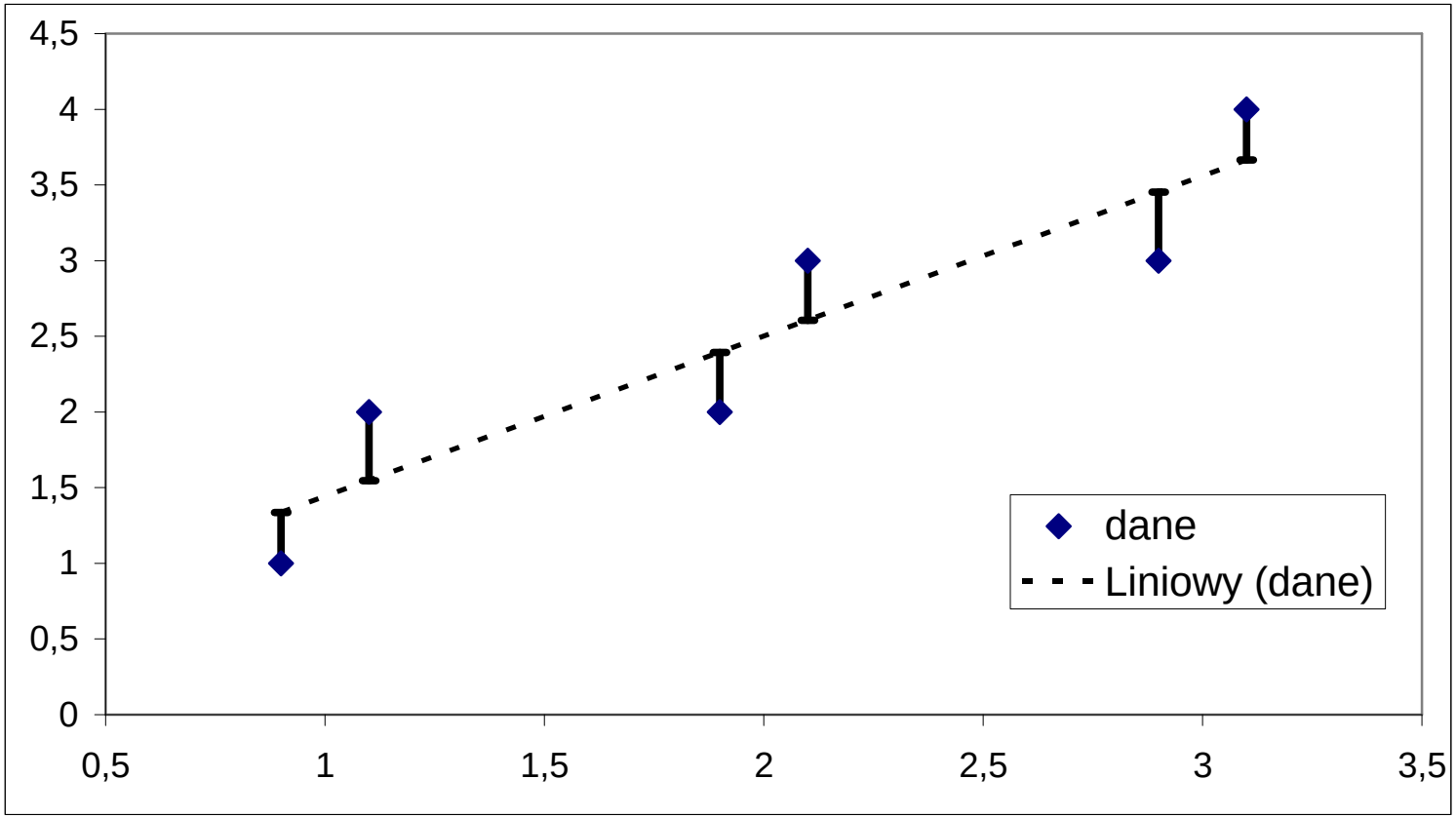
β – współczynnik regresji (*slope*)

α – stała regresji (*intercept*)

Współczynnik regresji β informuje o spodziewanej **zmianie** cechy **Y** jako reakcji na **wzrost** cechy **X** o **jedną** jednostkę.

Równanie prostej regresji należy tak wymodelować, aby było najlepiej dopasowane do danych empirycznych.

Współczynniki *a* i *b* są zwykle szacowane metodą najmniejszych kwadratów (MNK), która polega na takim ich doborze, aby suma kwadratów odchyleń rzędnych punktów empirycznych od wykresu prostej regresji była najmniejsza.



Wartości estymatorów regresji prostej oblicza się na podstawie próby ze wzorów:

$$\hat{b} = b = \frac{\operatorname{cov}(X,Y)}{\operatorname{var} X}, \qquad \hat{a} = a = \bar{Y} - \hat{b}\bar{X}$$

Współczynnik determinacji zmiennej Y przez X:

$$D = r^2 \cdot 100\%$$

Jest to liczba z przedziału (0%, 100%) i dopasowanie funkcji regresji jest tym lepsze im ten współczynnik jest wyższy.

Przykład: Badano zależność między zawartością witaminy C owocach czarnej porzeczki i jej zawartością w soku z nich produkowanego. Zbadano średnią zawartość witaminy w 15 partiach owoców oraz w soku wyprodukowanym z każdej z tych partii.

Przyjmujemy model liniowy dla tej zależności: zawartość witaminy C w soku zależy liniowo od zawartości witaminy w owocach. Y=a+bX

Hipoteza: Ten model nie nadaje się do wnioskowania o zawartości witaminy C w soku — wartościach cechy Y (**zmienność cechy Y wyjaśniana przez model jest taka sama jak zmienność losowa**, czyli model nie umożliwia lepszego przybliżenia wartości cechy Y niż jej średnia). Hipoteza alternatywna: model (a+bX) przybliża wartości cechy Y (**zmienność cechy Y wyjaśniana przez model jest większa niż zmienność losowa**). Rzeczywisty poziom istotności (P_{value}) dla hipotezy zerowej wynosi 0,0004. Odrzucamy hipotezę na rzecz hipotezy alternatywnej.

Model: $Y=aX+b$. Stawiamy hipotezę zerową, że współczynnik b jest równy 0 (czyli że $Y=aX+0$ –> $Y=a$ –> Y nie zależy od X) przy hipotezie alternatywnej, że a jest różne od 0. $P_{value}=0,0004$, więc odrzucamy hipotezę zerową na rzecz hipotezy alternatywnej (jest zależność liniowa).

Oceną wartości (estymatorem) dla współczynnika regresji między cechami w populacji jest wartość 0,43. Oznacza to, że jeżeli w jagodach zawartość witaminy C jest większa o 1, to w soku będzie średnio większa o 0,43.

Następnie stawiamy hipotezę zerową, że stała a jest równa 0 (czyli że $Y=0+bX$) przy hipotezie alternatywnej, że b jest różne od 0. $P_{value}=0,91$, więc nie mamy podstaw do odrzucenia hipotezy zerowej

(stała w regresji nie jest istotna, więc jej ocena, równa 0,48, nie jest ważna statystycznie).

Z tego wynika, że właściwym modelem mógłby być Y=*b*X. Dla takiego modelu zerowa zawartość witaminy w owocach (x=0) skutkowałaby zerową zawartością w soku (y=0). P_{value} dla hipotezy, że ten model **nie umożliwia lepszego przybliżenia wartości cechy Y niż jej średnia** wynosi 0,000, a ocena współczynnika regresji 0,44. Można powiedzieć, że zawartość witaminy C w jagodach wpływa na jej zawartość w soku wg zależności:

$C_{soku}=0,44*C_{jagód}$

Wyższa o 1 (10) zawartość witaminy C w jagodach powoduje wyższą o 0,44 (4,4) zawartość witaminy w soku.