Statystyka w pracy badawczej nauczyciela

Wykład 4: Analiza współzależności

dr inż. Walery Susłow
walery.suslow@ie.tu.koszalin.pl

Statystyczna teoria korelacji i regresji (1)

- Jest to dział statystyki zajmujący się badaniem zależności między cechami statystycznymi.
  - Dostarcza techniki do dokładnego określania stopnia, w jakim cechy są ze sobą powiązane.
  - Pozwala wykryć lub zweryfikować rozpoznane współzależności.
- Podstawowym problemem jest stwierdzenie:
  - Czy między cechami (zjawiskami, procesami, zdarzeniami) występuje jakiś związek (zależność).
  - Na ile związek ten jest ścisły.

Statystyczna teoria korelacji i regresji (2)

- *Analiza korelacyjna* daje możliwość stwierdzenia, czy istnieje związek miedzy badanymi cechami oraz jaka jest jego siła i kierunek.
  - Uwaga: niekoniecznie jest to związek przyczynowo-skutkowy!
  - Współczynnik korelacji jest to liczba określająca w jakim stopniu zmienne są współzależne.
- *Analiza regresyjna* daje możliwość oszacowania wartości jednej cechy na podstawie wartości przyjmowanych przez drugą cechę.
  - Funkcja regresji (równanie lub model regresji) opisuje *związek statystyczny* między zmiennymi.

Rodzaje zależności

- *Zależność funkcyjna* – zmiana wartości jednej cechy powoduje określoną zmianę wartości drugiej cechy (wiek i wzrost dziecka).
- *Zależność stochastyczna* – wraz ze zmianą jednej zmiennej zmienia się rozkład drugiej zmiennej (zmiana ciśnienia atmosferycznego i opady).
- *Zależność korelacyjna* – określonym wartościom jednej zmiennej odpowiadają określone średnie wartości drugiej zmiennej (sezonowe zmiany cen warzyw).

Analiza korelacyjna

Istota analizy korelacyjnej (1)

- Analiza korelacji ma sens głównie wtedy, gdy przypuszczamy, że między zmiennymi istnieje związek przyczynowo-skutkowy.
  - Wstępna *analiza jakościowa* umożliwia stwierdzenie takiego związku na podstawie merytorycznej analizy logicznej.
  - *Analiza ilościowa* określa siłę i kierunek związku.

- Uwaga: mimo ustalonego wpływu cechy x na cechę y, jednej i tej samej wartości x, może odpowiadać wiele różnych wartości y (chodzi o oddziaływanie innych niekontrolowanych czynników).

Istota analizy korelacyjnej (2)

- Rodzaje analizowanych cech: ilościowe i jakościowe
- Rodzaje związku: liniowy i nieliniowy
- Analizę korelacyjną zwykle rozpoczyna się od opracowania tablicy korelacyjnej oraz sporządzenia korelacyjnego wykresu rozrzutu.
- Mierniki współzależności dwóch cech ilościowych:
  - współczynnik korelacji liniowej Pearsona,
  - współczynnika korelacji rang Spearmana,
  - współczynnika korelacji nieliniowej.

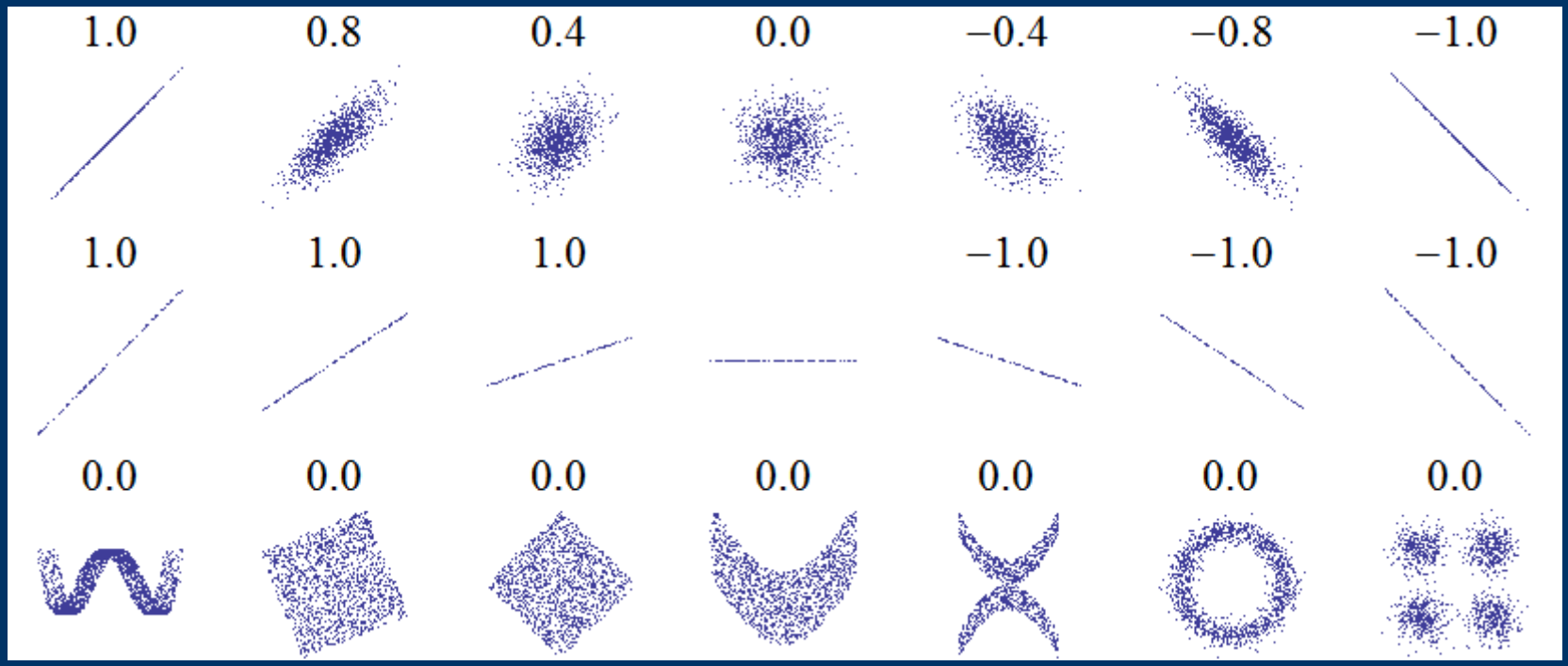
Współczynnik korelacji Pearsona (1)

- Jest to najbardziej popularny współczynnik określający poziom zależności liniowej między zmiennymi losowymi.
- Jest to iloraz kowariancji i iloczynu odchyleń standardowych badanych cech x, y:

$$r_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Kowariancja jest średnią arytmetyczną iloczynu odchyleń poszczególnych zmiennych od ich średnich arytmetycznych.

Współczynnik korelacji Pearsona (2)



Przykładowe wykresy danych (x, y) oraz odpowiadające im wartości współczynnika korelacji liniowej Pearsona.

Współczynnik korelacji Pearsona (3)

- Współczynnik r_{xy} przyjmuje wartość od –1 do +1, im większa jego wartość bezwzględną tym większa siła korelacji.
  - $r_{xy} = 0$ nie zawsze oznacza brak zależności, a jedynie brak zależności liniowej.
  - $r_{xy} = \pm 1$ oznacza ścisły dodatni/ujemny związek.
- Współczynnik r_{xy} wskazuje na korelację wzajemną: x względem y i y względem x.

Raz jeszcze warto przypomnieć: związek korelacji nie jest jednoznaczny z występowaniem związku przyczynowo-skutkowego, a badanie związku korelacji wymaga rozeznania merytorycznej strony badanych zjawisk. Współzmienność nie jest dowodem współzależności!

Współczynnik determinacji (określoności)

- Jest to kwadrat współczynnika korelacji r_{xy}^{2}.
- Jest opisową miarą siły związku między zmiennymi.
- Informuje on o tym, jaka część zmienności cechy zależnej *y* jest wyjaśniona zmiennością cechy niezależnej *x*.
    - $(1-r_{xy}^{2})$ nazywają współczynnikiem indeterminacji, bo informuje on o tym, jaka część zmienności nie została wyjaśniona.

Przykład: jeśli $r_{xy}=0,8$ to $r_{xy}^{2}=0,64$, to oznacza, że w 64% zmianę wartości *y* wyjaśnia zmiana *x*.

Weryfikacja hipotezy o istotności korelacji

- Jeśli rozkład zmiennych losowych Y i X jest normalny, to na podstawie próby z n elementów możemy zweryfikować hipotezę, że zmienne te są liniowo niezależne: $H_0 : \rho = 0$
- Jeżeli H_0 jest prawdziwa, to statystyka: $t = \frac{r}{\sqrt{1-r^2}}\sqrt{n-2}$ ma rozkład t Studenta z liczbą stopni swobody $v = n - 2$.

Analiza regresyjna

Istota analizy regresyjnej (1)

- Jest to badanie wpływu jednej lub kilku cech „objaśniających” na cechę, której kształtowanie się najbardziej nas interesuje, a więc na cechę „objaśnianą”.
- Metody regresji używane są zazwyczaj do opisu kształtowania się poziomu pewnego zjawiska w czasie, jak i na podstawie pobieranych z populacji generalnej prób losowych.

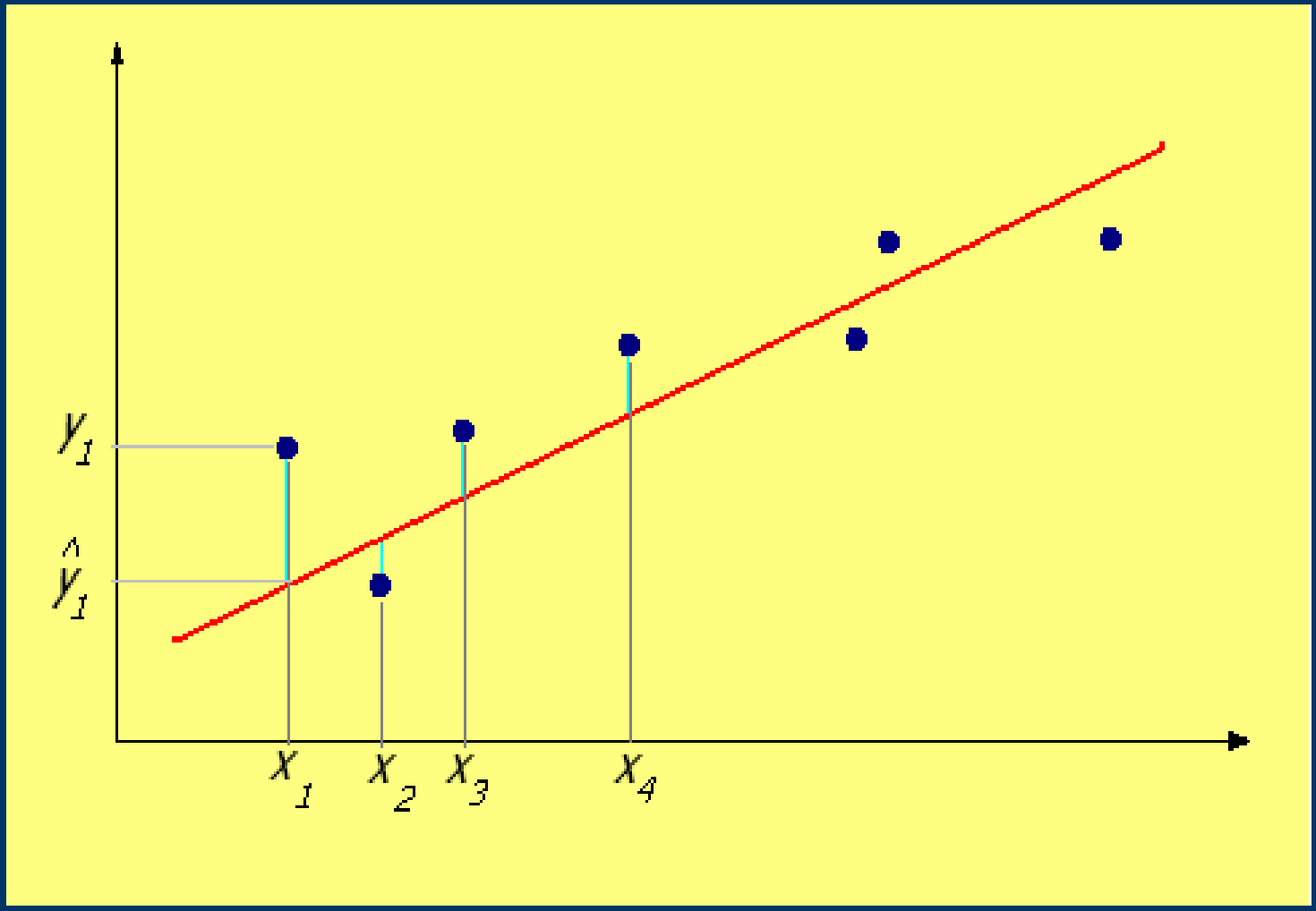
Istota analizy regresyjnej (2)

- Analiza regresji polega na estymacji parametrów równania teoretycznego, które odwzorowuje zależność.
- Ilustracją regresji jest wykres wartości rzeczywistych i teoretycznych zmiennej objaśnianej.
- Podstawowe modele regresji zakładają występowanie zależności liniowych istniejących pomiędzy zmienną objaśnianą, a zmiennymi ją objaśniającymi.

Analiza regresyjna: proces doboru zmiennych

1. Na podstawie wiedzy merytorycznej ustal listę potencjalnych zmiennych objaśniających.
2. Zbierz dane statystyczne, będące realizacjami zmiennej objaśnianej i zmiennych objaśniających.
3. Eliminuj zmienne objaśniające o zbyt niskim poziomie zmienności.
4. Oblicz współczynniki korelacji między zmiennymi.
5. Przeprowadź redukcję zbioru zmiennych.

Estymacja parametrów modelu



Nieznane parametry modelu $y = b + ax$ muszą być estymowane na podstawie próby losowej poprzez takie dobranie parametrów a, b aby suma kwadratów odległości każdego punktu empirycznego od prostej regresji była jak najmniejsza.

Metoda najmniejszych kwadratów

$$Y = a\,x + b$$

$$\Sigma\,[Y_i - (a\,x_i + b)]^2 = min$$

Zmienna objaśniana:

- dane teoretyczne z równania regresji Y,
- dane rzeczywiste Y_i.

Zmienna objaśniająca: x_i

Parametry strukturalne równania regresji:

- a - współczynnik regresji,
- b - wyraz wolny (tzw. parametr skali); podaje wartość zmiennej Y, gdy zmienna x przybiera wartość zero.

Równanie regresji

- Gdy obliczymy parametry równania *a* i *b*, otrzymamy empiryczne równanie regresji wyprowadzone z konkretnego szeregu danych statystycznych.
- Mając to równanie możemy obliczać zmienną zależną (objaśnianą) podstawiając konkretną wartość zmiennej niezależnej (objaśniającej).
- Wyniki te możemy wykorzystać do prognozowania kształtowania się konkretnego zjawiska w konkretnej przyszłości, badania wariantów rozwojowych.

Istotność równania regresji (1)

Istotność równania regresji ustalamy, weryfikując hipotezę zerową $H_0 : a = 0$ wobec $H_1 : a \neq 0$.

Przy prawdziwości H_0 statystyka: $t = \frac{\hat{a}}{s_{\hat{b}}} = \frac{\hat{a}}{\sqrt{\frac{s^2_{y/x}}{\operatorname{var} x}}}$

ma rozkład t Studenta z liczbą stopni swobody równej n - 2. Wyrażenie $s^2_{y/x}$ jest oszacowaniem wariancji odchyleń od regresji z próby.

Istotność równania regresji (2)

- Z tablic rozkładu Studenta odczytujemy, dla wcześniej przyjętego poziomu istotności α, wartość krytyczną $t_{kryt} = t_{n-2,\alpha}$.
- Jeżeli $|t| > t_{kryt}$, to hipotezę $H_0 : a = 0$ odrzucamy jako statystycznie mało prawdopodobną i mówimy o istotności wyznaczonego równania regresji.

- Jeżeli $|t| < t_{kryt}$, to wyniki próby nie przeczą hipotezie H_0 i funkcja regresji nie zależy od x.