



Statystyczna analiza danych SAD-2020/2021

Wykład 12



Analiza zależności dwóch zmiennych

Współczynnik korelacji próbkowej

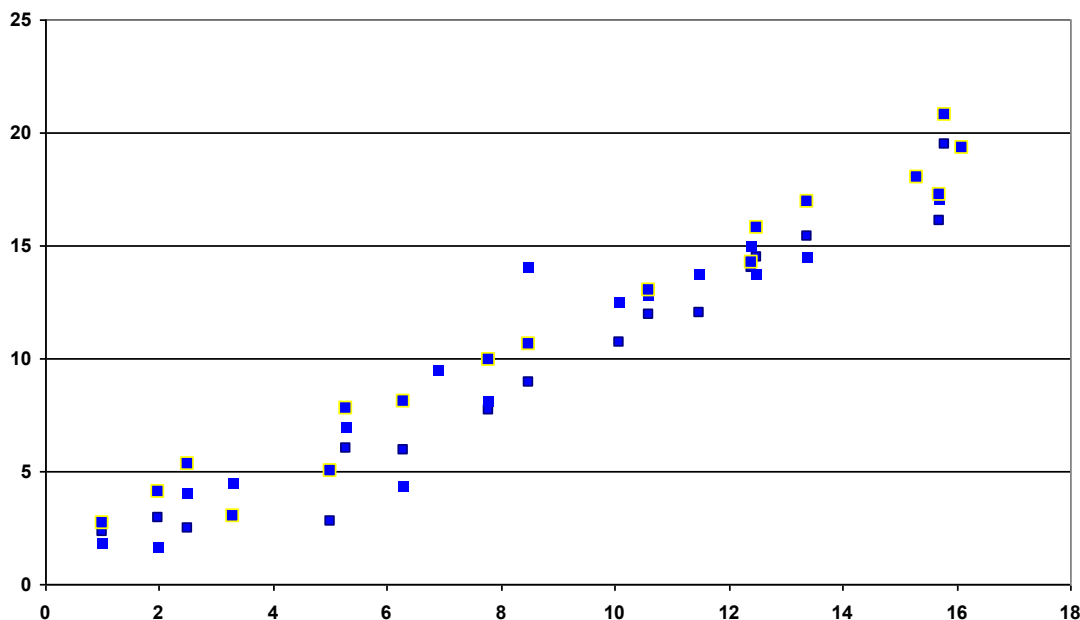
Niech $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ będzie próbką cechy dwuwymiarowej (X, Y) .

Będziemy badali zależność Y od X .

X = zmienna **niezależna** (objaśniająca),

Y = zmienna **zależna** (objaśniana),

Wykres rozproszenia – graficzne przedstawienie próbki w postaci punktów na płaszczyźnie Oxy .



Współczynnik korelacji z próby

Definicja. Niech $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ będzie próbą losową. **Współczynnikiem korelacji z próby** losowej nazywamy zmienną losową

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

gdzie \bar{X} i S_X oznaczają średnią i odchylenie standardowe dla X_1, X_2, \dots, X_n , a \bar{Y} i S_Y oznaczają średnią i odchylenie standardowe dla Y_1, Y_2, \dots, Y_n .

$$(\text{np. } \bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i, \quad S_Y^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2, \quad S_Y = \sqrt{S_Y^2})$$

Współczynnik korelacji z próby

$(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ – prosta próba losowa

$$\rho := \rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}, \quad R = \hat{\rho} = \frac{\widehat{\text{Cov}}(X, Y)}{\widehat{\sigma}_X \cdot \widehat{\sigma}_Y}$$

$\widehat{\sigma}_X := \sqrt{S_X^2}$ – estymator σ_X , bo

$$S_X^2 = \frac{1}{n-1} \sum_{j=1}^n (X_j - \bar{X})^2 \text{ – estymator } \sigma_X^2$$

Estymatorem $\text{Cov}(X, Y) = E((X - \mu_X)(Y - \mu_Y))$ jest

$$\widehat{\text{Cov}}(X, Y) = \frac{1}{n-1} \sum_{j=1}^n (X_j - \bar{X})(Y_j - \bar{Y})$$

Własności współczynnika korelacji

□ $-1 \leq \rho \leq 1$

□ Jeśli a i b są stałymi, $Y = a + bX$,

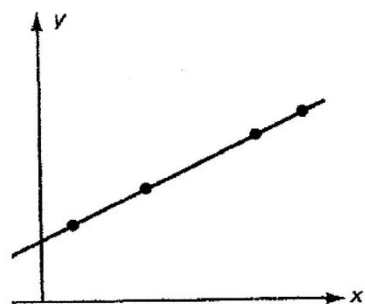
to
$$\rho = \begin{cases} 1 & \text{gdy } b > 0 \\ -1 & \text{gdy } b < 0 \end{cases}$$

□ Jeśli $|\rho| = 1$, to między zmiennymi losowymi X , Y

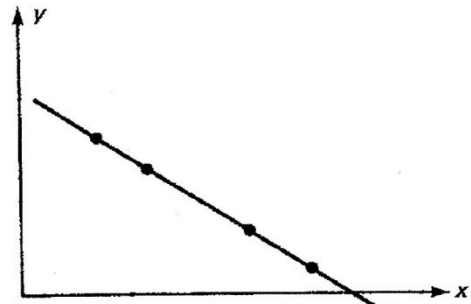
istnieje **liniowa zależność funkcyjna**

□ Jeśli zmienne losowe X i Y są **niezależne**, to

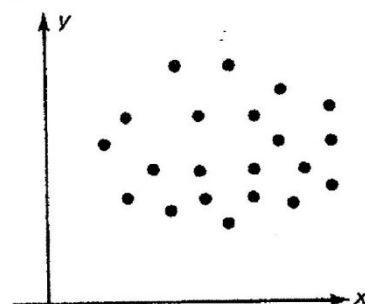
$\rho = 0$



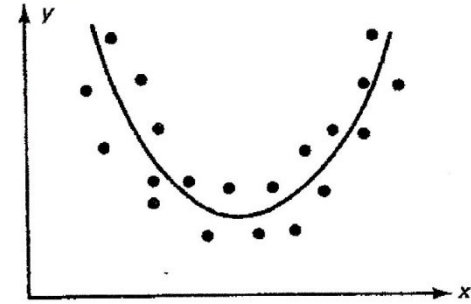
(a) $\rho = 1$, perfect positive correlation.



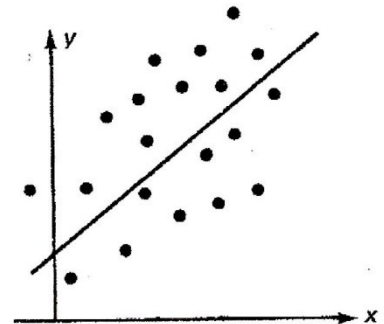
(b) $\rho = -1$, perfect negative correlation.



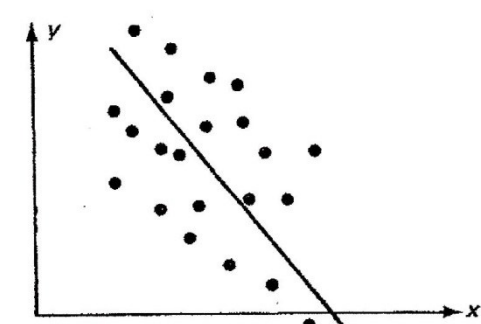
(c) ρ is close to zero and variables are not related.



(d) ρ is close to zero, but the variables are related through a curvilinear relation.



(e) $0 < \rho < 1$, low positive correlation.



(f) $-1 < \rho < 0$, low negative correlation.

Współczynnik korelacji próbkowej

Współczynnikiem korelacji próbkowej nazywamy wartość współczynnika R obliczoną dla próbki $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$:

$$r = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_X} \right) \left(\frac{y_i - \bar{y}}{s_Y} \right)$$

Własności współczynnika korelacji próbkowej

- $-1 \leq r \leq 1$.
- Jeśli $r = 1$, to wszystkie punkty wykresu rozproszenia leżą na prostej o dodatnim współczynniku kierunkowym, tzn. istnieje dodatnia zależność liniowa między zmiennymi x i y próbki.
- Jeśli $r = -1$, to wszystkie punkty wykresu rozproszenia leżą na prostej o ujemnym współczynniku kierunkowym, tzn. istnieje ujemna zależność liniowa między zmiennymi x i y próbki.
- Wartości r bliskie -1 lub 1 wskazują, że wykres rozproszenia jest skupiony wokół pewnej prostej.

Współczynnik korelacji liniowej z próbki

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} =$$

$$= \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{\sqrt{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} \sqrt{n \sum_{i=1}^n y_i^2 - \left(\sum_{i=1}^n y_i\right)^2}}$$

Do obliczenia r potrzebujemy:

$$\sum_{i=1}^n x_i, \sum_{i=1}^n y_i, \sum_{i=1}^n x_i^2, \sum_{i=1}^n y_i^2, \sum_{i=1}^n x_i y_i, n$$

Regresja prostoliniowa

Prosta regresji. Metoda najmniejszych kwadratów

Problem: W jaki sposób do wykresu rozproszenia, tzn. do punktów $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ dopasować „najlepiej” linię prostą?

Niech $y = b_0 + b_1x$, $-\infty < x < \infty$, będzie równaniem prostej „dopasowanej” do punktów (x_i, y_i) , $i = 1, \dots, n$, wykresu rozproszenia. (b_1 - **współczynnik kierunkowy**, b_0 - **wyraz wolny**).

Wówczas $\hat{y}_i = b_0 + b_1x_i$ będzie przybliżeniem wartości y_i na podstawie zmiennej niezależnej x_i uzyskanym z zależności liniowej.

Błąd przybliżenia, czyli różnicę $y_i - \hat{y}_i$ nazywamy **wartością resztową**, lub **rezyduum**.

Regresja prostoliniowa

Miarą dopasowania prostej do próbki (punktów wykresu rozproszenia) jest suma kwadratów błędów (rezyduów):

$$S(b_0, b_1) = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - (b_0 + b_1 x_i))^2$$

Prostą dla której $S(b_0, b_1)$ osiąga wartość minimalną nazywamy **prostą regresji** lub też prostą wyznaczoną **metodą najmniejszych kwadratów**.

Współczynniki prostej regresji b_0, b_1 wyznaczamy z warunku koniecznego minimum funkcji $S(b_0, b_1)$, tzn. przyrównując do zera obie pochodne cząstkowe.

Regresja prostoliniowa

Rozwiązując układ dwóch równań liniowych otrzymujemy:

$$b_1 = \frac{\sum_{i=1}^n x_i (y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (1)$$

$$b_0 = \frac{1}{n} \left(\sum_{i=1}^n y_i - b_1 \sum_{i=1}^n x_i \right) = \bar{y} - b_1 \bar{x} \quad (2)$$

gdzie $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$, $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$.

Wartość $\hat{y} = b_0 + b_1 x$ nazywamy **wartością przewidywaną** zmiennej objaśnianej (zależnej), przy pomocy prostej regresji, na podstawie zmiennej objaśniającej (niezależnej) x .

Wzory obliczeniowe

Współczynnik regresji liniowej b_1 :

$$b_1 = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2}$$

Wyraz wolny b_0 :

$$b_0 = \bar{y} - b_1 \bar{x}$$

Prosta regresji y względem x :

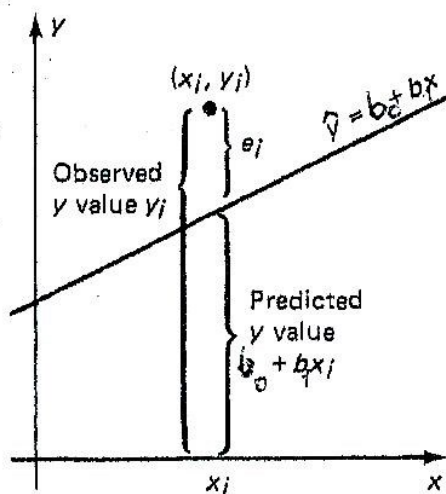
$$\hat{y} = b_0 + b_1 x$$

Błędy prognozy – rezydua

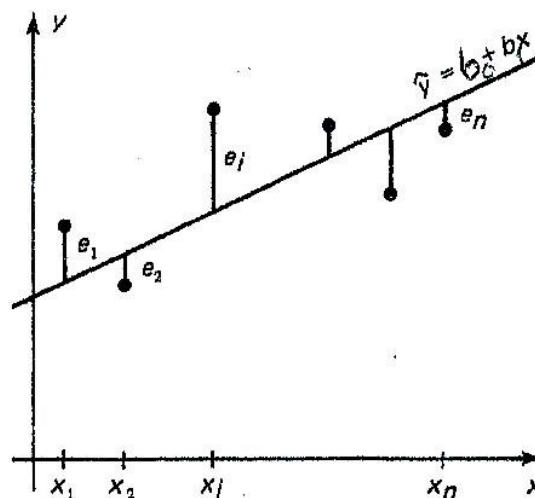
y_j = obserwowana wartość cechy y

$\hat{y}_j = b_0 + b_1 x_j$ = przewidywana wartość y na podstawie x_j

$e_j = y_j - \hat{y}_j$ = błąd prognozy = rezyduum



(a)



(b)

Deviations of the observed y values from the predicted y values $\hat{y} = b_0 + b_1 x$.

Wyznaczenie liniowej funkcji regresji y względem x

$$\sum_{j=1}^n \hat{y}_j = n(\bar{y} - b_1 \bar{x}) + \sum_{j=1}^n b_1 x_j = \sum_{j=1}^n y_j - b_1 \sum_{j=1}^n x_j + \sum_{j=1}^n b_1 x_j$$

$$\sum_{j=1}^n \hat{y}_j = \sum_{j=1}^n y_j \quad . \quad \text{Stąd } \bar{\hat{y}} = \bar{y} \quad (*)$$

$$\sum_{j=1}^n \hat{y}_j = \sum_{j=1}^n b_0 + b_1 x_j = nb_0 + b_1 \sum_{j=1}^n x_j$$

Stąd $\bar{\hat{y}} = b_0 + b_1 \bar{x}$. Ponadto z $(*)$ $\bar{\hat{y}} = \bar{y}$. Zatem

$$\bar{y} = b_0 + b_1 \bar{x}$$

Punkt (\bar{x}, \bar{y}) leży na wykresie liniowej funkcji regresji



Liniowa funkcja regresji x względem y

Prosta regresji x względem y :

$$\hat{x} = a_0 + a_1 y,$$

gdzie

$$a_1 = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n y_i^2 - \left(\sum_{i=1}^n y_i \right)^2}$$

$$a_0 = \bar{x} - a_1 \bar{y}$$

Liniowa funkcja regresji y względem x

Przykład. W tabeli podano wzrost i wagę grupy osób objętych badaniem. (a) Wyznaczyć liniową funkcję regresji wagi względem wzrostu. (b) Określić dobroć dopasowania modelu (c) Podać interpretację współczynnika regresji liniowej.

wzrost	175	168	173	185	180	188	180
waga	74	69	84	84	71	100	86

(a)

$$\sum_{i=1}^7 x_i = 1249, \sum_{i=1}^7 y_i = 568, \sum_{i=1}^7 x_i^2 = 223147, \sum_{i=1}^7 y_i^2 = 46786,$$

$$\sum_{i=1}^7 x_i y_i = 101674$$

Wyznaczenie liniowej funkcji regresji y względem x

$$b_1 = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2} = \frac{7 \cdot 101674 - 1249 \cdot 568}{7 \cdot 223147 - 1249^2} = 1,127$$

$$b_0 = \bar{y} - b_1 \bar{x} = \frac{568}{7} - 1,127 \cdot \frac{1249}{7} = -119,9$$

$$\hat{y} = b_0 + b_1 x = -119,9 + 1,127x$$

x_j	175	168	173	185	180	188	180
y_j	74	69	84	84	71	100	86
\hat{y}_j	77	69	75	88	83	92	83
e_j	-3	0	9	-4	-12	8	3

Współczynnik korelacji próbkowej

$$r = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{\sqrt{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} \sqrt{n \sum_{i=1}^n y_i^2 - \left(\sum_{i=1}^n y_i\right)^2}}$$

$$r = \frac{7 \cdot 101674 - 1249 \cdot 568}{\sqrt{7 \cdot 223147 - 1249^2} \sqrt{7 \cdot 46786 - 568^2}} = 0,73$$

Próbkowy współczynnik korelacji $r = 0,73$

Współczynnik determinacji $= R^2 = r^2 = 0,53$

Regresja prostoliniowa

Ocena „dobroci” dopasowania prostej regresji

Wprowadzamy oznaczenia:

$$SST = \sum_{i=1}^n (y_i - \bar{y})^2$$

- **całkowita suma kwadratów** (*Total Sum of Squares*) (miara zmienności samych y_1, \dots, y_n).

$$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

- **suma kwadratów błędów** (*Error Sum of Squares*).

$$SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

- **regresyjna (modelowa) suma kwadratów** (*Regression/Model Sum of Squares*) (miara zmienności $\hat{y}_1, \dots, \hat{y}_n$).

Regresja prostoliniowa

Można pokazać, że zachodzi równość :

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2.$$

$$SST = SSE + SSR$$

Współczynnik determinacji określony wzorem

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

jest miarą stopnia dopasowania prostej regresji do wykresu rozproszenia.

Określa stopień, w jakim zależność liniowa, między zmienną objaśnianą a objaśniającą, wyjaśnia zmienność wykresu rozproszenia.

Im mniejsze SSE , tym wykres rozproszenia jest bardziej skupiony wokół prostej regresji.

Regresja prostoliniowa

Wartość współczynnika determinacji jest ściśle związana z wartością współczynnika korelacji próbkowej.

Stwierdzenie.

Zachodzi równość

$$r^2 = \frac{SSR}{SST} = R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

zmienność wyjaśniona
przez model/ zmienność całkowita

Regresja prostoliniowa

Przykład. Zanotowano miesięczne wydatki na reklamę (w 10 000 PLN) pewnego artykułu oraz miesięczne dochody ze sprzedaży artykułu (w 100 000 PLN)

Miesiąc i	1	2	3	4	5
Reklama x_i	5	6	7	8	9
Dochód y_i	4,5	6,5	8,4	7,6	8,4

Wyznaczyć liniową funkcję regresji oraz przewidywaną wartość dochodu przy wydatkach na reklamę 100 000 PLN (10 x 10 000).

Kolejno obliczamy:

$$\bar{x} = 7,0, \quad \bar{y} = 7,08, \quad s_X = 1,58, \quad s_Y = 1,64$$

$$r = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_X} \right) \left(\frac{y_i - \bar{y}}{s_Y} \right) = 0,858$$

(współczynnik korelacji próbkowej)

Regresja prostoliniowa

Współczynniki prostej regresji

$$b_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \underline{0,89}$$

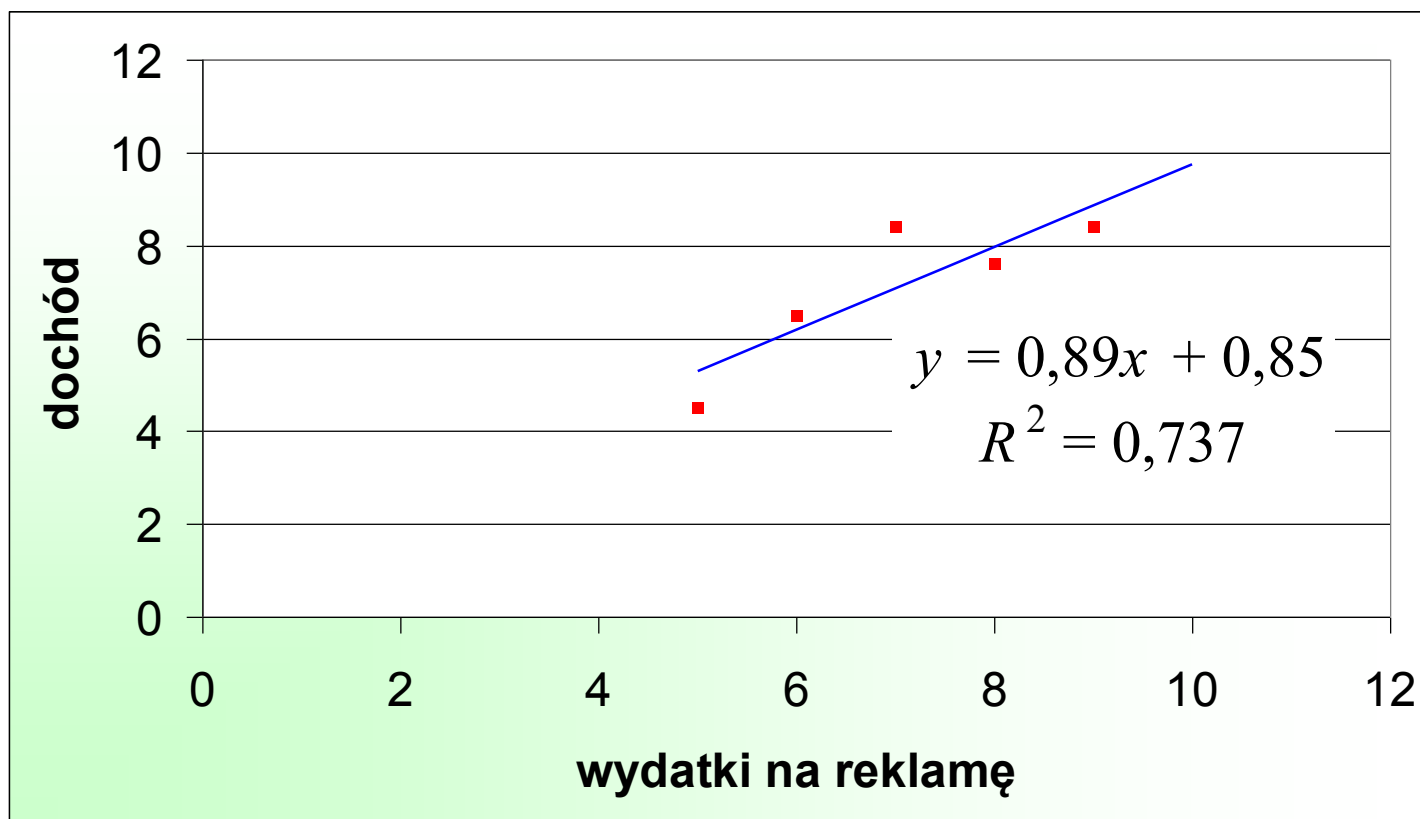
$$b_0 = \bar{y} - b_1 \bar{x} = 7,08 - 0,89 \times \underline{7} = \underline{0,85}$$

Przewidywany dochód ze sprzedaży, przy wydatku na reklamę $x = 10$ (x 10 000 PLN) wynosi:

$$\hat{y} = b_0 + b_1 x = 0,85 + 0,89 \cdot 10 = 9,75 (\times 100000 \text{ PLN})$$

Regresja prostoliniowa

Wykres rozproszenia oraz empiryczna prosta regresji



Regresja prostoliniowa

$$SST = \sum_{i=1}^n (y_i - \bar{y})^2 = 10,748$$

$$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = 2,827$$

$$SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = 7,921$$

współczynnik determinacji

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

$$R^2 = 0,737$$

Zmienność dochodu jest w prawie 74% wyjaśniona przez zmienność wydatków na reklamę (zmienność wydatków na reklamę w 74% określa zmienność dochodu).

Regresja prostoliniowa

Model zależności liniowej (model regresji liniowej)

Założmy, że próbka $(x_1, y_1), \dots, (x_n, y_n)$ jest realizacją próby losowej $(x_1, Y_1), \dots, (x_n, Y_n)$, gdzie

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

oraz $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ są niezależnymi zmiennymi losowymi o wartości oczekiwanej 0 i wariancji σ^2 , a znane liczby x_1, \dots, x_n nie wszystkie są jednakowe.

Prostą $y = \beta_0 + \beta_1 x$ nazywamy **prostą regresji**.

Współczynnik β_0 - **wyraz wolny** prostej regresji.

Współczynnik β_1 - **współczynnik kierunkowy** prostej regresji.

Zmienne losowe $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ - **losowe błędy** w modelu

$$\text{Var}(\varepsilon_i) = \sigma^2.$$

Regresja prostoliniowa

Własności zmiennej losowej Y_i , $i = 1, \dots, n$,

$$E(Y_i) = E(\beta_0 + \beta_1 x_i) + E(\varepsilon_i) = \beta_0 + \beta_1 x_i .$$

$$\text{Var}(Y_i) = \text{Var}(\beta_0 + \beta_1 x_i + \varepsilon_i) = \text{Var}(\varepsilon_i) = \sigma^2$$

Założenia:

- x_1, \dots, x_n są znane.
- Obserwujemy wartości zmiennych Y_1, \dots, Y_n .
- $\beta_0, \beta_1, \sigma^2$ są nieznanymi parametrami modelu.

Cel eksperymentu – wnioskowanie na temat parametrów modelu

Regresja prostoliniowa

Naturalne **estymatory** parametrów β_0, β_1 otrzymujemy metodą najmniejszych kwadratów, wstawiając we wzorach (1), (2) zmienne losowe Y_i zamiast ich wartości y_i , $i = 1, \dots, n$,

$$\hat{b}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

$$\hat{b}_0 = \bar{Y} - \hat{b}_1 \bar{x}.$$

Regresja prostoliniowa

Własności estymatorów \hat{b}_0 , \hat{b}_1 podane są w następującym twierdzeniu.

Twierdzenie

$$(i) \quad E(\hat{b}_0) = \beta_0, \quad E(\hat{b}_1) = \beta_1$$

$$(ii) \quad \text{Var}(\hat{b}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) \quad (3)$$

$$(iii) \quad \text{Var}(\hat{b}_1) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (4)$$

(iv) Jeśli $\varepsilon_i \sim N(0, \sigma)$, $i = 1, \dots, n$, to \hat{b}_0 , \hat{b}_1 mają rozkłady normalne o wartościach średnich i wariancjach określonych w (i) - (iii).

Regresja prostoliniowa

Estymator σ^2

Definicja

Błędem średniokwadratowym S^2 nazywamy estymator wariancji σ^2 określony następująco

$$S^2 = \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n - 2} = \frac{SSE}{n - 2}$$

Liczbę $n - 2$ nazywamy liczbą stopni swobody rezyduów.

Stwierdzenie

S^2 jest nieobciążonym estymatorem σ^2 , tzn.

$$E(S^2) = \sigma^2.$$

$$S = \sqrt{S^2} = \text{estymator } \sigma.$$

Regresja prostoliniowa

Wniosek (i) Nieobciążonym estymatorem wariancji $\text{Var}(\hat{b}_0)$ jest

$$[SE(\hat{b}_0)]^2 = S^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right),$$

Stąd naturalnym estymatorem odchylenia standardowego $\sigma_{\hat{b}_0}$ jest

$$SE(\hat{b}_0) = S \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

nazywany **błędem standardowym estymatora** \hat{b}_0 , gdyż na mocy (3)

$$SE(\hat{b}_0) = \text{estymator } \sigma_{\hat{b}_0} = (\sqrt{\text{Var}(\hat{b}_0)})$$

Regresja prostoliniowa

(ii) Nieobciążonym estymatorem $\text{Var}(\hat{b}_1)$ jest

$$[SE(\hat{b}_1)]^2 = \frac{S^2}{\sum_{i=1}^n (x_i - \bar{x})^2},$$

Stąd naturalnym estymatorem odchylenia standardowego $\sigma_{\hat{b}_1}$ jest

$$SE(\hat{b}_1) = \frac{S}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

nazywamy **błędem standardowym** estymatora \hat{b}_1 , gdyż na mocy (4)

$$SE(\hat{b}_1) = \text{estymator } \sigma_{\hat{b}_1} = \sqrt{\text{Var}(\hat{b}_1)}.$$

Regresja prostoliniowa

Twierdzenie Jeśli $\varepsilon_i \sim N(0, \sigma)$, $i = 1, \dots, n$, to:

(i)

$$\hat{b}_1 \sim N\left(\beta_1, \frac{\sigma}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}\right),$$

(tzn. ma rozkład normalny ze wskazanymi parametrami),

$$\frac{\hat{b}_1 - \beta_1}{SE(\hat{b}_1)} \sim t_{n-2}$$

gdzie

$$SE(\hat{b}_1) = \frac{S}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}.$$

(tzn. ma rozkład Studenta o $n-2$ stopniach swobody).

Regresja prostoliniowa

(ii)
$$\hat{b}_0 \sim N\left(\beta_0, \sigma \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}\right),$$

(tzn. ma rozkład normalny ze wskazanymi parametrami),

$$\frac{\hat{b}_0 - \beta_0}{SE(\hat{b}_0)} \sim t_{n-2}$$

gdzie
$$SE(\hat{b}_0) = S \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

(tzn. ma rozkład Studenta o $n-2$ stopniach swobody).

Regresja prostoliniowa

Do końca wykładu zakładamy, że szумы w modelu regresji liniowej mają rozkład normalny (założenie ostatniego twierdzenia).

Przedział ufności na poziomie ufności $1 - \alpha$ dla **współczynnika** β_1 :

$$[\hat{b}_1 - t_{1-\alpha/2, n-2} \times SE(\hat{b}_1), \quad \hat{b}_1 + t_{1-\alpha/2, n-2} \times SE(\hat{b}_1)]$$

Przedział ufności na poziomie ufności $1 - \alpha$ dla **współczynnika** β_0 :

$$[\hat{b}_0 - t_{1-\alpha/2, n-2} \times SE(\hat{b}_0), \quad \hat{b}_0 + t_{1-\alpha/2, n-2} \times SE(\hat{b}_0)]$$

Regresja prostoliniowa

Testowanie hipotezy o wartości współczynnika β_0

(A) $H_0 : \beta_0 = \beta_{0,0},$

gdzie $\beta_{0,0}$ jest ustaloną liczbą.

Statystyka testowa

$$T = \frac{\hat{b}_0 - \beta_{0,0}}{SE(\hat{b}_0)} = (\hat{b}_0 - \beta_{0,0}) / (S \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}})$$

Jeśli H_0 **prawdziwa**, to $T \sim t_{n-2}$.

(ma rozkład Studenta o $n-2$ stopniach swobody).

Zbiory krytyczne dla różnych postaci hipotez alternatywnych

(a) $H_1 : \beta_0 \neq \beta_{0,0}$.

Zbiór krytyczny

$$C = \{t : |t| \geq t_{1-\alpha/2, n-2}\}.$$

(b) $H_1 : \beta_0 > \beta_{0,0}$.

Zbiór krytyczny

$$C = \{t : t \geq t_{1-\alpha, n-2}\}.$$

(c) $H_1 : \beta_0 < \beta_{0,0}$.

Zbiór krytyczny

$$C = \{t : t \leq -t_{1-\alpha, n-2}\}.$$

Regresja prostoliniowa

Testowanie hipotezy o wartości współczynnika β_1

(B) $H_0 : \beta_1 = \beta_{1,0}$,

gdzie $\beta_{1,0}$ jest ustaloną liczbą.

Statystyka testowa

$$T = \frac{\hat{b}_1 - \beta_{1,0}}{SE(\hat{b}_1)} = \frac{(\hat{b}_1 - \beta_{1,0}) \sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}{S}$$

Jeśli H_0 prawdziwa, to

$$T \sim t_{n-2}$$

(ma rozkład Studenta o $n-2$ stopniach swobody).

Regresja prostoliniowa

Zbiory krytyczne dla różnych postaci hipotez alternatywnych

(a) $H_1 : \beta_1 \neq \beta_{1,0}$.

Zbiór krytyczny $C = \{t : |t| \geq t_{1-\alpha/2, n-2}\}$.

(b) $H_1 : \beta_1 > \beta_{1,0}$.

Zbiór krytyczny $C = \{t : t \geq t_{1-\alpha, n-2}\}$.

(c) $H_1 : \beta_1 < \beta_{1,0}$.

Zbiór krytyczny $C =$

$$\{t : t \leq -t_{1-\alpha, n-2}\}.$$

Test istotności regresji liniowej

c) $H_0 : \beta_1 = 0, \quad H_1 : \beta_1 \neq 0,$

Statystyka testowa

$$F = \frac{SSR / 1}{SSE / (n - 2)}$$

Jeśli H_0 prawdziwa, to F ma **rozkład F Snedecora** o $(1, n-2)$ stopniach swobody.

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2.$$

$$\begin{array}{rclcl} SST & = & SSE & + & SSR \\ n-1 & = & n-2 & + & 1 \end{array}$$

(Liczby stopni swobody SSx = liczba niezależnych zmiennych zmniejszona o liczbę ograniczeń występujących w określeniu SSx).

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

Regresja prostoliniowa

Zbiór krytyczny testu

$$C = \{F_{obl} : F_{obl} \geq f_{1-\alpha,1,n-2}\}.$$

Zauważmy, że

$$F = T^2,$$

stąd test jest szczególnym przypadkiem testu z **(B)** gdy $\beta_{1,0} = 0$.