

Wykład z Ekonometrii

Robert Pietrzykowski

robert.pietrzykowski@ekonometria.info

www.ekonometria.info

EKONOMETRIA:

Nauka zajmująca się ustalaniem za pomocą metod matematyczno–statystycznych ilościowych prawidłowości w życiu gospodarczym.

Encyklopedia Popularna PWN, Warszawa 1982

Dział ekonomii, który zajmuje się mierzaniem zależności występujących między różnymi wielkościami ekonomicznymi

Mały Ilustrowany Leksykon PWN, Warszawa 1997

Podstawowe pojęcia

Populacja.

Zbiorowość z przynajmniej jedną wspólną własnością i przynajmniej jedną cechą, której wartościami jednostki populacji mogą się różnić.

Zmienna losowa.

Cecha, której wartości są przyjmowane przez jednostki populacji w sposób losowy.

Rozkład prawdopodobieństwa.

Określa prawdopodobieństwo przyjęcia wartości przez zmienną losową.

Próba.

Losowo wybrana część populacji na podstawie której wnioskujemy o rozkładzie prawdopodobieństwa zmiennej losowej w całej populacji.

Hipoteza statystyczna.

Dowolne przypuszczenie dotyczące rozkładu prawdopodobieństwa zmiennej losowej.

Test statystyczny.

Procedura statystyczna mająca na celu weryfikację hipotezy statystycznej (odrzuć, nie odrzuć).

Błąd I rodzaju.

Odrzucenie hipotezy statystycznej, która w rzeczywistości jest prawdziwa.

Błąd II rodzaju.

Nie odrzucenie hipotezy statystycznej, która w rzeczywistości jest fałszywa.

Poziom istotności.

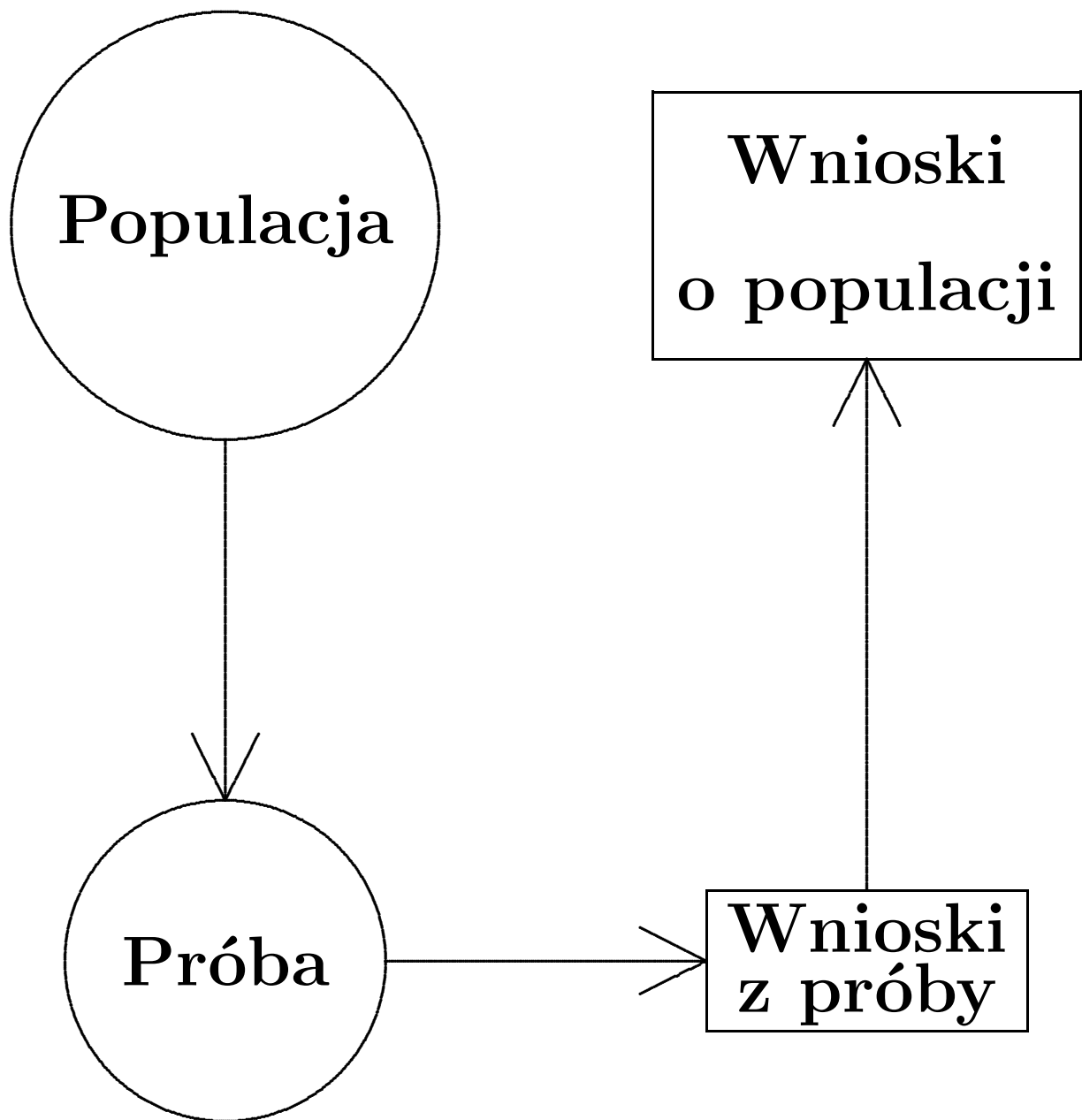
Dopuszczane ryzyko popełnienia błędu I rodzaju.

Estymacja punktowa.

Oszacowanie na podstawie próby za pomocą jednej liczby nieznaney wartości parametru zmiennej losowej.

Estymacja przedziałowa.

Wyznaczenie na podstawie próby przedziału pokrywającego nieznaną wartość parametru ze z góry zadanyim prawdopodobieństwem (poziomem ufności).

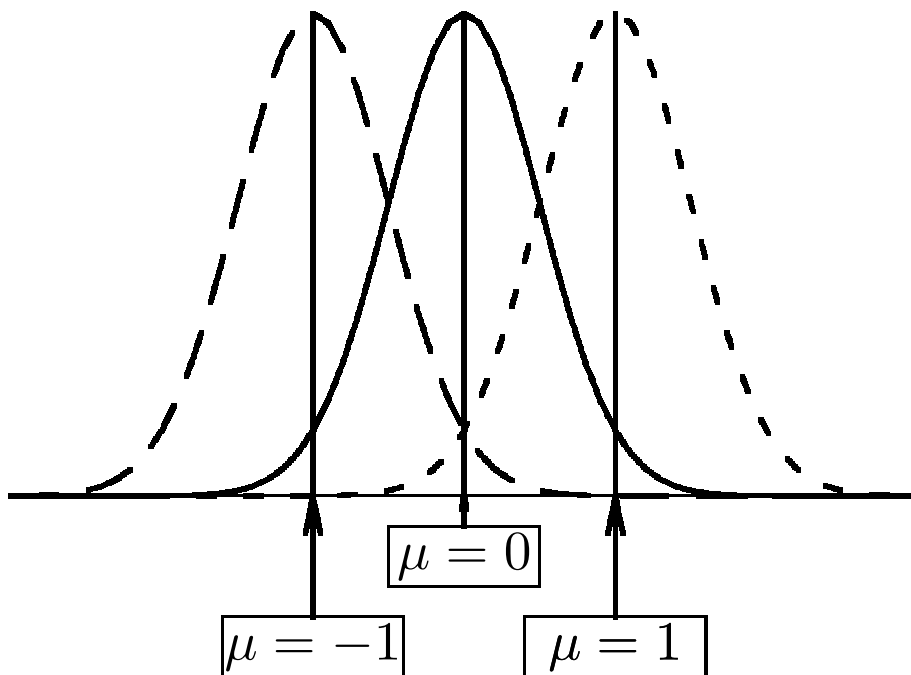


Rozkład normalny

Zmienna losowa X ma rozkład normalny $N(\mu, \sigma^2)$ o wartości średniej μ i wariancji σ^2 , jeżeli jej funkcja gęstości wyraża się wzorem

$$f_{\mu, \sigma^2}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad -\infty < x < \infty.$$

$$EX = \mu \quad D^2X = \sigma^2.$$



Y_1, \dots, Y_m — zmienne zależne

X_1, \dots, X_p — zmienne niezależne (czynniki)

Czy istnieje zależność między Y -ami a X -ami?

Jaki jest charakter zależności?

Ilościowe przedstawienie zależności.

Analiza dynamiki zjawisk.

Założenia

1. Wszystkie zmienne są zmiennymi ilościowymi.
2. Zmienne zależne są losowe o rozkładach normalnych.
3. Czynniki są losowe o rozkładach normalnych.
4. Czynniki są deterministyczne.

Rozważane sytuacje

Y_1, \dots, Y_m — zmienne zależne

X_1, \dots, X_p — zmienne niezależne (czynniki)

1. $m = 1, p = 1$

2. $m = 1, p = 2$

Losowe czynniki

$$m = 1 \quad p = 1$$

Obserwujemy dwie cechy: X oraz Y
Obiekt $\longrightarrow (X, Y)$

1. Czy cechy X oraz Y są niezależne?
2. Opis ilościowy zależności.
3. Wnioski.

Założenie:

Łączny rozkład cech X, Y jest normalny

Współczynnik korelacji

Opis jakościowy zależności

Współczynnik korelacji jest miernikiem zależności między dwiema cechami

Oznaczenie: ρ

Własności współczynnika korelacji

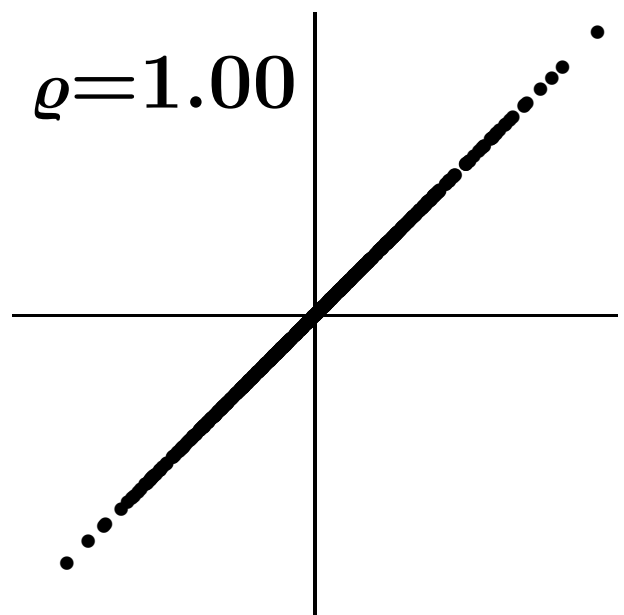
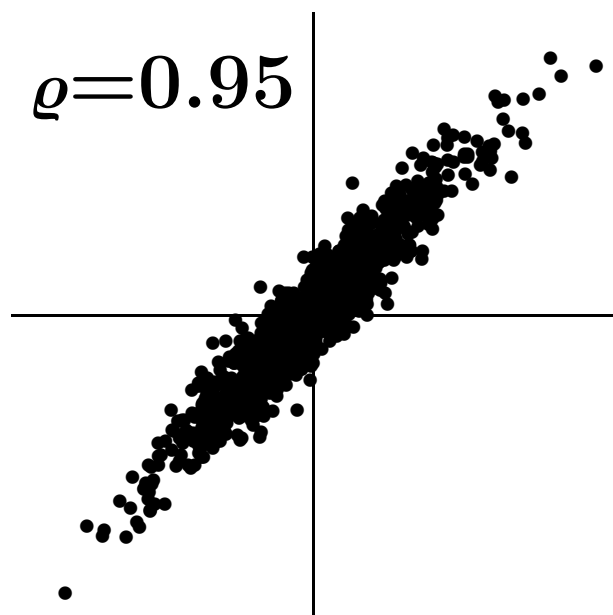
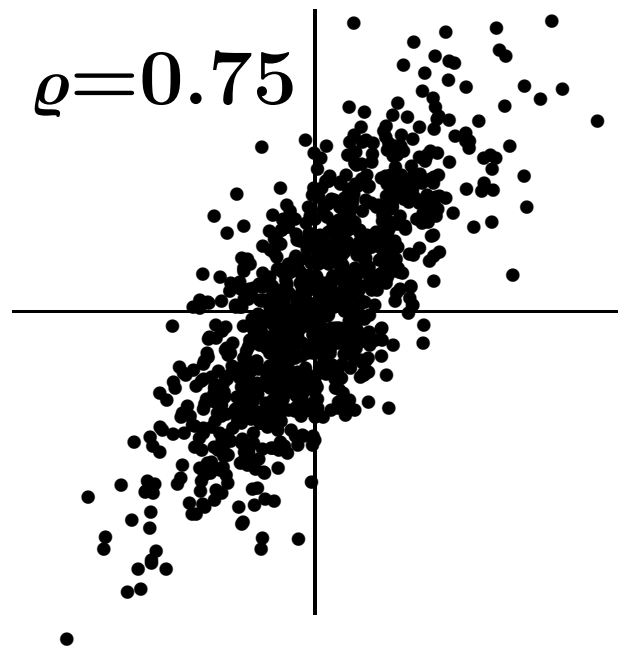
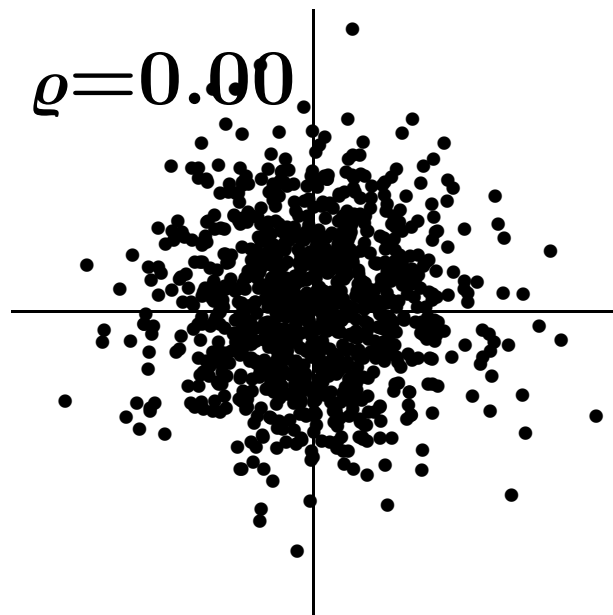
1. Współczynnik korelacji jest liczbą niemianowaną
2. $\rho \in \langle -1, 1 \rangle$
3. Jeżeli $\rho > 0$, to większym wartościom jednej cechy odpowiadają (średnio) większe wartości drugiej cechy. Zależność dodatnia (rosnąca, stymulująca).
4. Jeżeli $\rho < 0$, to większym wartościom jednej cechy odpowiadają (średnio) mniejsze wartości drugiej cechy. Zależność ujemna (malejąca, limitująca).

5. Jeżeli $\rho = 0$, to bez względu na wartości przyjmowane przez jedną z cech, średnie wartości drugiej cechy są takie same. Cechy nieskorelowane.
6. Jeżeli $\rho = \pm 1$, to istnieją takie liczby rzeczywiste a oraz b , że $Y = aX + b$. Jeżeli $\rho = 1$, to $a > 0$. Jeżeli $\rho = -1$, to $a < 0$.

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

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

7. Jeżeli (X, Y) ma dwuwymiarowy rozkład normalny, to $\rho = 0$ jest równoważne **niezależności** cech X, Y .



Współczynnik korelacji między cechami X i Y

$$\rho = \frac{\text{kowariancja}(X, Y)}{\sqrt{D^2 X \cdot D^2 Y}}$$

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

Współczynnik korelacji z próby (próbkowy)

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

Współczynnik korelacji Pearsona

Suma iloczynów odchyleń

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

$$\sum_{i=1}^n X_i Y_i - n\bar{X}\bar{Y} = \sum_{i=1}^n X_i Y_i - \frac{1}{n} \left(\sum_{i=1}^n X_i \right) \left(\sum_{i=1}^n Y_i \right)$$

$$H_0 : \text{Cechy } X \text{ oraz } Y \text{ są niezależne}$$
$$\Updownarrow$$
$$H_0 : \varrho = 0$$

Test współczynnika korelacji Pearsona (poziom istotności α)

Próba: $(X_i, Y_i), i = 1, \dots, n$
Statystyka testowa

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

Wartość krytyczna $r(\alpha; n)$ (dwustronna)

Jeżeli $|R| > r(\alpha; n)$,
to hipotezę H_0 odrzucamy.

Ilościowy opis zależności

Zakładamy niezerowość współczynnika korelacji:

$$\rho \neq 0$$

Ilościowy opis zależności Y od X :

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

Funkcja f nosi nazwę **funkcji regresji**

Przy założeniu normalności

$$f(x) = \beta_0 + \beta_1 x$$

β_1 — współczynnik regresji

β_0 — stała regresji

Zadanie: oszacować parametry funkcji regresji

Ocena parametrów funkcji regresji

Próba $(X_1, Y_1), \dots, (X_n, Y_n)$

Niezbędne rachunki

$$\bar{X}, \text{var}X, \bar{Y}, \text{var}Y, \text{cov}(X, Y)$$

$$\hat{\beta}_1 = \frac{\text{cov}(X, Y)}{\text{var}X} \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

Resztowa suma kwadratów

$$RSS = \text{var}Y(1 - R^2)$$

Resztowa suma kwadratów reprezentuje rozrzut obserwacji cechy Y wokół dopasowanej funkcji regresji

Wariancja resztowa

$$S^2 = \frac{RSS}{n - 2}$$

Przedział ufności dla współczynnika regresji (poziom ufności $1 - \alpha$)

Wariancja estymatora $\hat{\beta}_1$

$$S_{\beta_1}^2 = \frac{S^2}{\text{var}X}$$

Przedział ufności

$$\beta_1 \in (\hat{\beta}_1 - t(\alpha; n - 2)S_{\beta_1}; \hat{\beta}_1 + t(\alpha; n - 2)S_{\beta_1})$$

Przedział ufności dla stałej regresji (poziom ufności $1 - \alpha$)

Wariancja estymatora $\hat{\beta}_0$

$$S_{\beta_0}^2 = \frac{S^2}{\text{var}X} \left(\frac{\text{var}X}{n} + \bar{X}^2 \right)$$

Przedział ufności

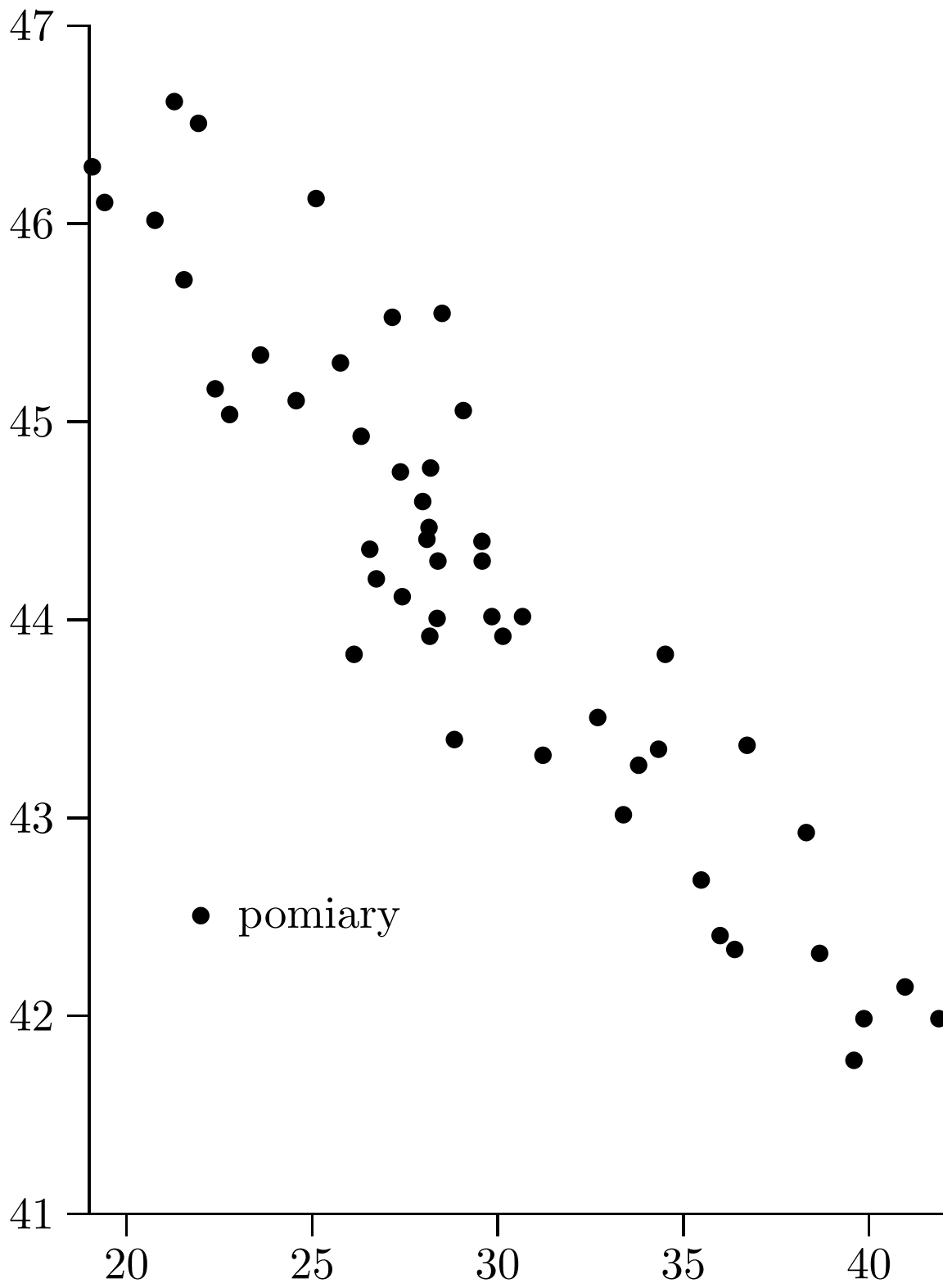
$$\beta_0 \in (\hat{\beta}_0 - t(\alpha; n - 2)S_{\beta_0}; \hat{\beta}_0 + t(\alpha; n - 2)S_{\beta_0})$$

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. Jakich średnich wydatków na artykuły spożywcze możemy się spodziewać jeżeli na używki wydajemy 30?

Uż	Sp	Uż	Sp	Uż	Sp	Uż	Sp	Uż	Sp
28.50	45.54	26.55	44.35	28.37	44.00	38.31	42.92	22.78	45.03
23.61	45.33	21.55	45.71	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

Plan działania

1. Istnienie zależności
2. Ilościowy opis
3. Wnioski



Populacja:

Cechy:

Uż-(X): tygodniowe wydatki na używki

Sp-(Y): tygodniowe wydatki na artykuły spożywcze

Założenie:

normalność rozkładów badanych cech

Techniki statystyczne:

1. Weryfikacja hipotezy $H_0 : \rho = 0$
2. Dopasowanie funkcji regresji $y = \beta_0 + \beta_1 x$

Poziom istotności $\alpha = 0.05$

Poziom ufności $1 - \alpha = 0.95$

Weryfikacja hipotezy $H_0 : \rho = 0$

Obliczenie próbkowego współczynnika korelacji

$$R = \frac{\text{cov}(x, y)}{\sqrt{\text{var}x} \sqrt{\text{var}y}}$$

$$n = 50$$

$$\bar{x} = 29.4652 \quad \sum x_i^2 = 45067.8076$$

$$\bar{y} = 44.2002 \quad \sum y_i^2 = 97762.2887$$

$$\sum x_i y_i = 64782.5974$$

$$\text{var}x = \sum x_i^2 - n (\bar{x})^2 = 1657.907048$$

$$\text{var}y = \sum y_i^2 - n (\bar{y})^2 = 79.404698$$

$$\text{cov}(x, y) = \sum x_i y_i - n \bar{x} \bar{y} = -335.789252$$

Próbkowy współczynnik korelacji

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

Wartość krytyczna współczynnika korelacji Pearsona

$$r(0.05; 50) = 0.2787$$

Ponieważ $|R| > r(0.05; 50)$, więc hipotezę o zerowości współczynnika korelacji odrzucamy.

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.

Ilościowy opis zależności

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

$$\hat{\beta}_1 = \frac{\text{cov}(x, y)}{\text{var}x} = \frac{-335.789252}{1657.907048} = -0.2025$$

$$\begin{aligned}\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ &= 44.2002 - (-0.2025) \cdot 29.4652 = 50.1680\end{aligned}$$

Zależność między średnimi wydatkami na artykuły spożywcze, a wydatkami na używki opisana jest wzorem

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

Równanie to obowiązuje tylko w zakresie między $\min x_i = 19.41$ a $\max x_i = 41.88$.

Wariancja resztowa

$$s^2 = \frac{\text{var}Y(1 - R^2)}{(n - 2)} = 0.2374$$

Przedziały ufności

Wartość krytyczna $t(0.05; 48) = 2.0106$

Przedział ufności dla współczynnika regresji

$$s_{\beta_1}^2 = \frac{s^2}{\text{var}x} = 0.0001432$$

$$\beta_1 \in (\hat{\beta}_1 - t(\alpha; n - 2)S_{\beta_1}; \hat{\beta}_1 + t(\alpha; n - 2)S_{\beta_1})$$

$$\beta_1 \in (-0.2266, -0.1785)$$

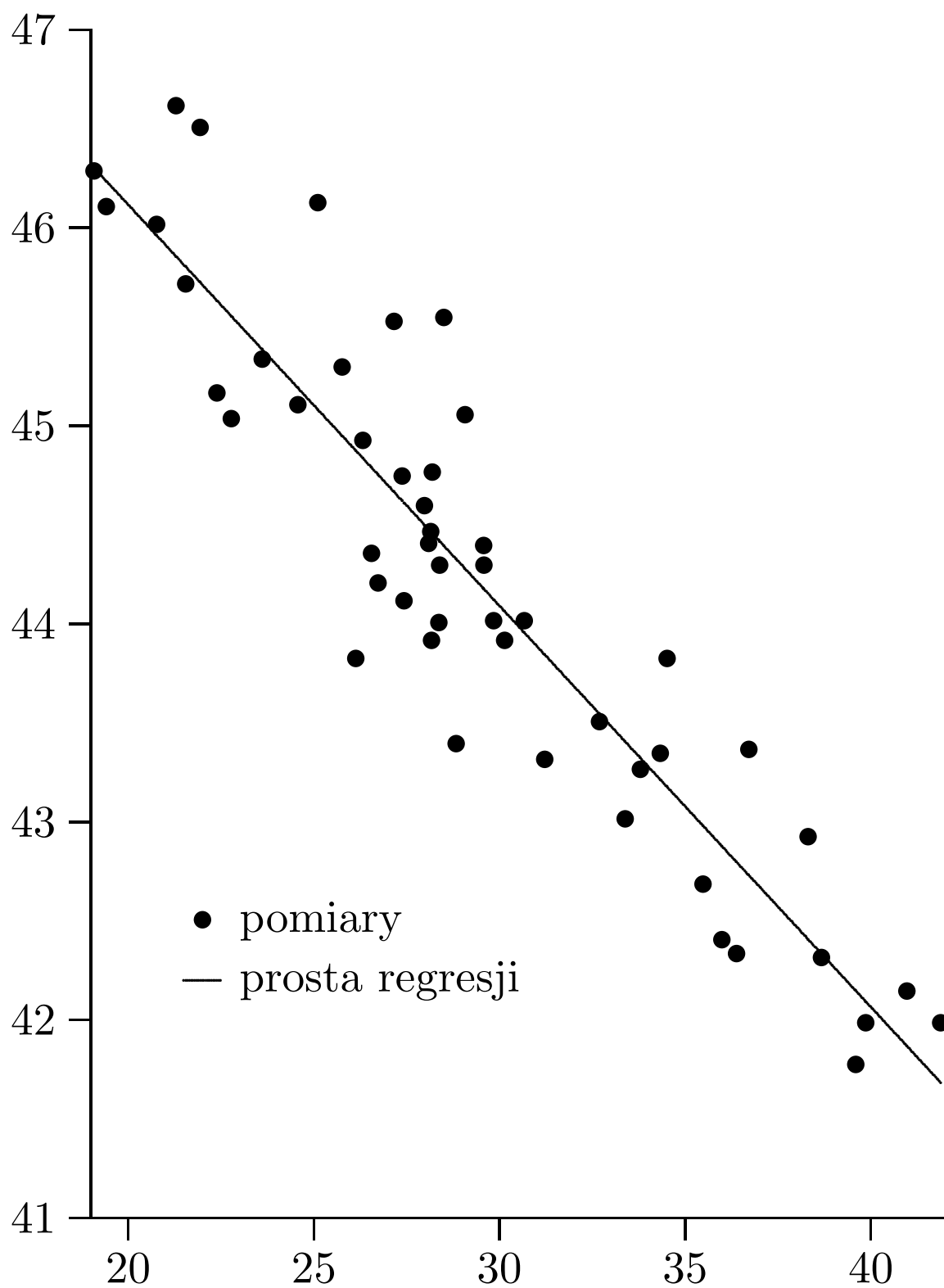
Wniosek: zwiększenie wydatków na używki o jednostkę spowoduje przeciętny spadek wydatków na artykuły spożywcze o co najmniej 0.1785, ale nie więcej niż 0.2266 (zaufanie do tak sformułowanego wniosku wynosi 95%).

Przedział ufności dla stałej regresji

$$s_{\beta_0}^2 = \frac{s^2}{\text{var}x} \left(\frac{\text{var}x}{n} + \bar{x}^2 \right) = 0.1291$$

$$\beta_0 \in (\hat{\beta}_0 - t(\alpha; n - 2)s_{\beta_0}; \hat{\beta}_0 + t(\alpha; n - 2)s_{\beta_0})$$

$$\beta_0 \in (49.4457, 50.8903)$$



Zastosowanie wyznaczonej funkcji regresji:

$$\hat{y}(x) = 50.1680 - 0.2025x$$

$$\hat{y}(30) = 50.1680 - 0.2025 \cdot 30 = 44.0930$$

Jeżeli na używki przeznaczymy 30 to na artykuły spożywcze wydamy średnio 44.0930.

Powyższe równanie nie musi dobrze opisywać zależności poza zakresem obserwowanych x -ów.

Przykład:

Jakich średnio wydatków na artykuły spożywcze możemy się spodziewać jeżeli na używki wydamy 300?

$$\hat{y}(300) = 50.1680 - 0.2025 \cdot 300 = -10.5820$$