Czynniki deterministyczne

$$m = 1 \qquad p = 1$$

Obserwujemy cechę Y oraz zmienną X
Obiekt $\longrightarrow (X, Y)$

1. Propozycja funkcji regresji f.

2. Dopasowanie zaproponowanej funkcji.

3. Ocena jakości dopasowania.

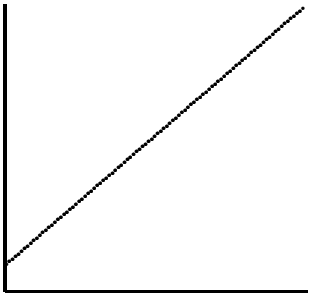
4. Wnioski.

Założenie:
Cecha Y ma rozkład normalny

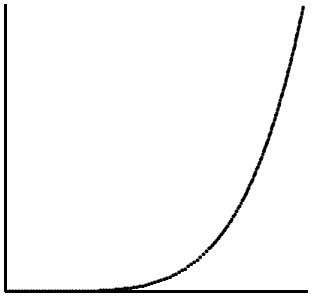
Regresja prosta

Funkcja regresji zależna tylko od jednego argumentu

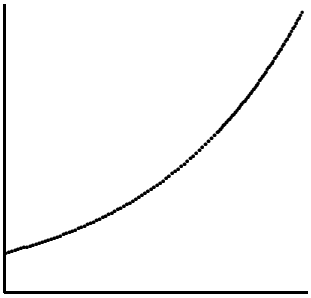
Nazwa funkcji	Wzór funkcji	Model
Liniowa	$a + bx$	$y = a + bx$
Potęgowa	ax^b	$\ln y = \ln a + b \ln x$
Wykładnicza	$\exp(a + bx)$	$\ln y = a + bx$
Typu S	$\exp(a + \frac{b}{x})$	$\ln y = a + b\frac{1}{x}$
Hiperboliczna	$\frac{1}{a+bx}$	$\frac{1}{y} = a + bx$
Podwójnie hiperboliczna	$\frac{1}{a+b/x}$	$\frac{1}{y} = a + b\frac{1}{x}$
Pierwiastkowa	$a + b\sqrt{x}$	$y = a + b\sqrt{x}$
Logarytmiczna	$a + b \ln x$	$y = a + b \ln x$



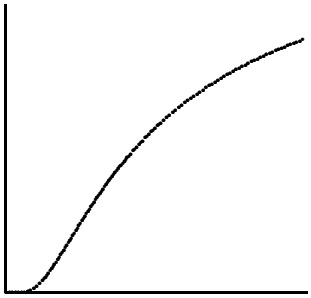
regresja liniowa



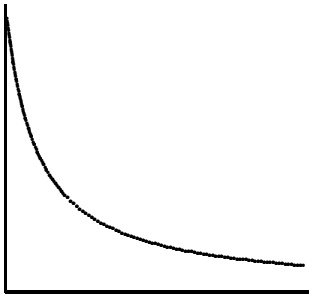
regresja potęgowa



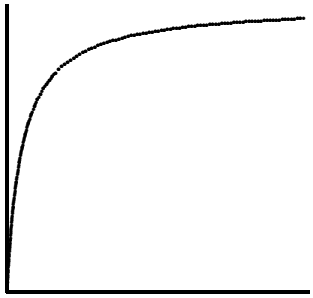
regresja wykładnicza



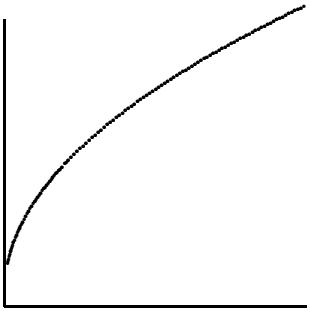
regresja typu S



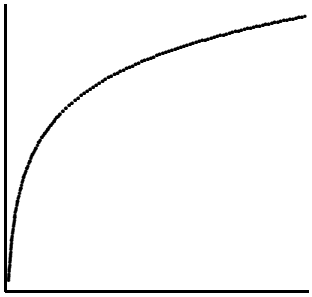
regresja hiperboliczna



regresja podwójnie hiperboliczna



regresja pierwiastkowa



regresja logarytmiczna

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

$(Y_1, X_1), \ldots, (Y_n, X_n)$ — obserwacje
Model

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

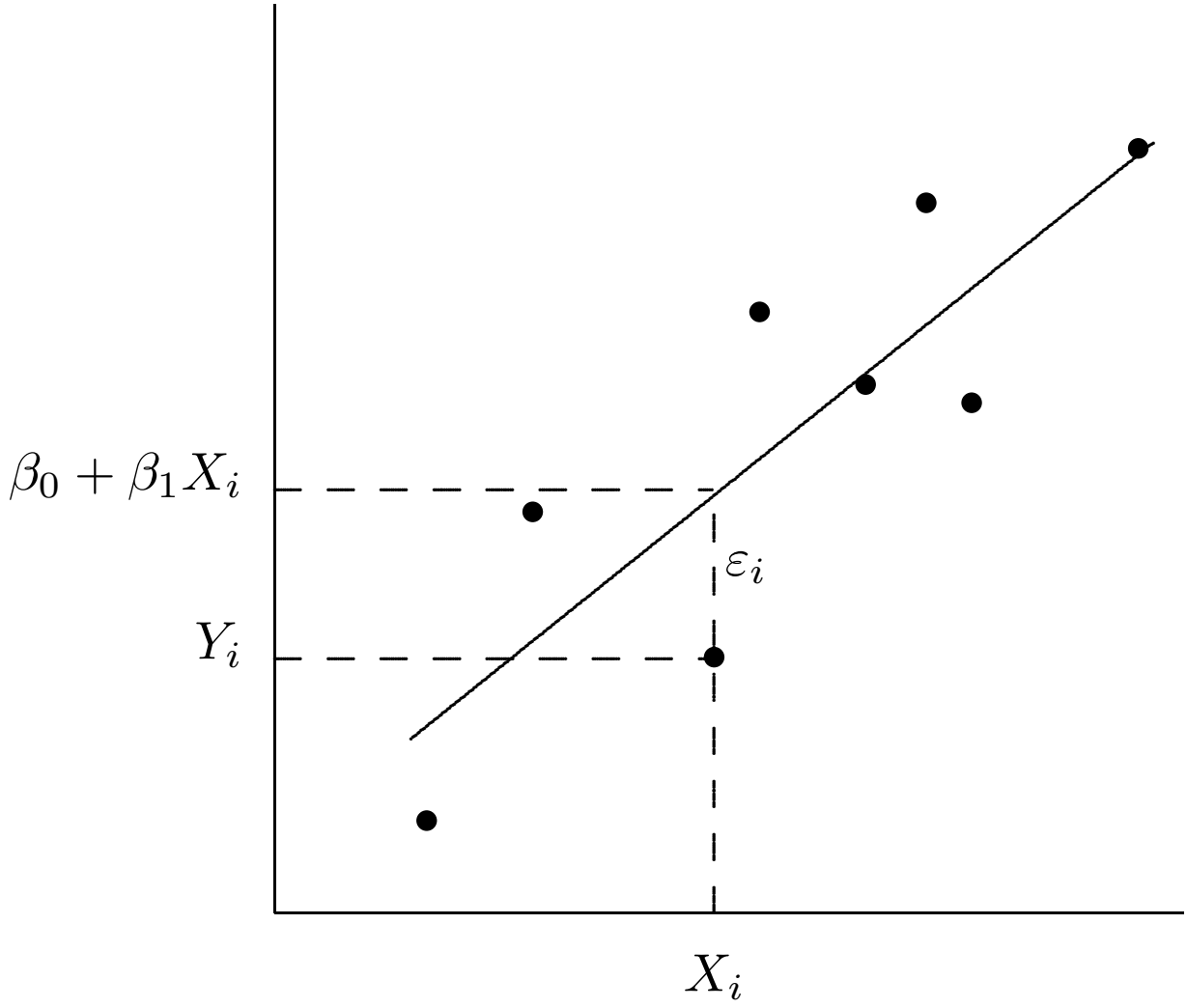
ε_i są niezależnymi zmiennymi losowymi o tym samym rozkładzie normalnym $N(0, \sigma^2)$.

Estymacja współczynników metodą najmniejszych kwadratów

Znaleźć takie β_0 i β_1 by

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

Metoda najmniejszych kwadratów



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

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

Rozwiązanie

$$\widehat{\beta}_0 = \bar{Y} - \widehat{\beta}_1 \bar{x}$$
$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\mathrm{cov}(X,Y)}{\mathrm{var}X}$$

Resztowa suma kwadratów

$$\mathrm{RSS} = \sum_{i=1}^{n}(Y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_i))^2$$

Ocena wariancji σ^2

$$S^2 = \left(\frac{\mathrm{var}Y - \widehat{\beta}_1 \mathrm{cov}(X,Y)}{n-2} \right) = \frac{\mathrm{var}Y - \mathrm{var}R}{n-2}$$

Istnienie zależności
Weryfikacja hipotezy $H_0 : \beta_1 = 0$

Źródło zmienności	Suma kwadratów	Stopnie swobody	Średnie kwadraty	F
Regresja	var$R$	1	s_R^2=var$R$	s_R^2/s^2
Błąd	RSS	$n-2$	s^2=RSS$/(n-2)$	
Całkowita	var$Y$	$n-1$		

$$\text{var}R = \widehat{\beta}_1 \text{cov}(X, Y), \ \text{var}Y = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

$$\text{var}Y = \text{var}R + RSS \Rightarrow RSS = \text{var}Y - \text{var}R$$

Jeżeli hipoteza $H_0 : \beta_1 = 0$ jest prawdziwa, to

$$F = \frac{s_R^2}{s^2}$$

ma rozkład F z $(1, n-2)$ stopniami swobody

Hipotezę odrzucamy, jeżeli $F > F(\alpha; 1, n-2)$
$F(\alpha; 1, n-2)$ — wartość krytyczna rozkładu F.

Rozkład zmienności cechy Y:

Niech

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

oraz niech

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i, i = 1, \ldots, n$$

$$\mathrm{var}Y = RSS + \mathrm{var}R$$

$$\sum (Y_i - \bar{Y})^2 = \sum (Y_i - \widehat{Y}_i)^2 + \sum (\widehat{Y}_i - \bar{Y})^2$$

var$R$ określa tą część zmienności cechy Y, która jest opisana przez funkcję regresji, natomiast RSS dotyczy zmienności opisanej przez reszty e_{ij}

Dla par $(Y_i, \widehat{Y}_i)$ wyznaczamy

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

Współczynnik determinacji

$$D = R^2 \cdot 100\% = \frac{\sum(\widehat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} \cdot 100\%.$$

Jest to liczba z przedziału $(0\%, 100\%)$ i dopasowanie funkcji regresji jest tym lepsze, im ten współczynnik jest wyższy.

$$D = \frac{\mathrm{var}R}{\mathrm{var}Y} \cdot 100\%$$

Współczynnik determinacji — procent zmienności cechy Y wyjaśniany przez funkcję regresji.

Jeżeli funkcja regresji jest funkcją liniową, to

$$D = \frac{(\mathrm{cov}(Y, X))^2}{\mathrm{var}X\,\mathrm{var}Y} \cdot 100\%$$

Przedział ufności dla β_1
Poziom ufności $1-\alpha$

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

$$S^2_{\beta_1} = \frac{S^2}{\mathrm{var}X}$$

Interpretacja współczynnika regresji β_1
jeżeli wartość zmiennej niezależnej x wzrośnie o jednostkę, to średnia wartość cechy Y zmieni się (wzrośnie lub zmaleje) o około $\widehat{\beta}_1$ jednostek, a dokładniej zmieni się o $\widehat{\beta}_1 \pm t(\alpha; n-2)S_{\beta_1}$ jednostek.

Przedział ufności dla β_0
Poziom ufności $1-\alpha$

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

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

Obszar ufności dla prostej regresji

$$y = \beta_0 + \beta_1 x$$

Średnia wartość cechy Y dla ustalonego $X = x$

$$\widehat{y}(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x$$

Obszar ufności (poziom ufności $1 - \alpha$)

$$E(Y|x) \in (\widehat{y}(x) - t(\alpha; n-2)S_Y; \widehat{y}(x) + t(\alpha; n-2)S_Y)$$

$$S_Y^2 = S^2 \left(\frac{1}{n} + \frac{(x - \bar{X})^2}{\mathrm{var}X} \right)$$

Na podstawie obszaru ufności wnioskujemy o **wartościach średnich** cechy Y jednocześnie dla wielu wybranych wartości cechy X

Predykcja wartości zmiennej $Y(x)$

Wartość cechy Y dla ustalonego $X = x$

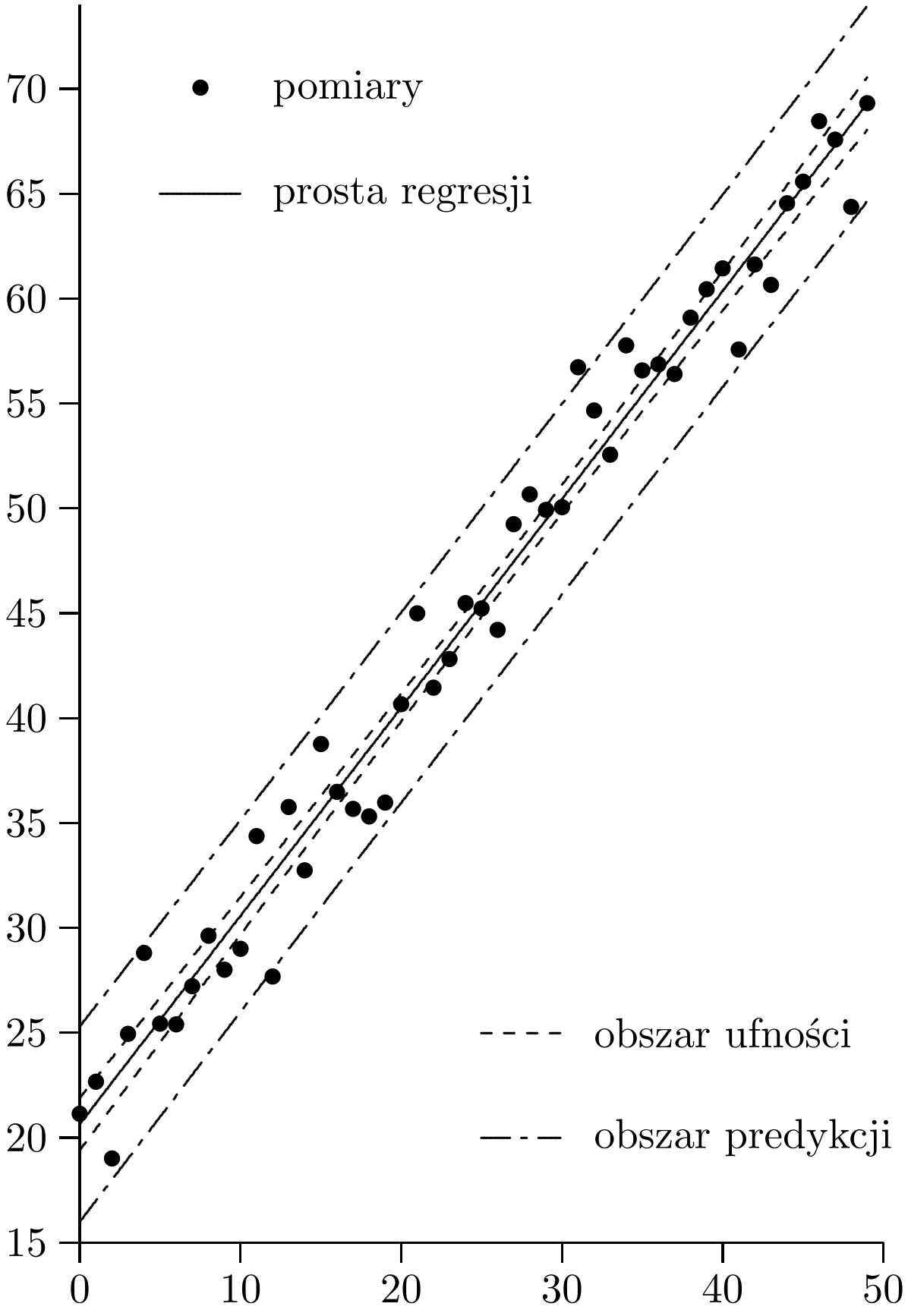
$$\widehat{y}(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x$$

Obszar predykcji (poziom ufności $1 - \alpha$)

$$Y(x) \in (\widehat{y}(x) - t(\alpha; n-2) S_{y(x)}; \widehat{y}(x) + t(\alpha; n-2) S_{y(x)})$$

$$S^2_{y(x)} = S^2 \left(1 + \frac{1}{n} + \frac{(x - \bar{X})^2}{\mathrm{var} X} \right)$$

Na podstawie obszaru predykcji wnioskujemy o **wartościach** cechy Y jednocześnie dla wielu wybranych wartości cechy X

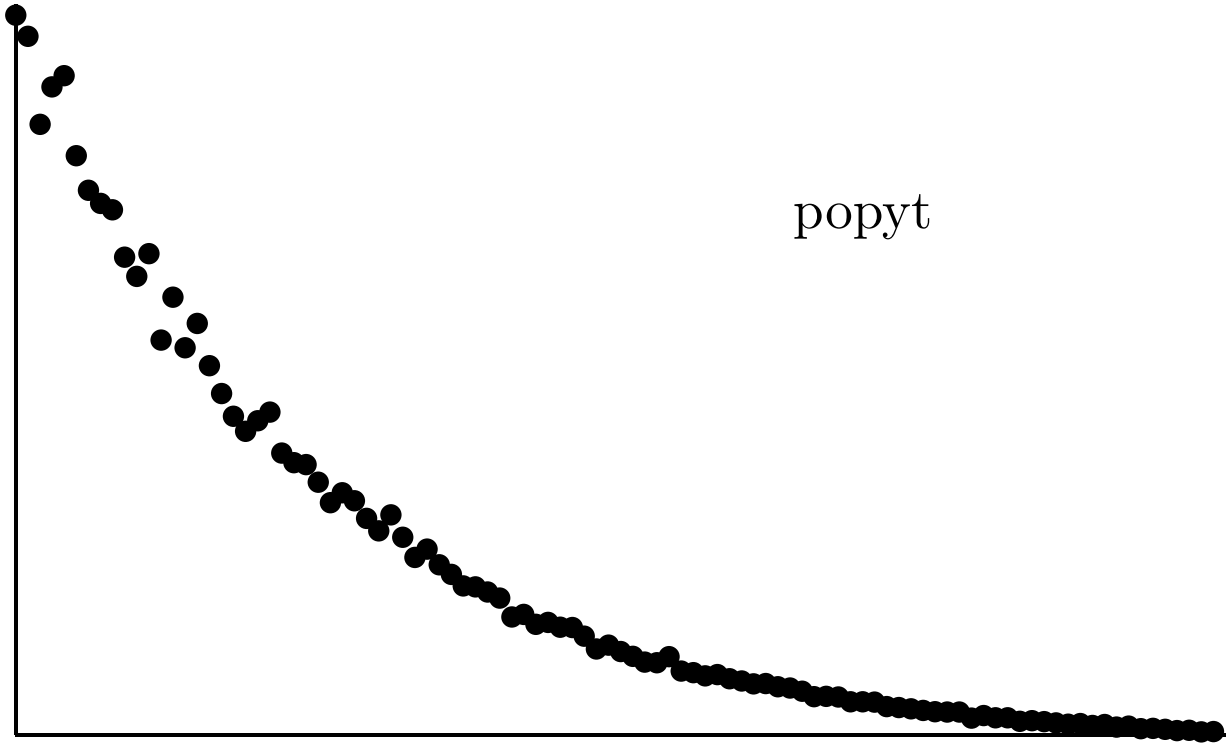

pomiary
prosta regresji
obszar ufności
obszar predykcji
70
65
60
55
50
45
40
35
30
25
20
15
0
10
20
30
40
50

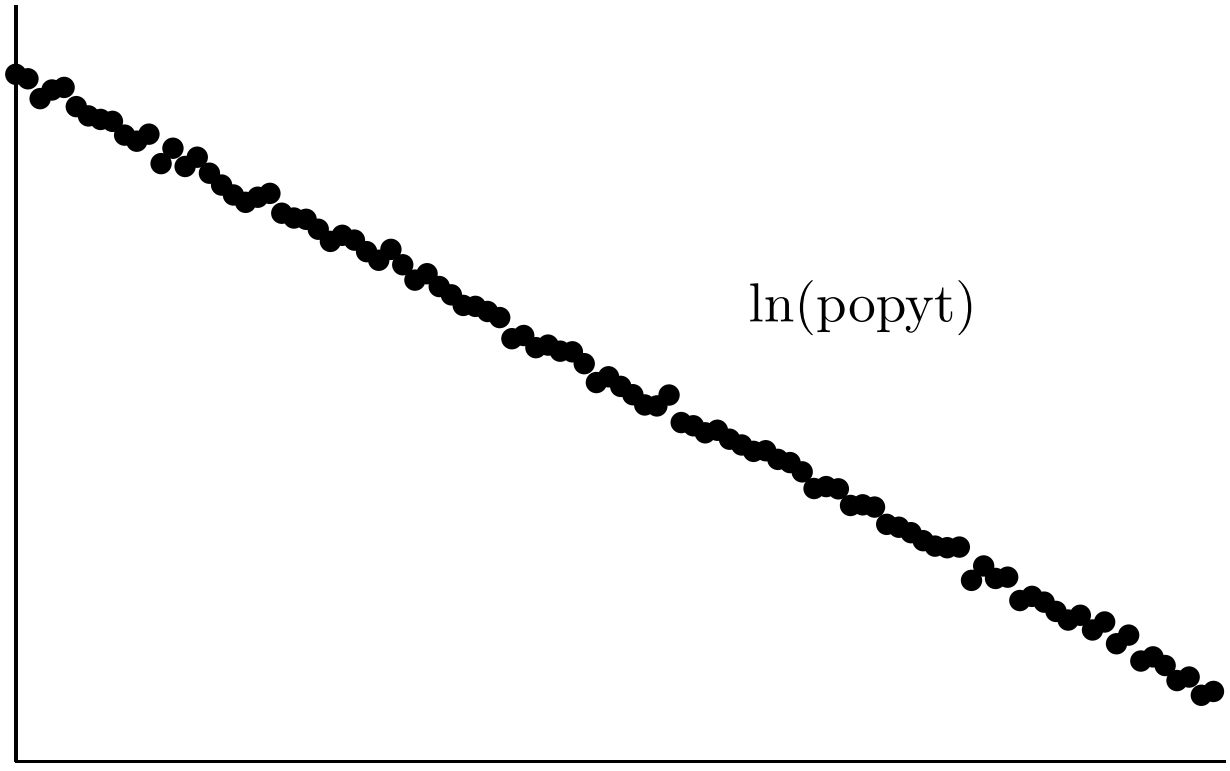
Przykład. Badano zależność między ceną towaru a popytem na ten towar. Na podstawie poniższych danych przeprowadzić analizę regresji.

Cena	Popyt	Cena	Popyt	Cena	Popyt	Cena	Popyt	Cena	Popyt
0	688.6660	20	305.2350	40	137.6316	60	59.2920	80	26.6664
1	668.5006	21	313.3761	41	119.8710	61	56.8174	81	24.5516
2	585.2404	22	274.8090	42	122.3169	62	57.1199	82	24.7466
3	620.7191	23	265.8115	43	112.8871	63	53.9587	83	21.1797
4	631.3066	24	263.9127	44	114.8166	64	52.7557	84	21.7877
5	555.6787	25	247.1726	45	110.2780	65	49.6954	85	20.9799
6	523.0394	26	227.9754	46	109.9141	66	44.5124	86	19.7234
7	510.7502	27	237.3264	47	101.6284	67	44.9967	87	18.6367
8	504.5449	28	229.8074	48	89.6586	68	44.3595	88	19.2405
9	459.9711	29	213.1303	49	93.1468	69	39.7275	89	17.4347
10	441.8514	30	201.1405	50	87.3825	70	39.8829	90	18.3870
11	463.2068	31	216.3138	51	82.7416	71	39.3322	91	15.9288
12	381.5119	32	195.3679	52	77.3328	72	35.1059	92	16.8617
13	422.1652	33	176.2760	53	76.7411	73	34.4407	93	14.2136
14	374.1622	34	184.1000	54	82.4336	74	33.2335	94	14.6192
15	397.3657	35	169.1835	55	68.8193	75	31.4761	95	13.7955
16	357.4174	36	160.1516	56	67.3987	76	30.4003	96	12.4930
17	331.1023	37	149.3296	57	64.2539	77	30.0409	97	12.7864
18	309.4546	38	148.2669	58	65.4210	78	30.1891	98	11.3360
19	295.2351	39	143.3767	59	61.5881	79	24.2257	99	11.6303

Plan działania

1. Propozycja funkcji regresji
2. Dopasowanie funkcji regresji
3. Istnienie zależności
4. Jakość dopasowania
5. Wnioski


popyt


ln(popyt)

Funkcja regresji

$$\text{popyt} = \exp(a + b \cdot \text{cena})$$

Model

$$\ln(\text{popyt}) = \beta_0 + \beta_1 \cdot \text{cena}$$

$$Y = \ln(\text{popyt}) \qquad x = \text{cena}$$

$$\beta_0 = a \qquad \beta_1 = b$$

Dopasowanie funkcji regresji

$$\bar{x} = 49.50 \qquad \bar{y} = 4.5053771$$

$$\text{var}x = 82500 \quad \text{var}y = 141.2006027$$

$$\text{cov}(x, y) = -3427.7924020$$

$$\widehat{\beta}_0 = 6.541689 \qquad \widehat{\beta}_1 = -0.04114$$

$$s^2 = 0.001932$$

Istnienie zależności

$$H_0 : \beta_1 = 0$$

Źródło zmienności	Suma kwadratów	Stopnie swobody	Średnie kwadraty	F
Regresja	141.0112	1	141.0112	72973.08
Błąd	0.189373	98	0.001932	
Całkowita	141.2006	99		

Wartość krytyczna

$$F(0.05; 1, 98) = 3.938$$

Wniosek:
zaproponowana funkcja regresji może opisywać zależność między ceną a popytem na towar

Współczynnik determinacji

$$D = 99.866\%$$

Zastosowanie

Obszar ufności dla oczekiwanego popytu na towar przy cenie $x = 30$

dla Y	$(5.297033, 5.318088)$
dla popytu	$(199.7434, 203.9935)$

Elastyczność cenowa popytu

$$\frac{d\ln(\exp(a + b \cdot \text{cena}))}{d\ln(\text{cena})} = b \cdot \text{cena} = -0.04114 \cdot \text{cena}$$