Statystyka opisowa

Robert Pietrzykowski

email: robert_pietrzykowski@sggw.pl

www.ekonometria.info

Na dziś…

- Sprawy bieżące

Na dziś…

- Wykład 5:
  - Badanie zależności cech

Zmienne losowe X i Y są niezależne wtedy i tylko wtedy gdy $\forall (x, y) \in R^2$:

$$F_{XY}(x, y) = F_X(x) F_Y(y).$$

Niech $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ będzie próbą

Współczynnik korelacji

Opis jakościowy zależności

$$\varrho = \frac{E(X - EX)(Y - EY)}{DX \cdot DY}$$

	1	2	3	4	5	6
1	0	0	1/12	1/12	0	0
2	0	1/12	0	0	1/12	0
3	1/12	0	0	0	0	1/12
4	1/12	0	0	0	0	1/12
5	0	1/12	0	0	1/12	0
6	0	0	1/12	1/12	0	0

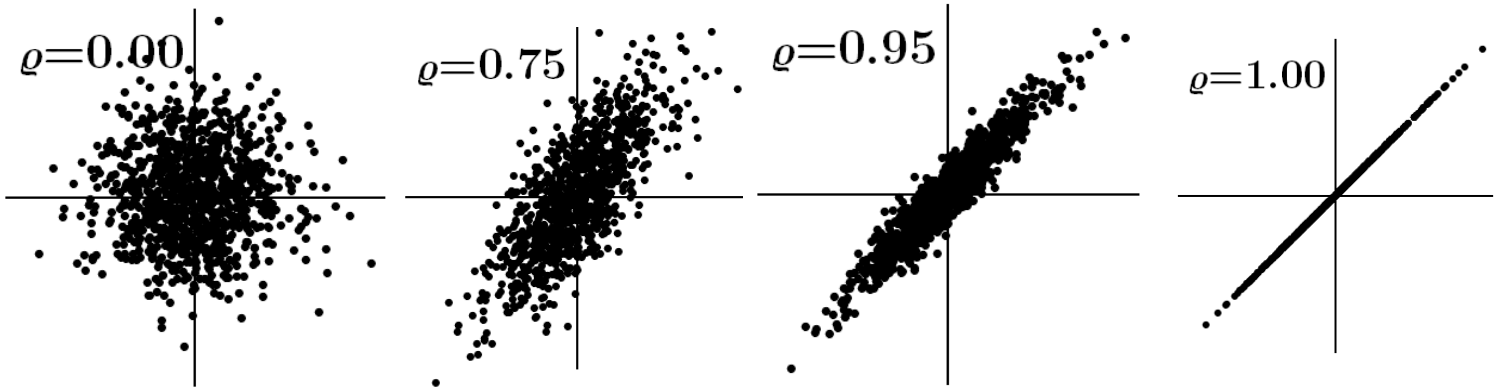
$$P\{X = 1 \ \& \ Y = 2\} \neq P\{X = 1\}P\{Y = 2\}$$

$$EX = 3.5 \qquad EY = 3.5$$

$$DX = 1.7078 \qquad DY = 1.7078$$

$$E(X - EX)(Y - EY) = 0$$

Współczynnik korelacji jest miernikiem zależności między dwiema cechami
Oznaczenie: ϱ



Im $|\varrho|$ jest bliższe 1, tym bardziej „liniowa” jest zależność między cechami.

Współczynnik korelacji jest miernikiem **liniowej** zależności między cechami X oraz Y.

Własności współczynnika korelacji

1. Współczynnik korelacji jest liczbą niemianowaną
2. $\varrho \in \langle -1, 1 \rangle$
3. Jeżeli $\varrho > 0$, to większym wartościom jednej cechy odpowiadają (średnio) większe wartości drugiej cechy. Zależność dodatnia (rosnąca, stymulująca).
4. Jeżeli $\varrho < 0$, to większym wartościom jednej cechy odpowiadają (średnio) mniejsze wartości drugiej cechy. Zależność ujemna (malejąca, limitująca).
5. Jeżeli $\varrho = 0$, to bez względu na wartości przyjmowane przez jedną z cech, średnie wartości drugiej cechy są takie same. Cechy nieskorelowane.
7. Jeżeli (X, Y) ma dwuwymiarowy rozkład normalny, to $\varrho = 0$ jest równoważne **niezależności** cech X, Y.

współczynnik korelacji Pearsona

współczynnik korelacji rangowej Spearmana
współczynnik korelacji rangowej Kendalla

test chi–kwadrat niezależności

Współczynnik korelacji Pearsona

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim N_2 \left(\begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \begin{bmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{bmatrix} \right)$$

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

Niech $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ będzie próbą

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

$$R = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

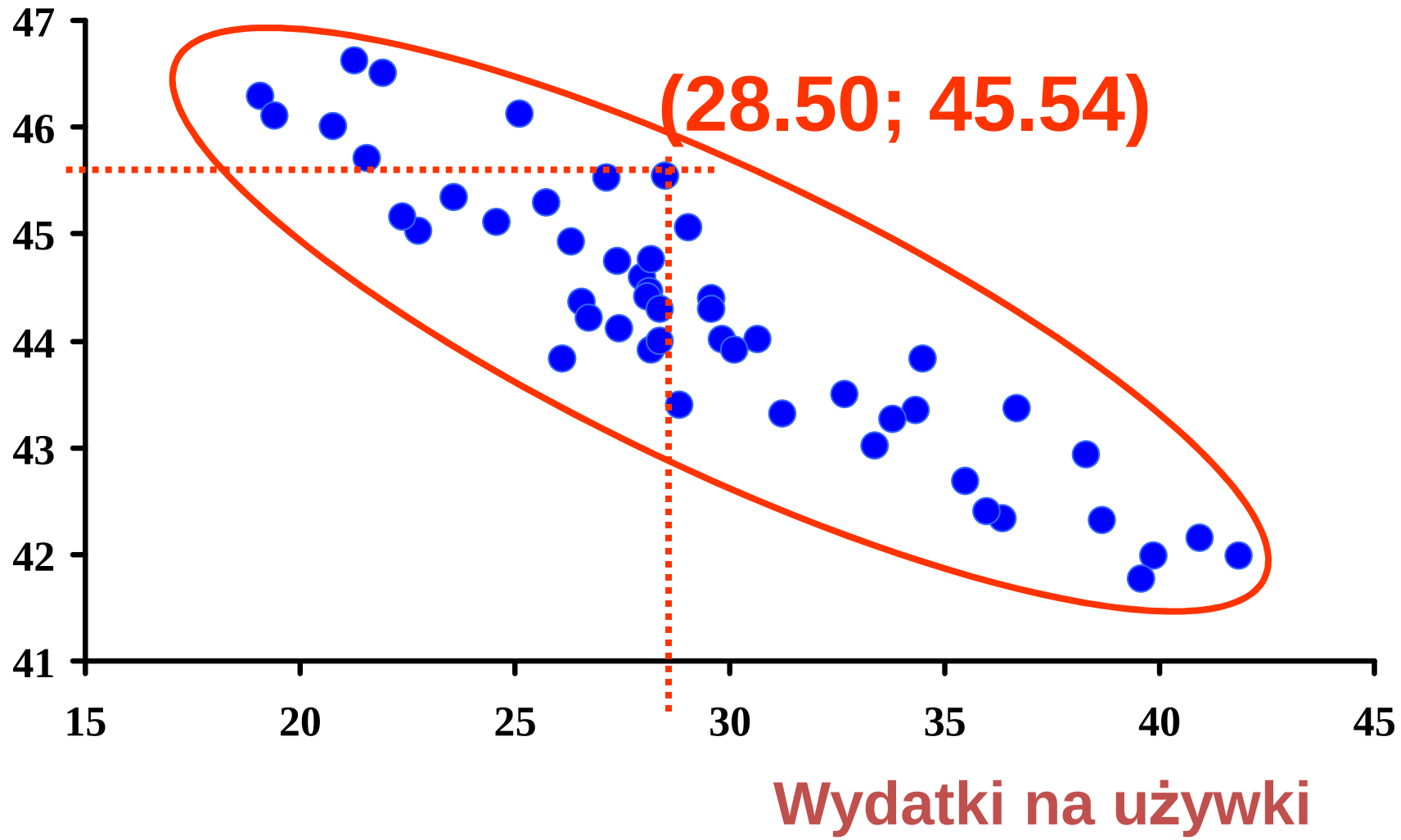
Wartość współczynnika korelacji by Pietrzykowski

- Poniżej 0,2 praktycznie brak korelacji
- Od 0,2 do 0,5 słaba korelacja
- Od 0,5 do 0,8 średnia korelacja
- od 0,8 do 0,9 silna korelacja
- Od 0,9 do 1,0 bardzo silna korelacja

Przykład. W pewnej rodzinie obserwowano tygodniowe wydatki na używki (Uż) i artykuły spożywcze (Sp). Na podstawie poniższych danych zbadać istnienie zależności. Jeżeli taka zależność istnieje, to opisać ją ilościowo.

Uż	Sp	Uż	Sp	Uż	Sp	Uż	Sp	Uż	Sp
				28.37	44.00	38.31	42.92	22.78	45.03
				28.15	44.46	21.94	46.50	25.76	45.29
31.22	43.31	20.77	46.01	36.71	43.36	32.69	43.50	22.39	45.16
36.38	42.33	25.11	46.12	29.57	44.39	34.51	43.82	28.19	44.76
35.99	42.40	26.13	43.82	29.07	45.05	39.59	41.77	29.84	44.01
38.67	42.31	19.41	46.10	27.43	44.11	29.58	44.29	30.14	43.91
19.08	46.28	27.16	45.52	39.86	41.98	27.38	44.74	28.39	44.29
28.83	43.39	27.98	44.59	34.33	43.34	33.38	43.01	40.97	42.14
35.48	42.68	30.67	44.01	41.88	41.98	28.09	44.40	21.29	46.61
24.57	45.10	28.17	43.91	26.73	44.20	33.79	43.26	26.32	44.92

(28.50; 45.54)


Wydatki na art. spożywcze
47
46
45
44
43
42
41
15
20
25
30
35
40
45
(28.50; 45.54)
Wydatki na używki

Cechy:

X: tygodniowe wydatki na używki

Y: tygodniowe wydatki na artykuły spożywcze

Założenie:

normalność rozkładów badanych cech

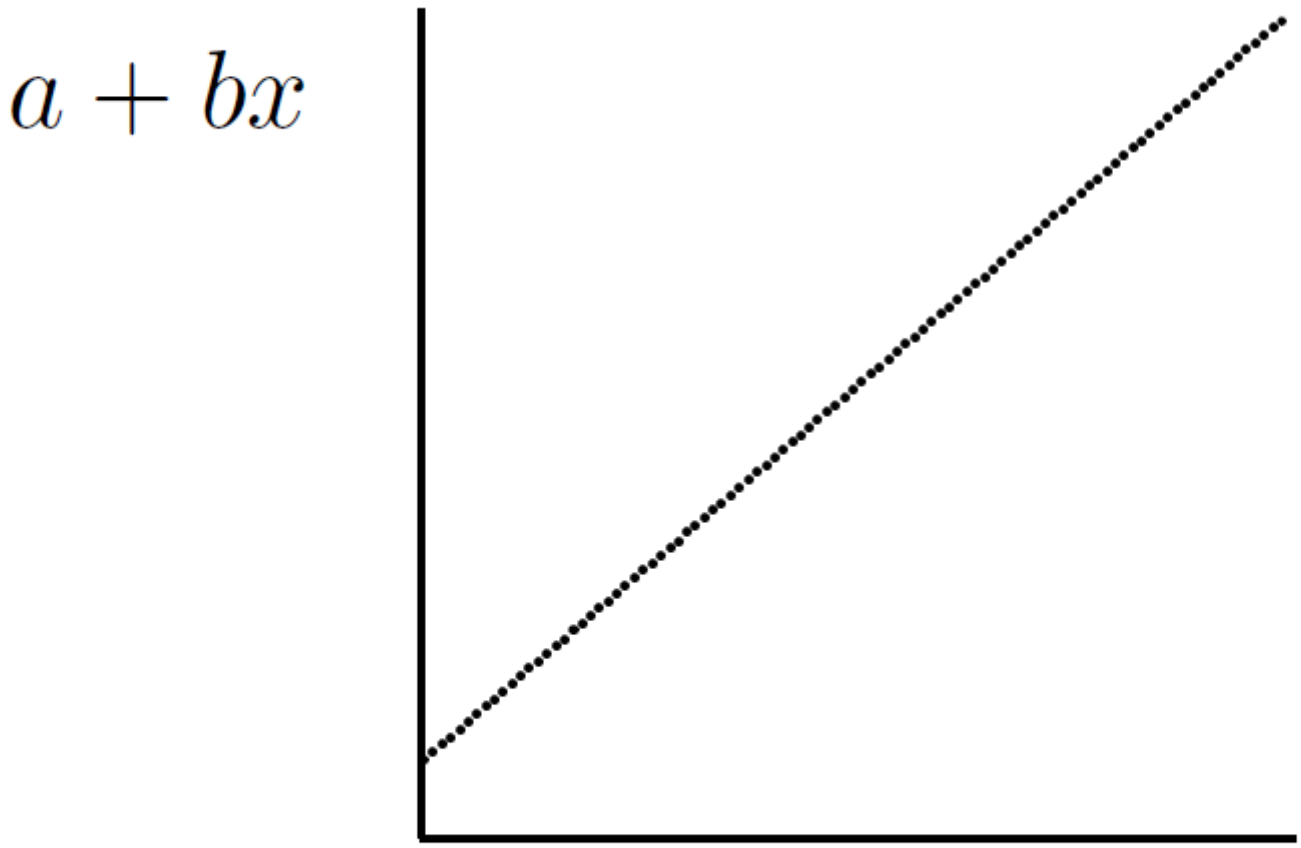
Weryfikacja hipotezy i wnioskowanie:

Próbkowy współczynnik korelacji

$$R = \frac{-335.789252}{\sqrt{1657.907048} \cdot \sqrt{79.404698}} = -0.9255$$

Wniosek.

Wydatki na używki (X) i wydatki na artykuły spożywcze (Y) są od siebie zależne. Ponieważ współczynnik korelacji jest ujemny, więc zależność ma charakter malejący, tzn. im większe są wydatki na używki, tym mniejsze (średnio) na artykuły spożywcze.



regresja liniowa

Funkcja regresji $E(Y|X = x) = \beta_0 + \beta_1 x$

Model

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \ \ i = 1, \ldots, n,$$

ε_i są niezależnymi zmiennymi losowymi o tym samym rozkładzie normalnym $N(0, \sigma^2)$.

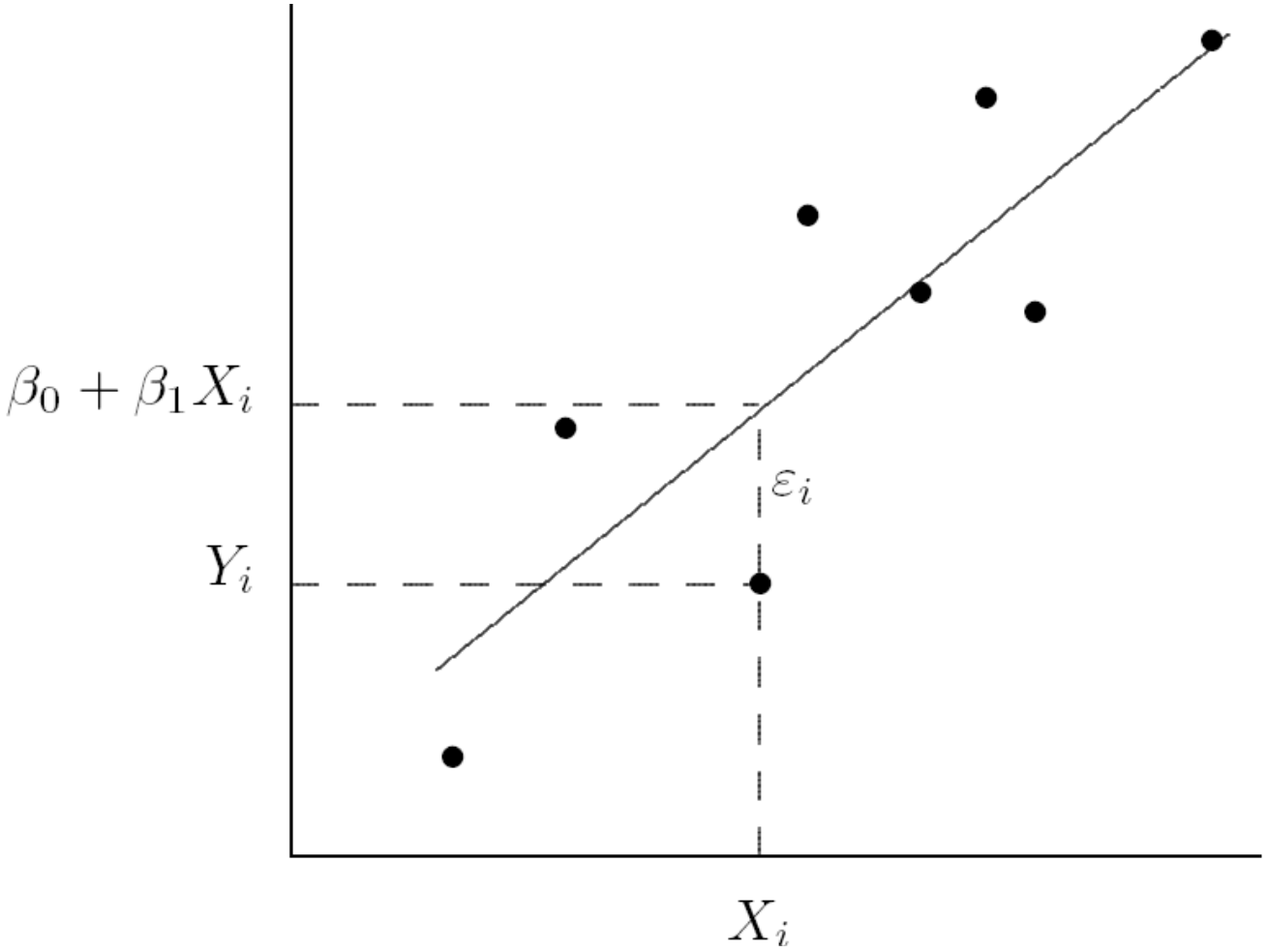
Estymacja współczynników metodą najmniejszych kwadratów

Znaleźć takie β_0 i β_1 by

$$\sum_{i=1}^{n}(Y_i - (\beta_0 + \beta_1 x_i))^2 = \min$$

Szukamy takich parametrów β_0, β_1 aby zminimalizować sumę kwadratów reszt, tzn.

$$\sum_{i=1}^{n} \varepsilon_i^2 = \min$$



$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}, \qquad \widehat{\beta}_0 = \bar{Y} - \widehat{\beta}_1 \bar{X}$$

Przykład. W pewnej rodzinie obserwowano tygodniowe wydatki na używki (Uż) i artykuły spożywcze (Sp). Na podstawie poniższych danych zbadać istnienie zależności. Jeżeli taka zależność istnieje, to opisać ją ilościowo.

				Uż	Sp	Uż	Sp	Uż	Sp
				28.37	44.00	38.31	42.92	22.78	45.03
				28.15	44.46	21.94	46.50	25.76	45.29
31.22	43.31	20.77	46.01	36.71	43.36	32.69	43.50	22.39	45.16
36.38	42.33	25.11	46.12	29.57	44.39	34.51	43.82	28.19	44.76
35.99	42.40	26.13	43.82	29.07	45.05	39.59	41.77	29.84	44.01
38.67	42.31	19.41	46.10	27.43	44.11	29.58	44.29	30.14	43.91
19.08	46.28	27.16	45.52	39.86	41.98	27.38	44.74	28.39	44.29
28.83	43.39	27.98	44.59	34.33	43.34	33.38	43.01	40.97	42.14
35.48	42.68	30.67	44.01	41.88	41.98	28.09	44.40	21.29	46.61
24.57	45.10	28.17	43.91	26.73	44.20	33.79	43.26	26.32	44.92

(28.50; 45.54)

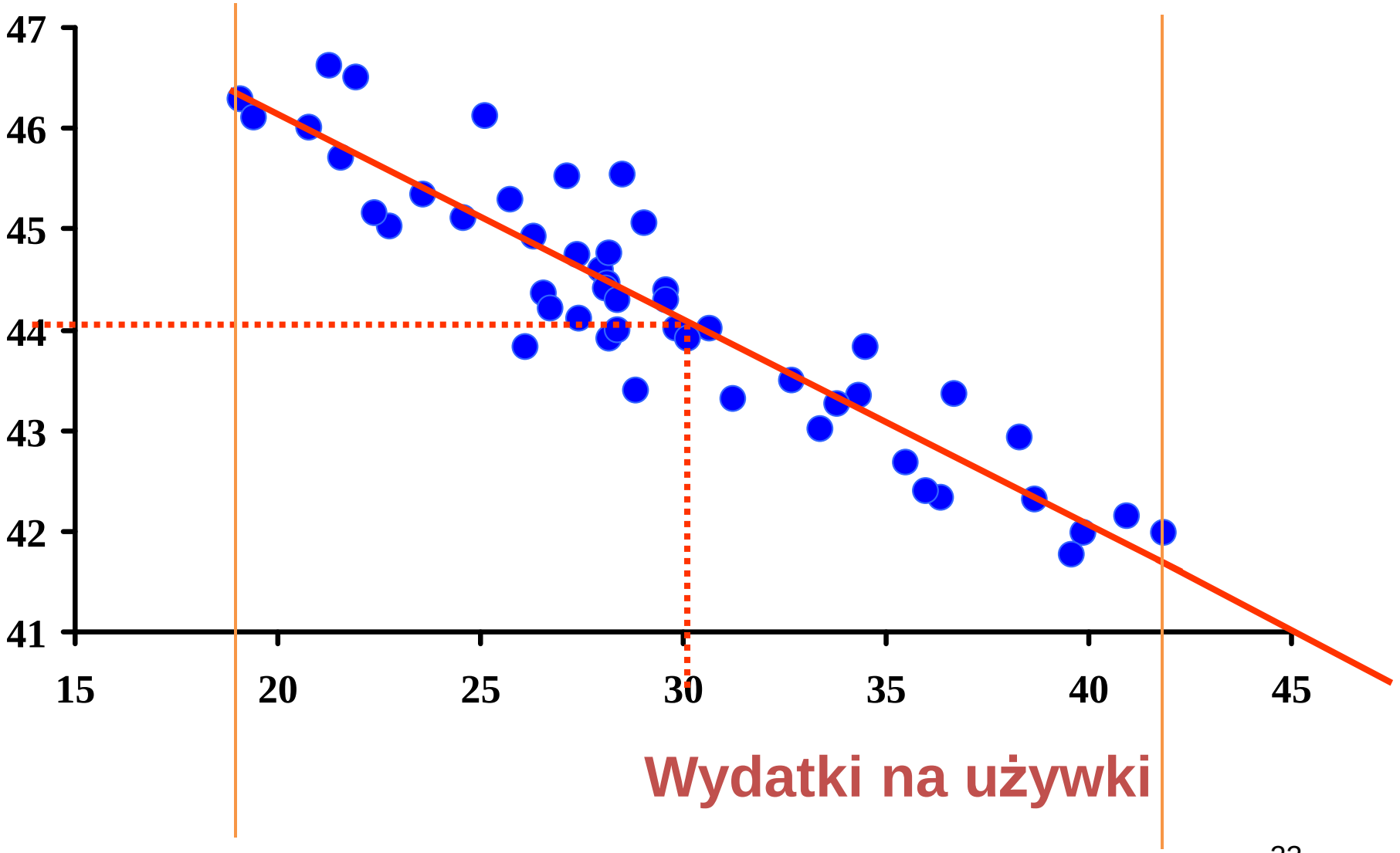
Ilościowy opis zależności

$$E(Y|X = x) = \beta_0 + \beta_1 x$$

$$\text{średni } y = 50.1680 - 0.2025x$$

Zależność ta obowiązuje w zakresie od: minimalnej wartości wydatków na używki 19.08 do maksymalnej wartości wydatków na używki 41.88

średnie wydatki na art. spożywcze =
50,1680 - 0,2025 wydatki na używki



PYTANIA

1. Co mierzy współczynnika korelacji?
2. Podaj interpretacje współczynnika korelacji
3. W dwudziestu gospodarstwach wiejskich badano zależność między spożyciem ziemniaków (cecha X) i artykułów zbożowych (cecha Y). Uzyskano współczynnik korelacji r=1,9983. Czy można na tej podstawie przyjąć, że istnieje zależność między spożyciem ziemniaków i artykułów zbożowych?
4. Co to znaczy jeżeli współczynnik korelacji przyjmuje wartość zerową?
5. Narysuj na oddzielnych wykresach uwzględniając prostą regresji, przypadek braku korelacji, korelację ujemną i dodatnią.
6. Kiedy współczynnik korelacji Pearsona jest złą miarą do określenia korelacji między dwiema cechami.
7. Podaj interpretację współczynnika regresji.
8. Wyjaśnij na czym polega metoda MNK.
9. Wyjaśnij co mierzy współczynnik determinacji.
10. Zapisz postać liniowego modelu funkcji regresji.