Prognoza wartości Y na podstawie x_0.

Obserwowane $Y_1,\ldots,Y_n$.

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

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

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

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

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

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

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

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

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

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

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

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

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

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 $\mathbf{x_0}$ = 10 (x 10000 zł.) oraz jednocześnie estymowana (przewidywana) wartość średnia dochodu na podstawie miesięcznych wydatków na reklamę $\mathbf{x_0}$ = 10 (x 10000 zł.)

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

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

(a) $\mu_{Y(10)}$ ma granice $9{,}75 \mp t_{0{,}95,3} SE_{\hat{Y}(10)}$,

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

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

$SE_{\widehat{Y}(10)}$ = 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 $Y(x_0)$:

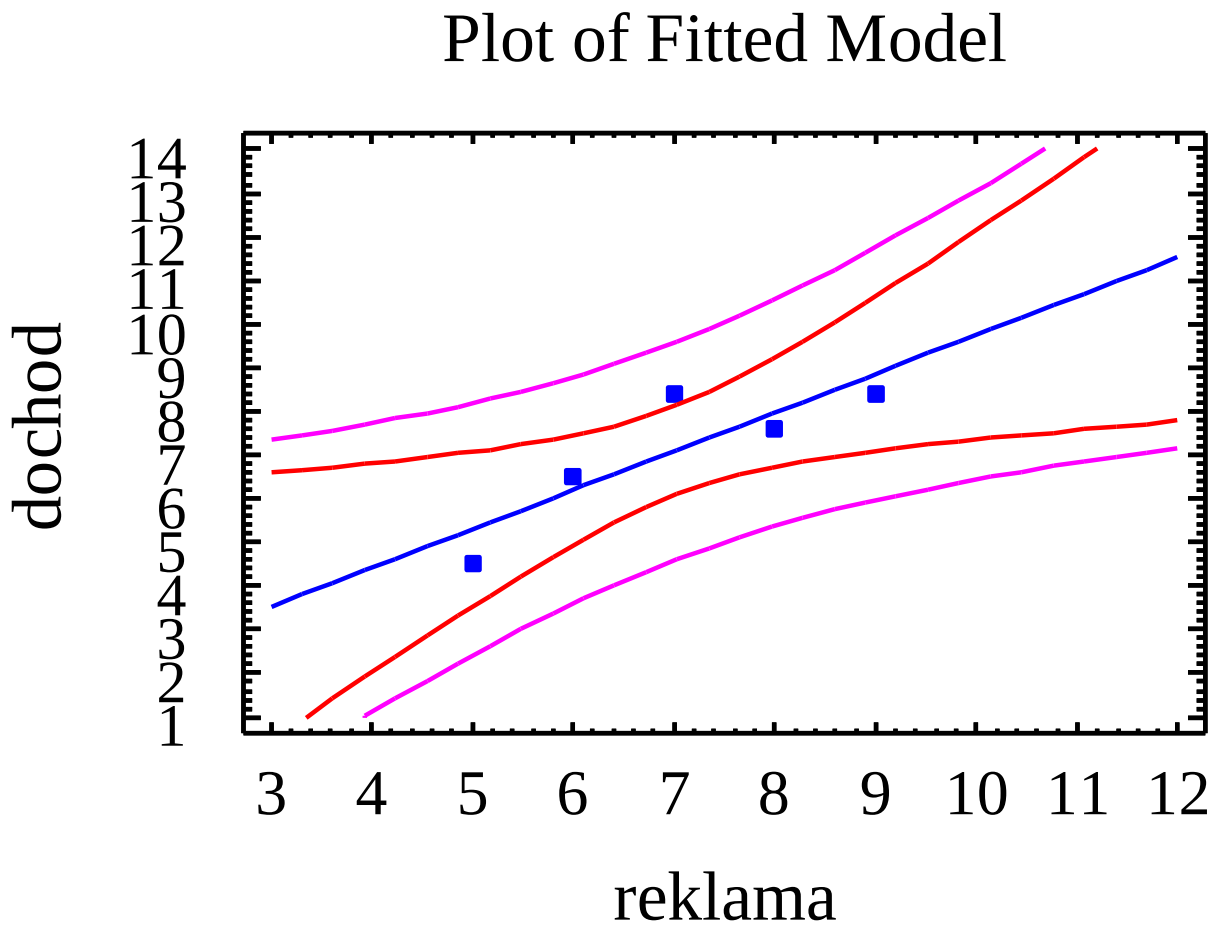
$$9{,}75 \mp t_{0{,}95,3} SE_{\widehat{Y}(10)-Y(10)},$$

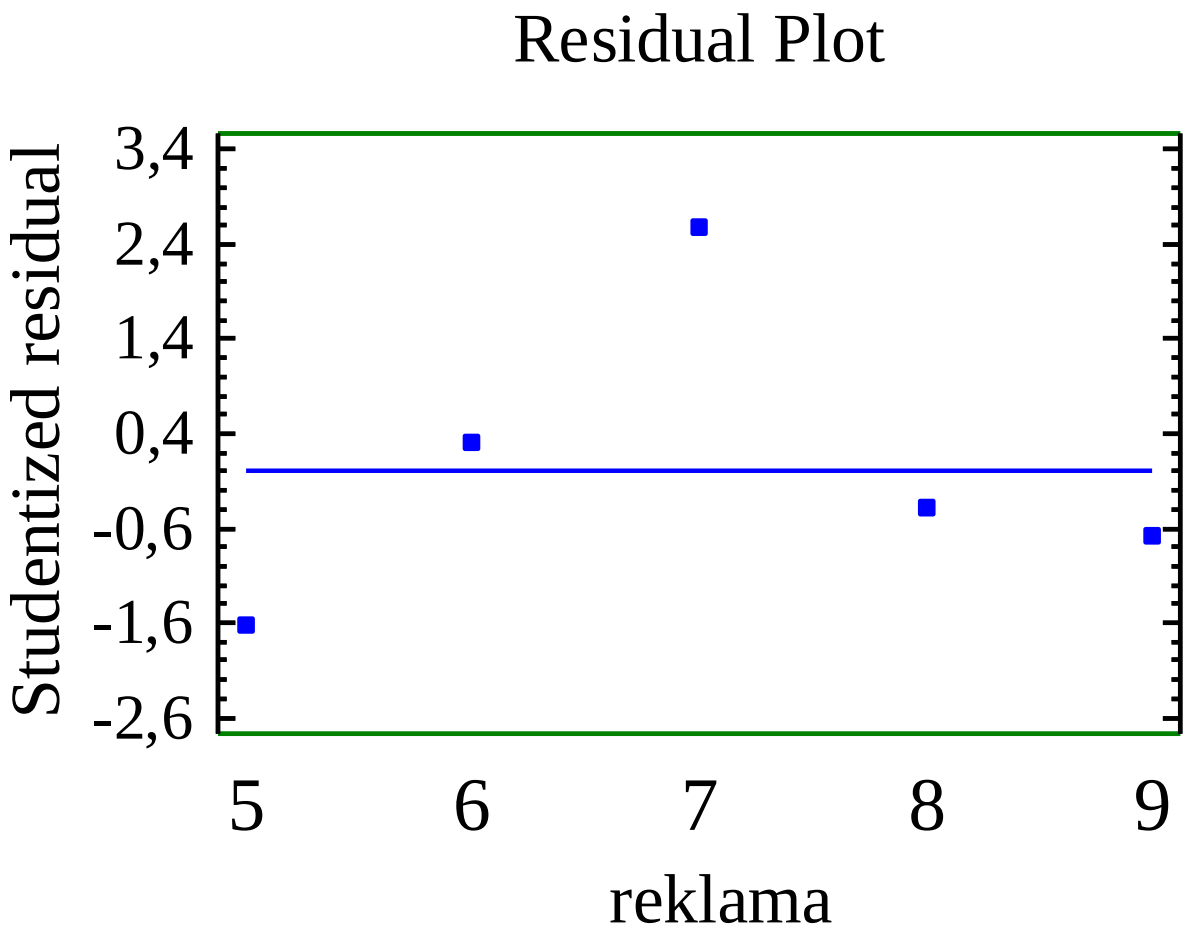
gdzie $$SE_{\widehat{Y}(x_0)-Y(x_0)} = S\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}} =$$

= 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
dochod
14
13
12
11
10
9
8
7
6
5
4
3
2
1
3
4
5
6
7
8
9
10
11
12
reklama


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: $r_i = \frac{e_i}{SE_{e_i}}$, $i = 1,\ldots,n.$

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}.

(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**: $\mathrm{Var}(\varepsilon_i) = \sigma_i^2$. 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:

PRODUKCJA = 6,40 + 2,20 × 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: $y = b_0 + b_1 x$

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

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

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

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

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

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

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

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

Mamy zatem przy spełnionej hipotezie H_0:

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