Statystyczna Analiza Danych
SAD-2020/2021

Wykład 12 i 13

Współczynnik korelacji próbkowej

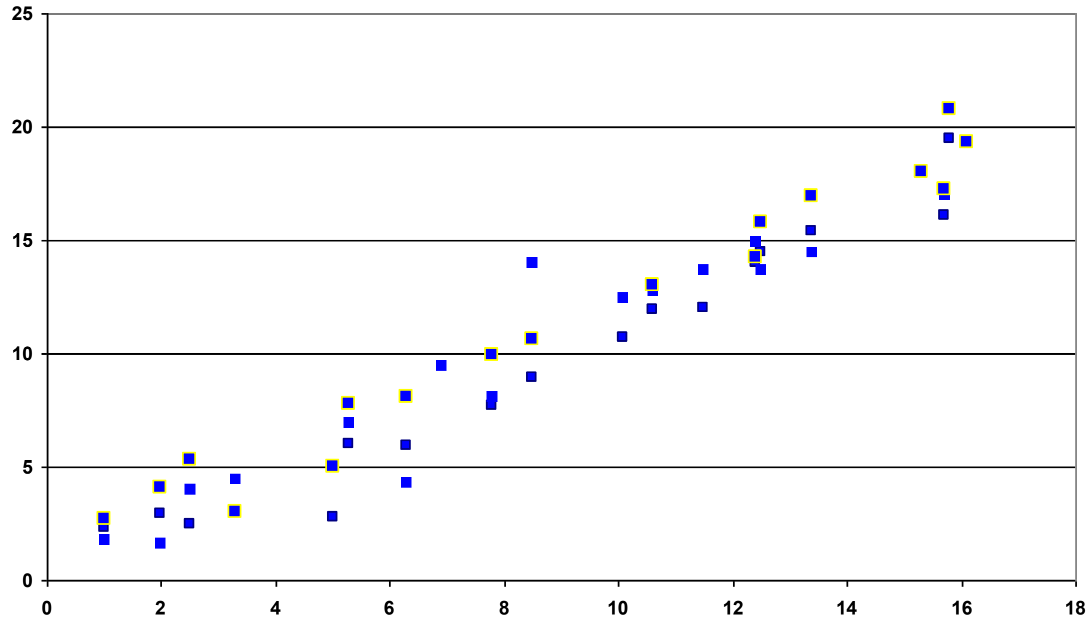
Niech $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ będzie próbką cechy dwuwymiarowej (X, Y).

Będziemy badali zależność Y od X.

X = zmienna **niezależna** (o**bjaśniająca**),

Y = zmienna **zależna** (**objaśniana**),

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



Współczynnik korelacji z próby

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

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

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

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

Współczynnik korelacji próbkowej

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

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

Własności współczynnika korelacji próbkowej

- $\boxed{-1 \le r \le 1}$.
- 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.

Regresja prostoliniowa

Prosta regresji. Metoda najmniejszych kwadratów

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

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

Wówczas $\hat{y}_i = b_0 + b_1 x_i$ 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):

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

Prostą dla której $S(b_0, b_1)$ osiąga wartość <u>minimalną</u> 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:

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

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

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

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

Regresja prostoliniowa

Ocena „dobroci” dopasowania prostej regresji

Wprowadzamy oznaczenia:

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

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

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

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

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

- **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 :

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

$$SST \quad = \quad SSE \quad + \quad SSR$$

Współczynnik determinacji określony wzorem

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

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

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

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:

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

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

(współczynnik korelacji próbkowej)

Regresja prostoliniowa

Współczynniki prostej regresji

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

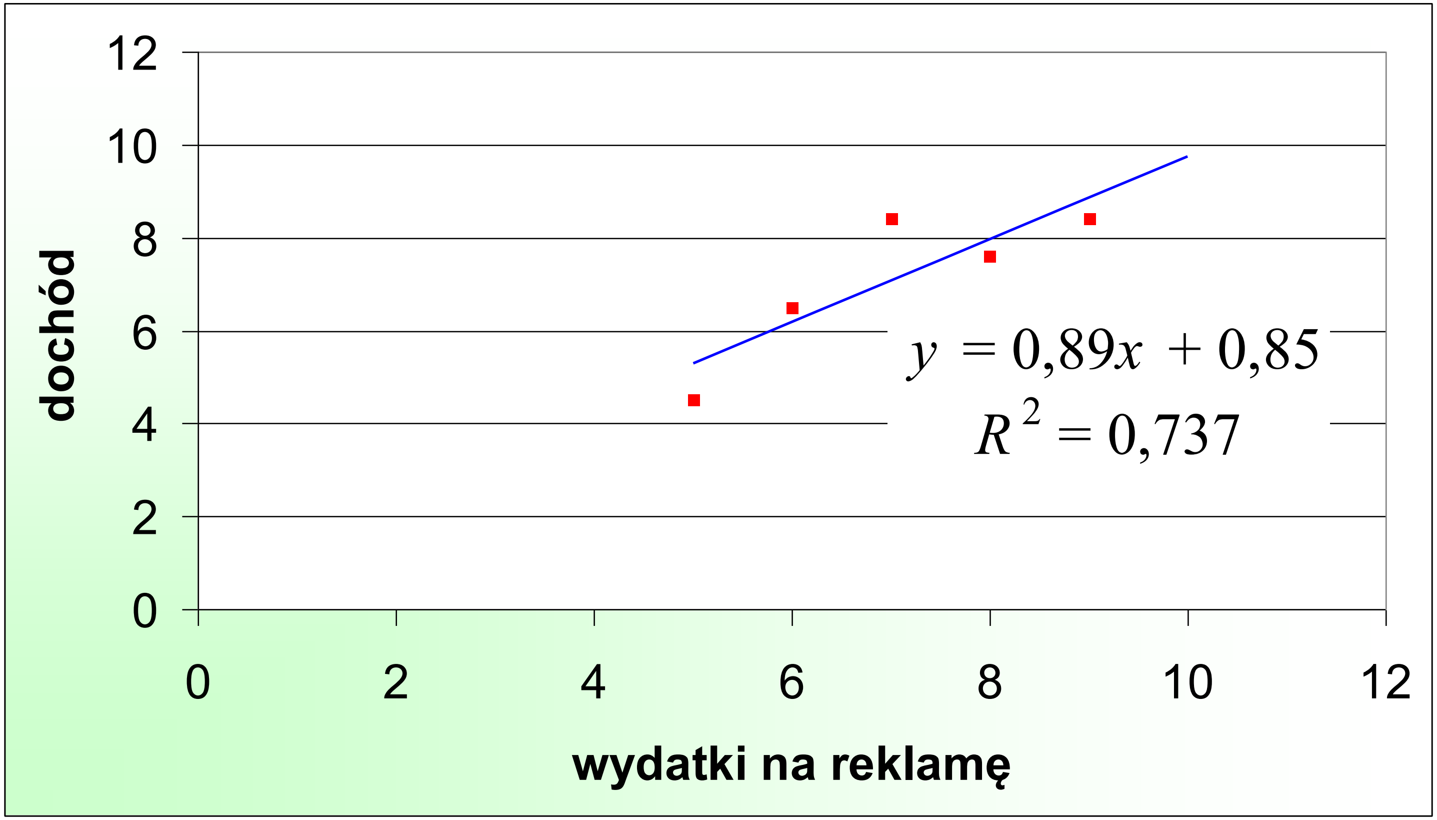
$$b_0 = \bar{y} - b_1\bar{x} = 7{,}08 - 0{,}89 \times 7 = \underline{0{,}85}$$

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

$$\hat{y} = b_0 + b_1 x = 0{,}85 + 0{,}89 \times 10 = 9{,}75\ (\times 100\ 000\ \text{PLN})$$

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 ma reklamę (zmienność wydatków na reklamę w 74% określa zmienność dochodu).

Regresja prostoliniowa

Model zależności liniowej (**model regresji liniowej**)

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

$$\boxed{Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i}, \quad i = 1,\ldots,n,$$

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

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

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

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

Zmienne losowe $\varepsilon_1, \varepsilon_2,\ldots,\varepsilon_n$ - **losowe błędy** w modelu $\mathrm{Var}(\varepsilon_i) = \sigma^2$.

Regresja prostoliniowa

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

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

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

Założenia:

- $x_1,\ldots,x_n$ są znane.
- Obserwujemy wartości zmiennych $Y_1,\ldots,Y_n$.
- $\beta_0, \beta_1, \sigma^2$ 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, $i = 1,\dots,n$,

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

$$\widehat{b}_0 = \overline{Y} - \widehat{b}_1\overline{x}.$$

Regresja prostoliniowa

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

Twierdzenie

(i) $E(\widehat{b}_0) = \beta_0, \qquad E(\widehat{b}_1) = \beta_1$

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

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

(iv) Jeśli $\varepsilon_i \sim N(0, \sigma)$, $i = 1,..,n$, to $\widehat{b}_0$, $\widehat{b}_1$ 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

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

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

Stwierdzenie

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

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

$S = \sqrt{S^2}$ = estymator σ.

Regresja prostoliniowa

Wniosek (i) Nieobciążonym estymatorem wariancji $\mathrm{Var}(\widehat{b}_0)$ jest

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

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

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

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

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

Regresja prostoliniowa

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

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

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

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

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

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

Regresja prostoliniowa

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

(i)

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

(tzn. ma rozkład normalny ze wskazanymi parametrami),

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

gdzie

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

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

Regresja prostoliniowa

(ii)

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

(tzn. ma rozkład normalny ze wskazanymi parametrami),

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

gdzie

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

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

Regresja prostoliniowa

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

$$\boxed{[\widehat{b_1} - t_{1-\alpha/2,n-2} \times SE(\widehat{b_1}), \quad \widehat{b_1} + t_{1-\alpha/2,n-2} \times SE(\widehat{b_1})]}$$

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

$$\boxed{[\widehat{b_0} - t_{1-\alpha/2,n-2} \times SE(\widehat{b_0}), \quad \widehat{b_0} + t_{1-\alpha/2,n-2} \times SE(\widehat{b_0})]}$$

Regresja prostoliniowa

Testowanie hipotezy o wartości współczynnika β_0

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

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

Statystyka testowa

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

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

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

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

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

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

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

Regresja prostoliniowa

Testowanie hipotezy o wartości współczynnika β_1

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

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

Statystyka testowa

$$\boxed{T = \frac{\widehat{b_1} - \beta_{1,0}}{SE(\widehat{b_1})} = \frac{(\widehat{b_1} - \beta_{1,0})\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}}{S}}$$

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

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

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

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

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

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

Regresja prostoliniowa

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

Statystyka testowa

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

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

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

$$SST \quad = \quad SSE \quad + \quad SSR$$

$$n\text{-}1 \quad = \quad n\text{-}2 \quad + \quad 1$$

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

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

Regresja prostoliniowa

Zbiór krytyczny testu

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

Zauważmy, że

$$\boxed{F = T^2},$$

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

Prognoza wartości Y na podstawie x_0.

Obserwowane $Y_1,\ldots,Y_n$.

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

Nieobserwowane $Y(x_0) = \beta_0 + \beta_1 x_0 + \varepsilon_0,$ (5)

gdzie $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n, \varepsilon_0$ są niezależnymi zmiennymi losowymi o rozkładach $N(0,\sigma)$.

Zadania:

(a) ocena (estymacja) **wartości średniej zmiennej objaśnianej $Y(x_0)$, tzn.**

$$\mu_{Y(x_0)} = E[Y(x_0)]$$

w sytuacji, gdy zmienna objaśniająca x jest równa x_0.

(b) przewidywanie (prognoza) wartości $Y(x_0)$.

(a) Obliczając wartość średnią obu stron (5) mamy:

$$\mu_{Y(x_0)} = E(\beta_0 + \beta_1 x_0) + E(\varepsilon_0) = \beta_0 + \beta_1 x_0.$$

Stąd naturalnym oszacowaniem $\mu_{Y(x_0)}$ jest

$$\widehat{\mu}_{Y(x_0)} := \widehat{Y}(x_0) = \widehat{b}_0 + \widehat{b}_1 x_0.$$

$$E[\widehat{Y}(x_0)] = E(\widehat{b}_0 + \widehat{b}_1 x_0) = \beta_0 + \beta_1 x_0 = \mu_{Y(x_0)} \qquad \textbf{(6)}$$

Zatem $\widehat{Y}(x_0)$ jest **nieobciążonym** estymatorem $\mu_{Y(x_0)}$.

$$\sigma^2_{\widehat{Y}(x_0)} = \mathrm{Var}(\widehat{b}_0 + \widehat{b}_1 x_0) = \mathrm{Var}(\overline{Y} + \widehat{b}_1(x_0 - \overline{x})).$$

Można pokazać, że $\widehat{b}_1, \overline{Y}$ są nieskorelowane, stąd

$$\sigma^2_{\widehat{Y}(x_0)} = \sigma^2_{\overline{Y}} + (x_0 - \overline{x})^2 \sigma_{\widehat{b}_1}{}^2 = \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \right) \qquad \textbf{(7)}$$

Błąd standardowy estymatora $\widehat{Y}(x_0)$ definiujemy jako

$$SE_{\widehat{Y}(x_0)} = S \sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2}}.$$

Twierdzenie. Estymator $\widehat{Y}(x_0)$ wartości średniej $\mu_{Y(x_0)}$ zmiennej objaśnianej *Y* dla wartości zmiennej objaśniającej x_0 ma rozkład normalny o wartości średniej i wariancji postaci (6) i (7), odpowiednio. Ponadto,

$$\frac{\widehat{Y}(x_0)-\mu_{Y(x_0)}}{SE_{\widehat{Y}(x_0)}} \sim t_{n-2}.$$

Wniosek. **Przedział ufności na poziomie ufności** $1-\alpha$ dla $\mu_{Y(x_0)} = \beta_0 + \beta_1 x_0$ ma krańce

$$\widehat{Y}(x_0) \mp t_{1-\alpha/2,n-2} SE_{\widehat{Y}(x_0)}.$$

Długość przedziału nie jest stała, (wynosi $2t_{1-\alpha/2,n-2} SE_{\widehat{Y}(x_0)}$) , zależy od x_0,

im **dalej od** $\bar{x}$ tym bardziej ocena staje się **niedokładna.**

(b) Prognoza (przewidywanie) $Y(x_0)$.

Niech $\widehat{Y}(x_0)$ będzie oceną (prognozą) $Y(x_0)$. Zmienne

losowe $\widehat{Y}(x_0)$, $Y(x_0)$ są niezależne, więc wariancja ich różnicy ma postać:

$$\sigma^2_{\widehat{Y}(x_0)-Y(x_0)} = \sigma^2_{\widehat{Y}(x_0)} + \sigma^2_{Y(x_0)} = \sigma^2_{\widehat{Y}(x_0)} + \sigma^2 =$$

$$= \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right).$$

Stąd naturalnym estymatorem standardowego odchylenia $\widehat{Y}(x_0) - Y(x_0)$ jest tzw. błąd standardowy $\widehat{Y}(x_0) - Y(x_0)$ jest

$$SE_{\widehat{Y}(x_0)-Y(x_0)} = S\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}.$$

Twierdzenie. Zmienna losowa $\widehat{Y}(x_0) - Y(x_0)$ ma rozkład normalny

$N(0, \sigma_{\widehat{Y}(x_0)-Y(x_0)})$, oraz

$$\frac{\widehat{Y}(x_0) - Y(x_0)}{SE_{\widehat{Y}(x_0)-Y(x_0)}} \sim t_{n-2}.$$

Wniosek. Przedział ufności na poziomie ufności $1-\alpha$

dla zmiennej $Y(x_0) = \beta_0 + \beta_1 x_0 + \varepsilon_0$ ma krańce

$$\widehat{Y}(x_0) \mp t_{1-\alpha/2, n-2} SE_{\widehat{Y}(x_0)-Y(x_0)}.$$

Przykład. (c.d.) Zanotowano miesięczne wydatki na reklamę (w 10000 złotych) pewnego artykułu oraz miesięczne dochody ze sprzedaży artykułu (w 100000 zł):

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

Prosta regresji dla miesięcznego dochodu ze sprzedaży artykułu w zależności od miesięcznego wydatku na reklamę:

$$y = 0{,}85 + 0{,}89x$$

Stąd prognozowany dochód przy wydatku na reklamę **x_0** = 10 (x 10000 zł.) oraz jednocześnie estymowana (przewidywana) wartość średnia dochodu na podstawie miesięcznych wydatków na reklamę **x_0** = 10 (x 10000 zł.)

$$\widehat{Y}(10) = 0,85 + 0,89 \times 10 = 9,75$$ (x 100000 zł.)

Przedział ufności na poziomie ufności **0,90** dla :

(a) $\mu_{Y(10)}$ to

$$[9.75 - t_{0.95,3} SE_{\widehat{Y}(10)}, 9.75 + t_{0.95,3} SE_{\widehat{Y}(10)}] \; ,$$

gdzie $t_{0,95,3}$ = 2,353, $SE_{\widehat{Y}(10)}$ = $S \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$,

S = $\sqrt{SSE/(5-2)}$ = 0,9423,

$SE_{\widehat{Y}(10)}$ = 0,9423 x (1/5 + (10 - 7)^2/10)$^{1/2}$ = 0,9883

granice 90% przedziału ufności dla $\mu_{Y(10)}$:

9,75 — 2,353 x 0,9883 = 7,354
9,75 + 2,353 x 0,9883 = 12,146.

(b) granice 90% przedziału ufności dla prognozy zmiennej $Y(x_0)$:

$$9{,}75 \mp t_{0,95,3} SE_{\widehat{Y}(10)-Y(10)},$$

gdzie

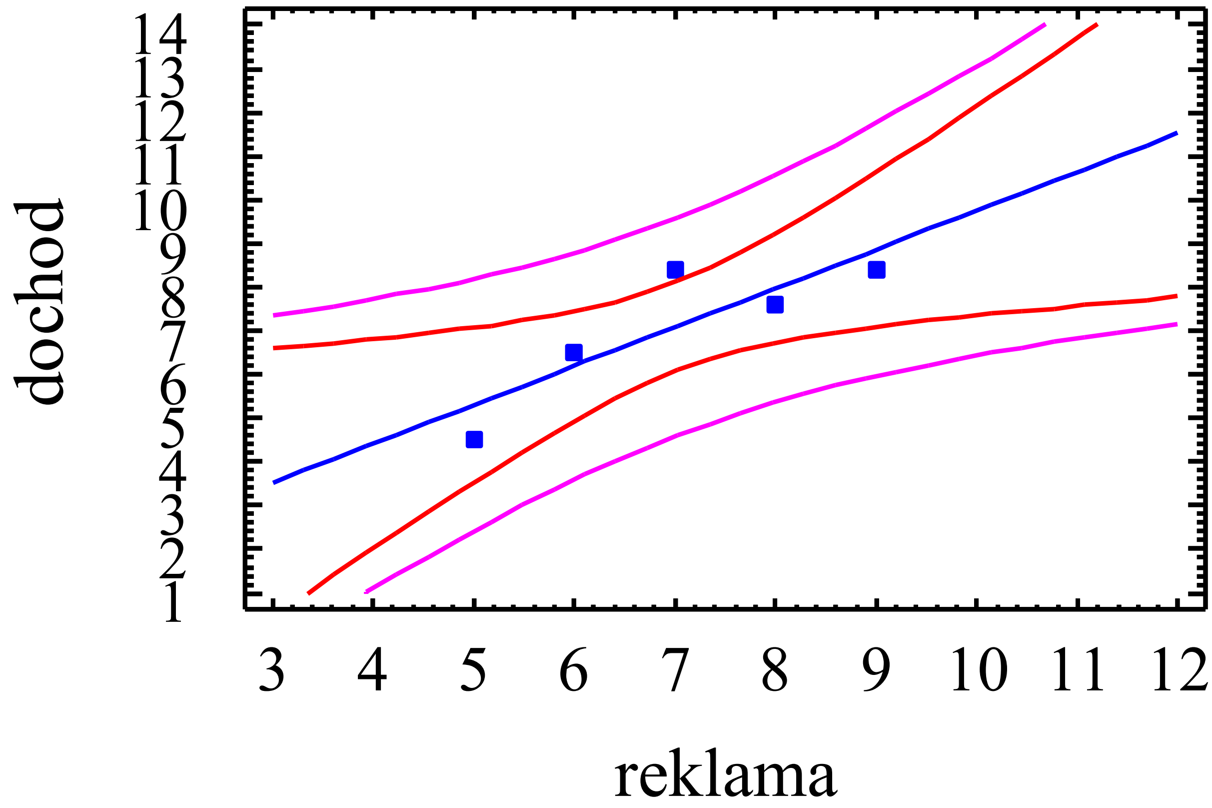
$$SE_{\widehat{Y}(x_0)-Y(x_0)} = S\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$$

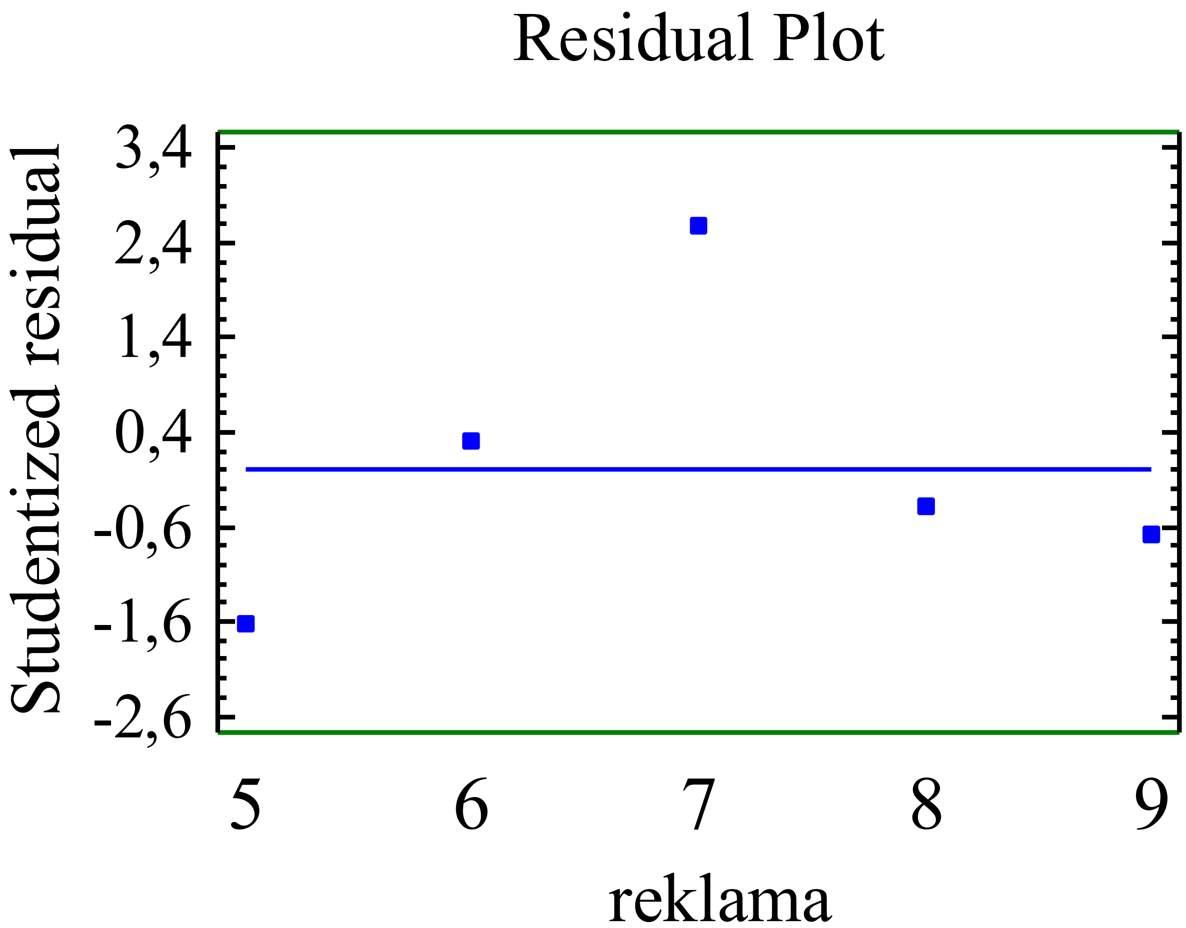
= 0,9423 x (1 +1/5 + (10 - 7)^2/10)$^{1/2}$ = 1,3655.

Predicted Values

X	Predicted Y	90,00% Prediction Limits Lower	90,00% Prediction Limits Upper	90,00% Confidence Limits Lower	90,00% Confidence Limits Upper
4,0	4,41	1,09942	7,72058	2,01398	6,80602
5,0	5,3	2,41029	8,18971	3,53042	7,06958
5,5	5,745	3,0179	8,4721	4,25568	7,23432
6,0	6,19	3,58525	8,79475	4,93872	7,44128
7,0	7,08	4,57744	9,58256	6,05833	8,10167
7,5	7,525	4,9965	10,0535	6,44136	8,60864
8,0	7,97	5,36525	10,5747	6,71872	9,22128
9,0	8,86	5,97029	11,7497	7,09042	10,6296
10,0	9,75	6,43942	13,0606	7,35398	12,146
12,0	11,53	7,13564	15,9244	7,77616	15,2838

Plot of Fitted Model




Residual Plot
Studentized residual
3,4
2,4
1,4
0,4
-0,6
-1,6
-2,6
5
6
7
8
9
reklama

Analiza wartości resztowych (rezyduów)

Poprawność testów dotyczących parametrów modelu oraz prognozy przyszłych zmiennych zależy istotnie od poprawności przyjętego modelu liniowego:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad (8)$$

Wartość resztowa (rezyduum):

$e_i = Y_i - \widehat{Y}_i = Y_i - (\widehat{b}_0 + \widehat{b}_1 x_i)$ jest przybliżeniem błędu

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

Jeśli model (8) jest poprawny, błędy mają rozkład normalny, to rezyduua zachowują się w przybliżeniu tak jak ciąg niezależnych zmiennych losowych o rozkładzie normalnym. W szczególności, wykres rezyduów względem numeru porządkowego powinien przedstawiać „chmurę” punktów skupioną wokół osi Ox, bez wyraźnej struktury czy tendencji.

Stwierdzenie. Wariancja rezyduum ma postać:

$$\sigma_{e_i}^2 = \sigma^2 \left(1 - \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right) \right).$$

Błąd standardowy rezyduum definiujemy:

$$SE_{e_i} = S \sqrt{1 - \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)}.$$

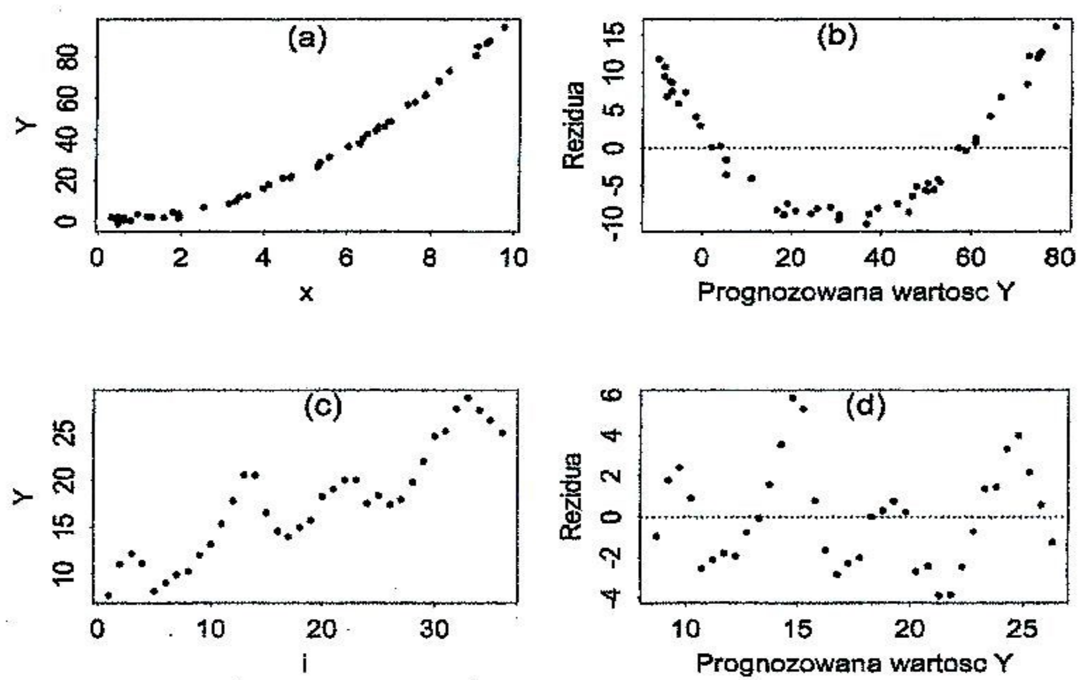
Studentyzowane rezyduum: $r_i = \dfrac{e_i}{SE_{e_i}}, \; i = 1, \ldots, n.$

Przy małej liczbie obserwacji i dużym rozproszeniu zmiennej objaśniającej błędy SE_{e_i} mogą odbiegać znacznie od błędu *S*.

Badanie odstępstw od modelu:

(a) Załóżmy, że model **liniowy jest prawdziwy** (zachodzi związek (8)), ale **rozkład błędów różni się** znacznie od normalnego rozkładu. Wówczas odkryjemy to analizując **histogram** oraz **wykres kwantylowy** rezyduów bądź studentyzowanych rezyduów. W przypadku rozkładu normalnego punkty wykresu kwantylowego będą skupiały się wokół pewnej prostej.

(b) Załóżmy, że model nie jest prawdziwy. Zachodzi związek $Y_i = f(x_i) + \varepsilon_i, i = 1, \ldots, n,$ ale funkcja regresji $f(x)$ nie jest postaci $\beta_0 + \beta_1 x$. Odstępstwo tego typu często udaje się odczytać z wykresu rezyduów. Rys. (a)-(b) sporządzone są dla obserwacji modelu $Y = x^2 + \varepsilon$. Rys. (c)-(d) wykonany dla obserwacji modelu $Y = 10 + 0{,}5i + \varepsilon_i$, gdzie regresja jest liniowa, ale błędy nie są niezależne, kolejne ε_i jest ujemnie zależne od ε_{i-1}.



Rys. 4.6. Charakter zależności zmiennej objaśnianej od objaśniającej

Źródło: J. Koronacki, J. Mielniczuk –*Statystyka dla studentów kierunków technicznych i przyrodniczych*

(c) Prawdziwy model zależności jest **sprowadzalny do modelu liniowego,** np. zależność $Y_i = \beta_0 + \beta_1 x_i^2 + \varepsilon_i$, i = 1, ... , n, sprowadzamy do modelu liniowego wprowadzając nowe zmienne objaśniające: $x_i^{'} = x_i^2$. Jeśli regresja jest liniowa względem współczynników β_0, β_1, to na ogół udaje się znaleźć przekształcenie $f(x)$, które prowadzi do modelu w przybliżeniu liniowego, np. jeśli zależność *y* od *x* jest dodatnia i opisana przez funkcję wklęsłą, to próbujemy zastosować funkcje $f(x) = \sqrt{x}$ lub $f(x) = \log(x)$.

(d) Funkcja regresji jest liniowa (równość (8) spełniona), ale **wariancja błędów nie jest stała**: $\boxed{\mathrm{Var}(\varepsilon_i) = \sigma_i^2.}$ Wówczas modyfikujemy kryterium najmniejszych kwadratów – zamiast minimalizacji sumy kwadratów błędów

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

minimalizujemy ważoną sumę kwadratów błędów:

$$\sum_{i=1}^{n} w_i (y_i - (b_0 + b_1 x_i))^2.$$

Waga w_i powinna być tym mniejsza im większa jest wariancja błędu σ_i^2.

Przyjmujemy: $w_i \approx \sigma_i^{-2}$ lub $w_i \approx \hat{\sigma}_i^{-2}$ (gdy σ nie jest znane).

Często $\hat{\sigma}_i$ = wartość przewidywana dla i-tej obserwacji w modelu regresji z tą samą zmienną objaśniającą, gdy za wartości zmiennej objaśnianej przyjmuje się wartości rezyduów.

(e) Model jest nieadekwatny ze względu na występowanie innych lub większej ilości zmiennych objaśniających

Zadanie. Dopasowano prostą regresji do zmiennej PRODUKCJA (wartość produkcji w 1000 zł) w oparciu o zmienną objaśniającą ENERGIA (wartość zużytej energii w 1000 zł) na podstawie zbioru 115 par obserwacji. Otrzymano następujące wyniki:

$$\text{PRODUKCJA} = 6{,}40 + 2{,}20 \times \text{ENERGIA},$$

wartości błędów standardowych estymatorów współczynników prostej regresji $SE(b_0)=2{,}20$, $SE(b_1)=0{,}11$, $R^2=0{,}86$.

(a) Jaka jest przewidywana wartość produkcji przy wartości zużytej energii 2000 zł?

(b) Podaj procent zmienności wartości produkcji wyjaśnionej przez zaproponowany model zależności liniowej.

(c) Zakładając, że model regresji liniowej jest właściwy, odpowiedz, czy na poziomie istotności 0,01 można stwierdzić, że współczynnik kierunkowy prostej regresji $y = \beta_0 + \beta_1 x$ jest istotny?

Wskazówka. Odpowiednia statystyka testowa T ma rozkład Studenta o 113 stopniach swobody, a więc można zastąpić go rozkładem $N(0,1)$. Sformułuj hipotezy i uzasadnij odpowiedź.

Dopasowana prosta regresji: $y = b_0 + b_1 x$

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = 2{,}20$$

$$b_0 = \overline{y} - b_1\overline{x} = 6{,}40.$$

$$SE(b_0) = S\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}} = 2{,}20, \quad SE(b_1) = \frac{S}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}} = 0{,}11$$

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

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

$$I_1 = [b_1 - t_{1-\alpha/2,n-2} \times SE(b_1),\ \ b_1 + t_{1-\alpha/2,n-2} \times SE(b_1)].$$

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

$$I_0 = [b_0 - t_{1-\alpha/2,n-2} \times SE(b_0),\ \ b_0 + t_{1-\alpha/2,n-2} \times SE(b_0)].$$

Mamy zatem przy spełnionej hipotezie H_0:

$$P(\beta_1 \in I_1) = 1 - \alpha \quad \text{i} \quad P(\beta_0 \in I_0) = 1 - \alpha\ .$$