Ocena zależności

Badanie związku między cechami

Cele:

- Stwierdzić, czy istnieje zależność między cechami w próbie
- Stwierdzić, czy zależność jest istotna
- Ocenić siłę zależności

Ze względu na rodzaje cech w próbach dwuwymiarowych oraz na rodzaj zależności, dzielimy na:

- połączenie cechy jakościowej lub ilościowej skokowej z dowolną
- połączenie cechy ilościowej ciągłej z cechą ilościową ciągłą

WARIANT 1

- połączenie cechy jakościowej lub ilościowej skokowej z dowolną

H_o: cechy są niezależne
H_1: cechy są zależne

		umaszczenie			$n_{i.}$
		n_{ij} czarne	czekolada	biszkopt	
miasto	Warszawa	26	4	50	80
	Kraków	34	6	30	70
	Gdańsk	20	10	20	50
	$n_{.j}$	80	20	100	200 N

$$p_{i.} = \frac{n_{i.}}{N} \qquad p_{.j} = \frac{n_{.j}}{N} \qquad \hat{n}_{ij} = p_{i.} \cdot p_{.j} \cdot N = \frac{n_{i.} \cdot n_{.j}}{N}$$

Teoretyczne liczebności

		umaszczenie			
		czarne	czekolada	biszkopt	
miasto	Warszawa	26 32	4 8	50 40	80
	Kraków	34 28	6 7	30 35	70
	Gdańsk	20 20	10 5	20 25	50
		80	20	100	200

$$\chi^2_{emp} = \sum_{i,j} \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}} = \frac{(26-32)^2}{32} + \ldots + \frac{(20-25)^2}{25} = 13{,}77$$

Weryfikacja hipotezy o niezależności cech

H_0 : cechy są niezależne

H_1 : cechy są zależne

$$\chi^2_{emp} = 13{,}77$$

Rozkład chi-kwadrat o czterech stopniach swobody,

gdyż v = (r-1)(k-1)

$$P(\chi^2 \geq 13{,}77) = 0{,}00807$$

Cecha jakościowa z inną

Współczynnik zbieżności T - Czuprowa

$$T = \sqrt{\frac{\chi^2_{emp}}{N \cdot \sqrt{(r-1)(k-1)}}}$$

Informuje o sile zależności,

Jego wartość zawiera się w przedziale od <0 ; 1>

$$T = \sqrt{\frac{13{,}77}{200 \cdot \sqrt{(3-1)(3-1)}}} = 0{,}1855$$

Zależność wysoko istotna, ale słaba

LICZEBNOŚCI EMPIRYCZNE

x \ y	duże miasto	małe miasto	wieś	n_i
podstawowe	25	30	42	97
zawodowe	36	38	29	103
średnie	43	33	26	102
wyzsze	51	32	23	106
n_j	155	133	120	408

LICZEBNOŚCI TEORYCZNE

x \ y	duże miasto	małe miasto	wieś	n_i
podstawowe	36,85	31,62	28,53	97
zawodowe	39,13	33,58	30,29	103
średnie	38,75	33,25	30,00	102
wyzsze	40,27	34,55	31,18	106
n_j	155	133	120	408

Chi-kwadrat $_{emp}$	**17,3366**
Liczba st.swobody	**6**
Istotność	**0,00812271**
WSPÓŁCZYNNIK ZBIEŻNOŚCI V	**0,131708489**
WSPÓŁCZYNNIK DETERMINACJI V^2	**1,7%**

WARIANT 2a:

- połączenie cechy ilościowej ciągłej z cechą ilościową ciągłą, gdy obie mają w populacji rozkład normalny

KOWARIANCJA

$$\mathrm{cov}_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})\cdot(y_i - \bar{y}) =$$

$$= \frac{1}{N-1}\left(\sum_{i=1}^{N} x_i \cdot y_i - \frac{\sum_{i=1}^{N} x_i \cdot \sum_{i=1}^{N} y_i}{N}\right)$$

KOWARIANCJA (cov_{xy})

Informuje

- o istnieniu lub braku liniowej zależności między cechami,
- o kierunku zależności,

Jej wartość zawiera się w przedziale $\langle -S_x S_y ; S_x S_y \rangle$

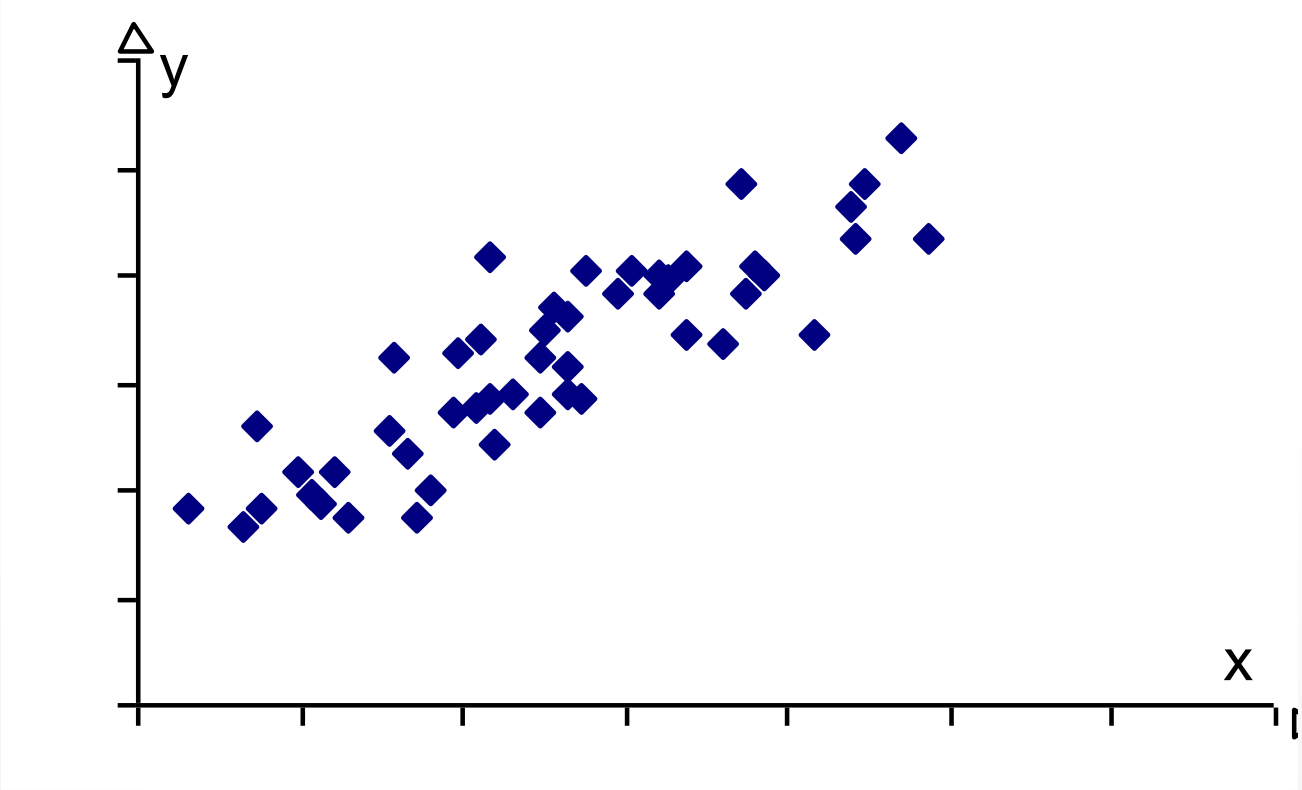
WSPÓŁCZYNNIK KORELACJI LINIOWEJ PEARSONA

$$r_{xy} = \frac{\mathrm{cov}_{xy}}{S_x \cdot S_y}$$

Informuje o sile i kierunku zależności,

Jego wartość zawiera się w przedziale od <-1 ; 1>

KORELACJA SILNA, DODATNIA



Przykład

Wydajność mleczna [kg]	Zawartość tłuszczu [%]
5400	3,04
4800	3,09
6200	3,06
4700	3,15
4750	3,06
4800	3,17
5800	3,07
6250	3,13
5200	3,14
5500	3,09

Przykład

Wydajność mleczna [kg]	Zawartość tłuszczu [%]
5400	3,04
4800	3,09
6200	3,06
4700	3,15
4750	3,06
4800	3,17
5800	3,07
6250	3,13
5200	3,14
5500	3,09

$$\mathrm{cov}_{xy} = -8{,}056$$

$$S_x^2 = 352111{,}1$$

$$S_y^2 = 0{,}001978$$

$$r_{xy} = \frac{-8{,}056}{\sqrt{352111{,}1 \cdot 0{,}001978}} = -0{,}3053$$

Hipoteza porównująca wartość współczynnika korelacji z stałą $H_0 : \rho_{XY} = \rho_0$

$$u_{emp} = \left(\frac{1}{2} \ln \frac{1+r}{1-r} - \frac{1}{2} \ln \frac{1+\rho_0}{1-\rho_0} - \frac{\rho_0}{2 \cdot (N-1)} \right) \cdot \sqrt{N-3}$$

Hipoteza porównująca wartość współczynnika korelacji z zerem $H_0 : \rho_{XY} = 0$

$$t_{emp} = \frac{r}{\sqrt{1-r^2}} \cdot \sqrt{(N-2)}$$

Przykład

$$t_{emp} = \frac{-0{,}3053}{\sqrt{1-(-0{,}3053)^2}} \cdot \sqrt{(10-2)} = -0{,}9068$$

$P(t \leq -0{,}9068) = 0{,}1955$ dla H_1 jednostronnej

$P(|t| \geq |-0{,}9068|) = 0{,}391$ dla H_1 dwustronnej

Korelacja ujemna, słaba i nieistotna

WARIANT 2b:

- połączenie cechy ilościowej ciągłej z cechą ilościową ciągłą, gdy przynajmniej jedna nie ma w populacji rozkładu normalnego

WSPÓŁCZYNNIK KORELACJI RANG SPEARMANA

$$r_s = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N(N^2 - 1)}$$

Przykład

I sędzia	ranga	II sędzia.	ranga	Różnica rang d_i
100	**1**	102	**1,5**	-0,5
101	**2**	102	**1,5**	0,5
107	**6**	108	**6**	0
111	**9**	110	**8**	1
109	**7**	110	**8**	-1
110	**8**	110	**8**	0
106	**5**	107	**5**	0
103	**3**	103	**3,5**	-0,5
112	**10**	111	**10**	0
104	**4**	103	**3,5**	0,5

Przykład

$$r_s = 1 - \frac{6 \cdot 3}{10(100-1)} = 1 - \frac{18}{990} = 0{,}9818$$

$$t_{emp} = \frac{0{,}9818}{\sqrt{\frac{1-0{,}9818^2}{10-2}}} \approx 14{,}622$$

$$P(|t| \geq 14{,}622) = 0{,}0000$$

Korelacja bardzo silna i wysoce istotna

ranga 1	ranga 2	d_i	d_i^2
1	12	-11	121
2	11	-9	81
3	10	-7	49
4	9	-5	25
5	8	-3	9
6	7	-1	1
7	6	1	1
8	5	3	9
9	4	5	25
10	3	7	49
11	2	9	81
12	1	11	121
		0	572

$$r_s = 1 - \frac{6 \cdot 572}{12(144-1)} = 1 - 2 = -1$$

Analiza regresji

Cele:

- Skonstruować funkcję regresji dla próby
- Ocenić dopasowania oraz sprawdzić czy funkcja jest istotna
- Stosować do przewidywania wartości cechy zależnej (pamiętać o błędzie prognozy)

Funkcja liniowa regresji ma postać:

$$\hat{y}_i = b_{yx} \cdot x_i + a_{yx}$$

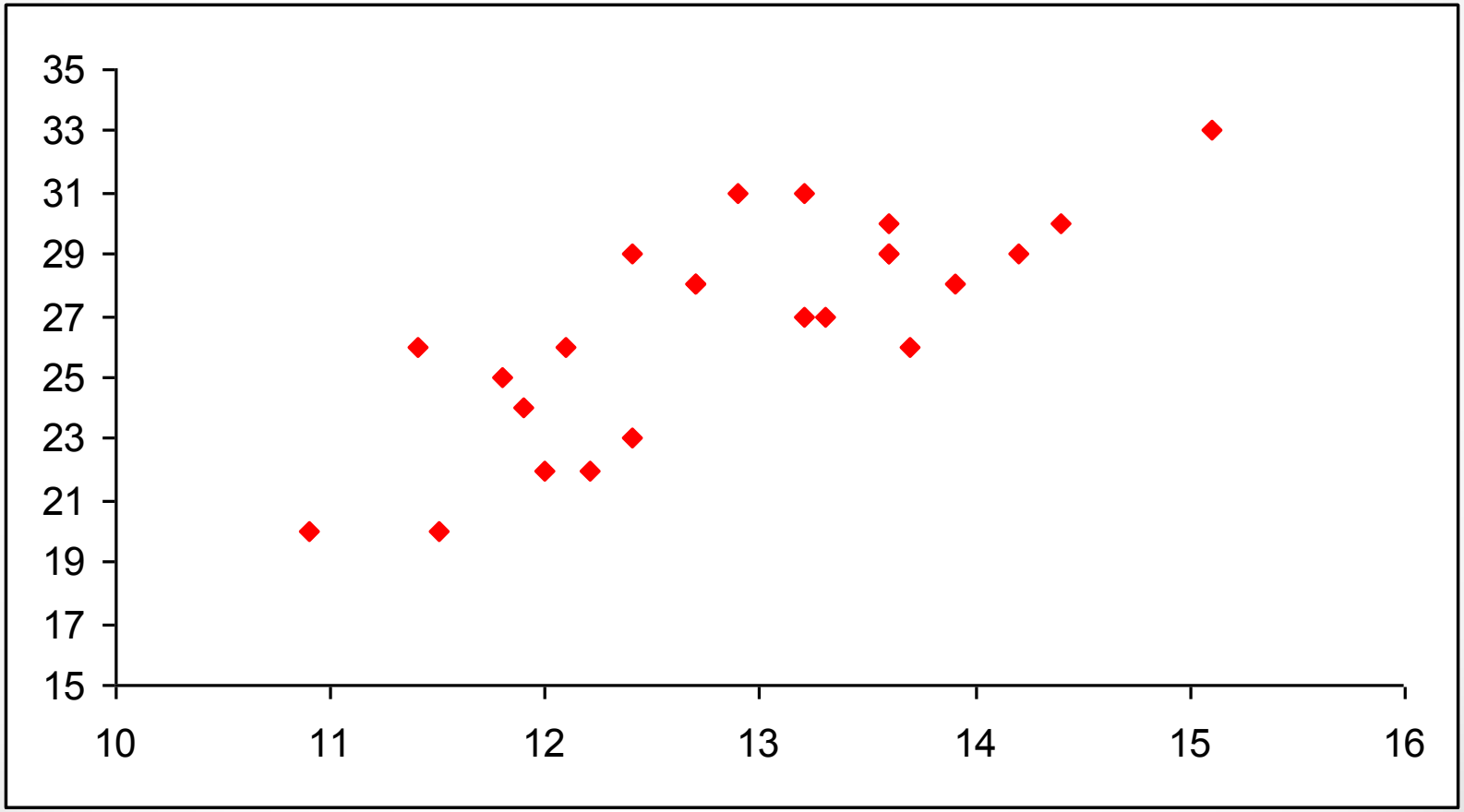
b_{yx} jest współczynnikiem regresji – informuje o ile zmieni się cecha zależna (y) jeśli cecha niezależna (x) wzrośnie o jeden

x	y
12,2	22
13,6	30
13,2	27
14,2	29
13,7	26
12,4	23
11,5	20
15,1	33
11,9	24
13,3	27
12	22
14,4	30
12,7	28
10,9	20
13,9	28
12,7	28
13,2	31
11,4	26
11,8	25
12,1	26
12,9	31
13,6	29
12,4	29
13,6	29
308,7	**643**
12,863	26,792

Przykład:

Zależność wyniku egzaminu (Y) od wieku nastolatka

Analizowane cechy x (wiek) y (umiejętności)



Metoda najmniejszych kwadratów

Współczynniki równania regresji liniowej są obliczone metodą najmniejszych kwadratów, w której minimalizowana jest suma kwadratów odchyleń wartości rzeczywistych od szacowanych cechy y (reszt):

$$\min\left(\sum_i (y_i - \hat{y}_i)^2\right)$$

$$b_{yx} = \frac{\mathrm{cov}_{xy}}{S_x^2}$$

$$a_{yx} = \bar{y} - b_{yx} \cdot \bar{x}$$

x	y	$x-x_{sr}$	$y-y_{sr}$	$(x-x_{sr})(y-y_{sr})$	$(x-x_{sr})^2$	$(y-y_{sr})^2$
12,2	22	-0,663	-4,792	3,174	0,439	22,960
13,6	30	0,737	3,208	2,366	0,544	10,293
13,2	27	0,337	0,208	0,070	0,114	0,043
14,2	29	1,338	2,208	2,954	1,789	4,877
13,7	26	0,837	-0,792	-0,663	0,701	0,627
12,4	23	-0,463	-3,792	1,754	0,214	14,377
11,5	20	-1,363	-6,792	9,254	1,856	46,127
15,1	33	2,238	6,208	13,891	5,006	38,543
11,9	24	-0,963	-2,792	2,687	0,926	7,793
13,3	27	0,437	0,208	0,091	0,191	0,043
12	22	-0,863	-4,792	4,133	0,744	22,960
14,4	30	1,538	3,208	4,933	2,364	10,293
12,7	28	-0,163	1,208	-0,196	0,026	1,460
10,9	20	-1,963	-6,792	13,329	3,851	46,127
13,9	28	1,038	1,208	1,254	1,076	1,460
12,7	28	-0,163	1,208	-0,196	0,026	1,460
13,2	31	0,337	4,208	1,420	0,114	17,710
11,4	26	-1,463	-0,792	1,158	2,139	0,627
11,8	25	-1,063	-1,792	1,904	1,129	3,210
12,1	26	-0,763	-0,792	0,604	0,581	0,627
12,9	31	0,037	4,208	0,158	0,001	17,710
13,6	29	0,737	2,208	1,629	0,544	4,877
12,4	29	-0,463	2,208	-1,021	0,214	4,877
13,6	29	0,737	2,208	1,629	0,544	4,877
308,7	**643**	**0,000**	**0,000**	**66,313**	**25,136**	**283,958**
12,863	26,792			**2,883**	**1,093**	**12,346**

Współczynniki

Współczynnik regresji

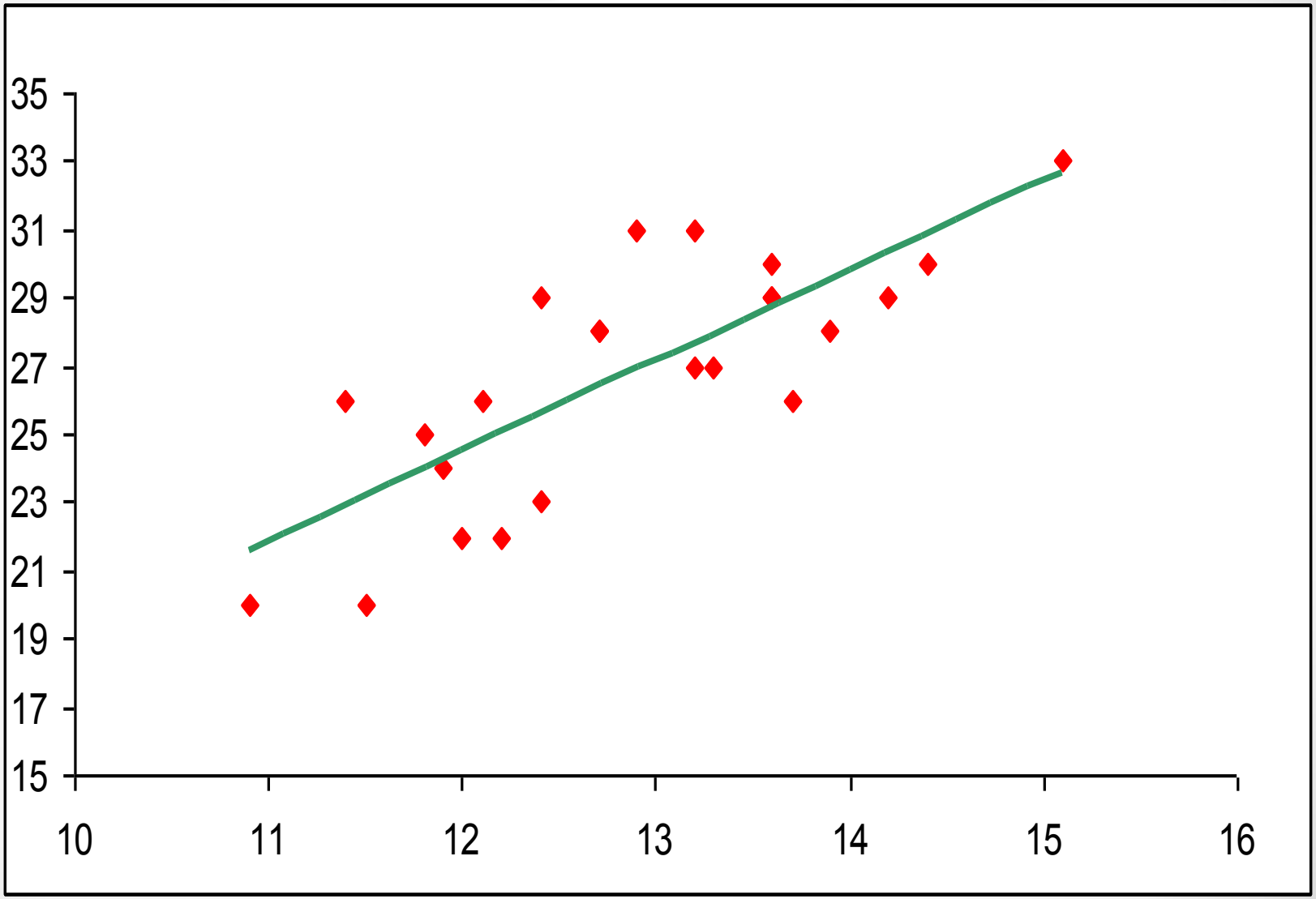
b = 2,638

Stała regresji

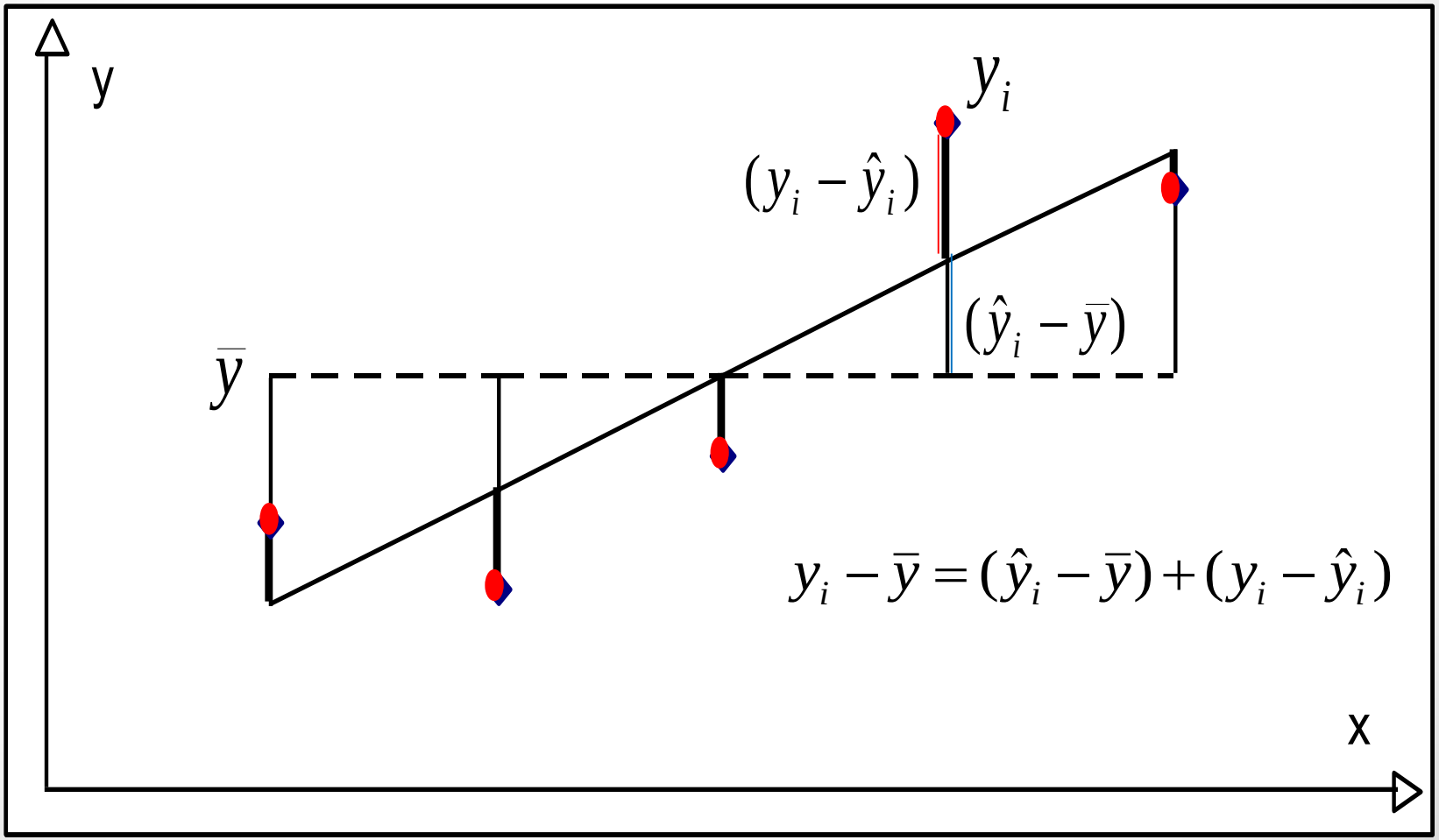
a = -7,141

$$\hat{y}_i = 2{,}368 \cdot x_i - 7{,}141$$

Wykres prostej regresji



Ocena dopasowania funkcji regresji



Ocena dopasowania funkcji regresji

$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$$

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

$$SST = SSR + SSE$$

x	y	$(y-y_{sr})^2$	y_{sz}	$y-y_{sz}$	$y_{sz}-y_{sr}$	$(y-y_{sz})^2$	$(y_{sz}-y_{sr})^2$
12,2	22	22,960	25,044	-3,044	-1,748	9,265	3,055
13,6	30	10,293	28,737	1,263	1,946	1,594	3,785
13,2	27	0,043	27,682	-0,682	0,890	0,465	0,793
14,2	29	4,877	30,320	-1,320	3,528	1,743	12,450
13,7	26	0,627	29,001	-3,001	2,209	9,007	4,882
12,4	23	14,377	25,572	-2,572	-1,220	6,613	1,489
11,5	20	46,127	23,197	-3,197	-3,594	10,222	12,920
15,1	33	38,543	32,694	0,306	5,903	0,093	34,843
11,9	24	7,793	24,252	-0,252	-2,539	0,064	6,447
13,3	27	0,043	27,946	-0,946	1,154	0,895	1,332
12	22	22,960	24,516	-2,516	-2,275	6,332	5,177
14,4	30	10,293	30,848	-0,848	4,056	0,719	16,452
12,7	28	1,460	26,363	1,637	-0,429	2,680	0,184
10,9	20	46,127	21,614	-1,614	-5,177	2,606	26,805
13,9	28	1,460	29,529	-1,529	2,737	2,337	7,491
12,7	28	1,460	26,363	1,637	-0,429	2,680	0,184
13,2	31	17,710	27,682	3,318	0,890	11,009	0,793
11,4	26	0,627	22,933	3,067	-3,858	9,404	14,886
11,8	25	3,210	23,989	1,011	-2,803	1,023	7,857
12,1	26	0,627	24,780	1,220	-2,012	1,488	4,046
12,9	31	17,710	26,891	4,109	0,099	16,887	0,010
13,6	29	4,877	28,737	0,263	1,946	0,069	3,785
12,4	29	4,877	25,572	3,428	-1,220	11,754	1,489
13,6	29	4,877	28,737	0,263	1,946	0,069	3,785
308,7	**643**	**283,958**	**643**	**0**	**0**	**109,018**	**174,940**
		SST				**SSE**	**SSR**

Ocena dopasowania funkcji regresji

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

Współczynnik determinacji R^2 informuje o stopniu dopasowania funkcji regresji

$$R^2 = \frac{174{,}94}{283{,}958} = 0{,}616$$

Ocena istotności liniowej funkcji regresji

H_0: Funkcja regresji jest nieistotna
H_1: Funkcja regresji istotna

$$F_{emp} = \frac{\frac{SSR}{k-1}}{\frac{SSE}{N-k}} = \frac{SSR \cdot (N-k)}{SSE \cdot (k-1)} = \frac{R^2}{\left(1-R^2\right)} \cdot \frac{(N-k)}{(k-1)}$$

$$F_{emp} = \frac{174{,}940 \cdot (24-2)}{109{,}018} = 35{,}303$$

Obszar krytyczny prawostronny

p = 0,0000056

hipotezę odrzucamy

Ocena istotności parametrów funkcji

$H_0 : \beta_{yx} = \beta_0$ dla współczynnika regresji

$H_0 : \alpha_{yx} = \alpha_0$ dla stałej regresji

Obydwie hipotezy weryfikujemy przy pomocy testu t-Studenta, a wartości empiryczne testu obliczane są według wzorów:

$$t_{emp} = \frac{b_{yx} - \beta_0}{S_b}$$

dla współczynnika regresji

$$t_{emp} = \frac{a_{yx} - \alpha_0}{S_a}$$

dla stałej regresji

$$S_b = \sqrt{\frac{S_e^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}} = \sqrt{\frac{S_y^2 - S_x^2 \cdot b_{yx}^2}{S_x^2 \cdot (N-2)}}$$

$$S_a = \sqrt{\frac{S_e^2 \cdot \sum_{i=1}^{N} x_i^2}{N \cdot \sum_{i=1}^{N}(x_i - \bar{x})^2}}$$

Predykcja i jej błąd

Po stwierdzeniu, że skonstruowana funkcja regresji jest istotna, możemy ją stosować do przewidywania (predykcji) wartości cechy zależnej, jeśli znamy wartość cechy niezależnej.

Czyli dla dowolnej wartości cechy X można oszacować przy pomocy funkcji liniowej wartość cechy Y

$$\hat{y}_o = f(x_o)$$

oraz błąd tego przewidywania

$$S_{\hat{y}_0} = S_e \cdot \sqrt{1 + \frac{1}{N} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}}$$

$$S_e = \sqrt{\frac{SSE}{(N-k)}}$$

$$\sum_{i=1}^{N}(x_i - \bar{x})^2 = (N-1) \cdot S_x^2$$

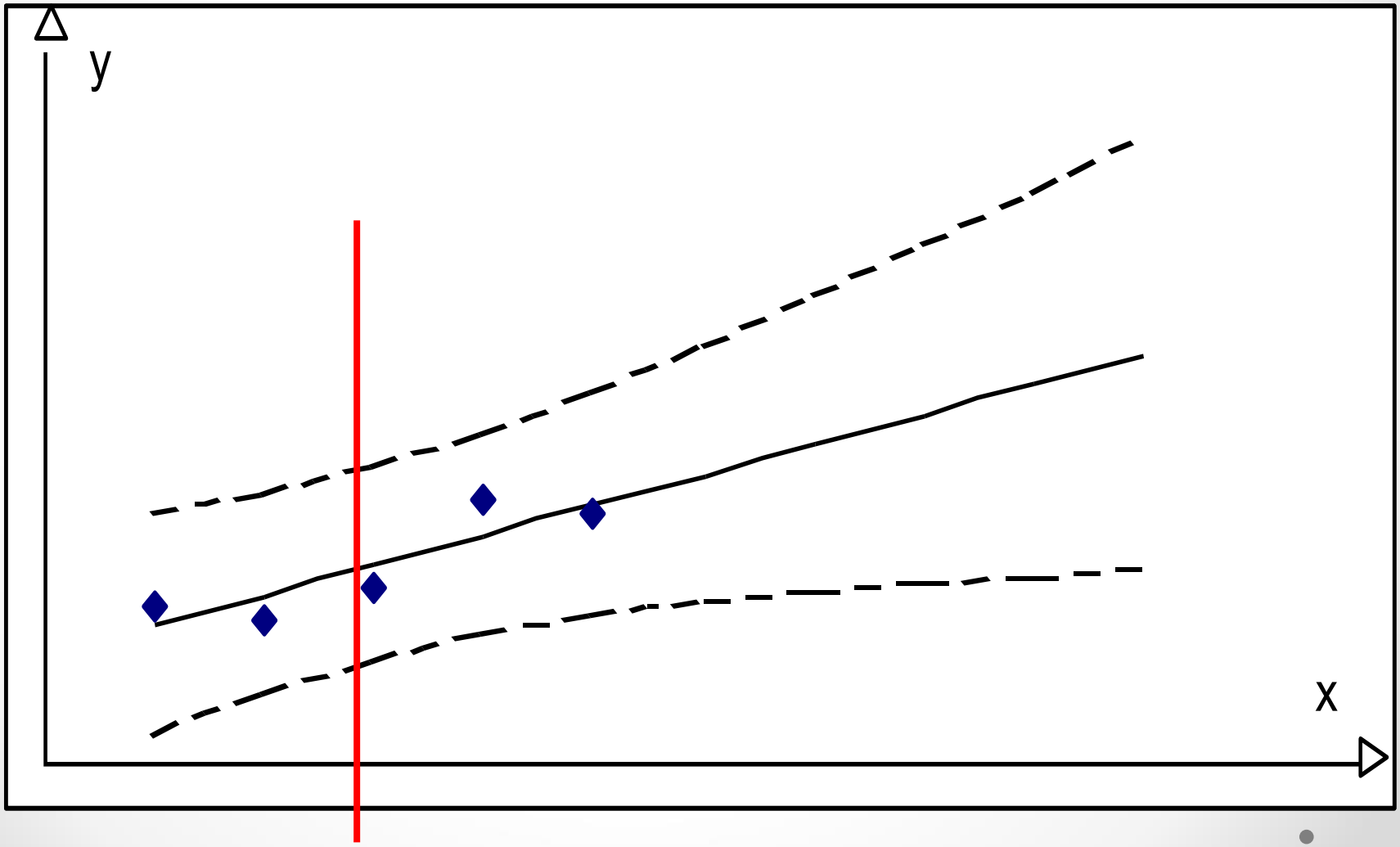
Predykcja

Dla x=16

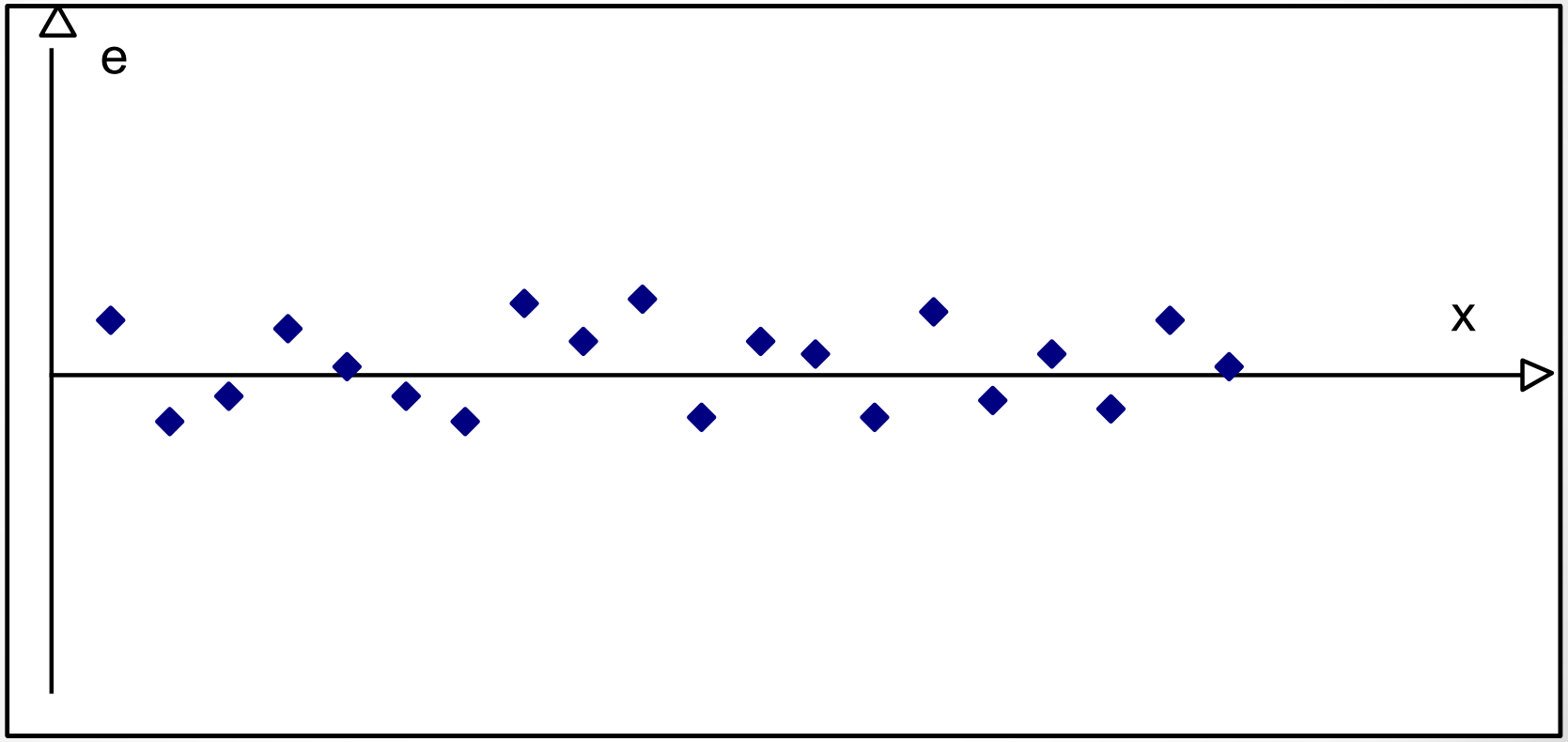
$\hat{y}_0$ =35,069

$$S_{\hat{y}_0} = \sqrt{\frac{109{,}018}{(24-2)}}\sqrt{1+\frac{1}{24}+\frac{(16-12{,}863)^2}{25{,}136}} = 2{,}266 \cdot 1{,}197 = 2{,}665$$

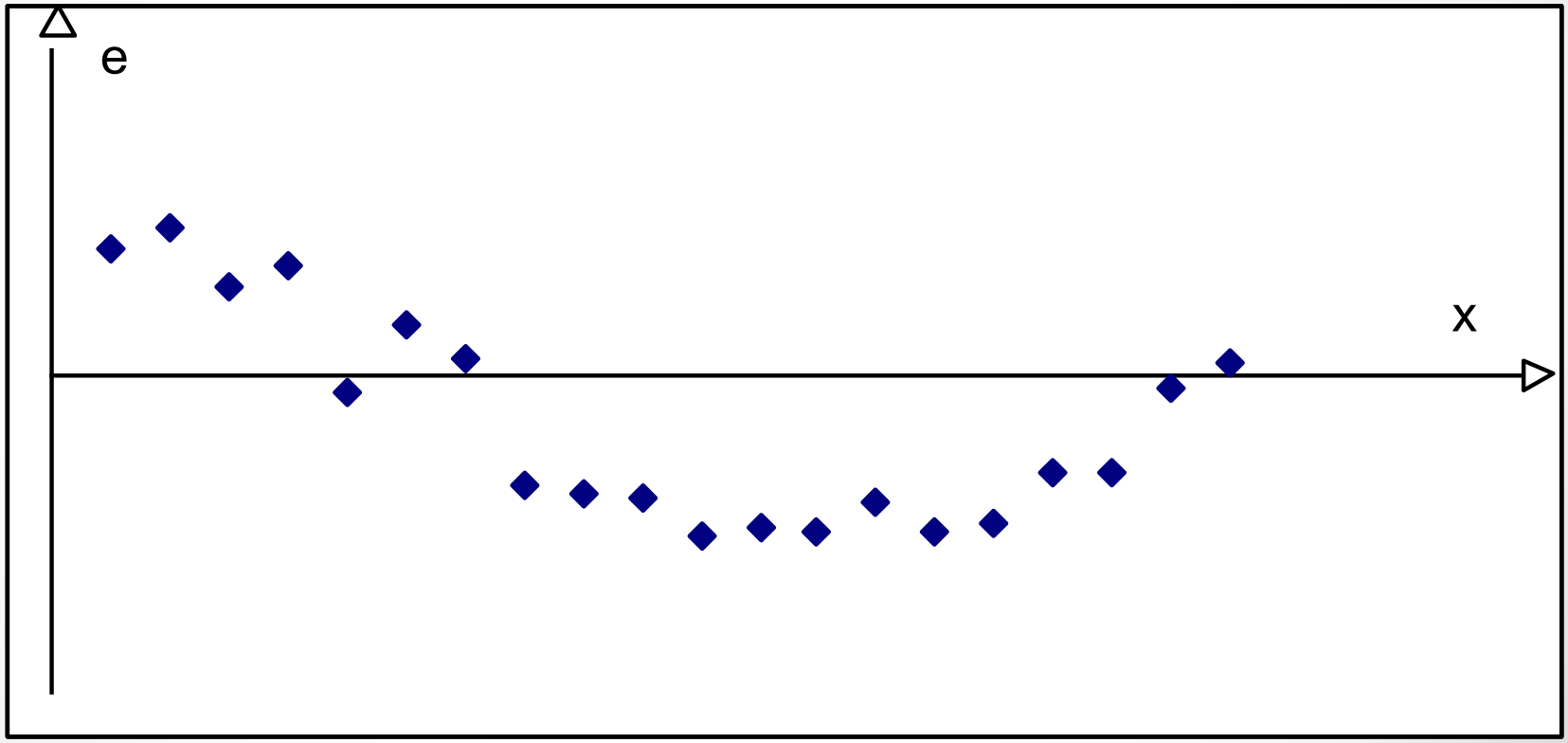
Błąd predykcji



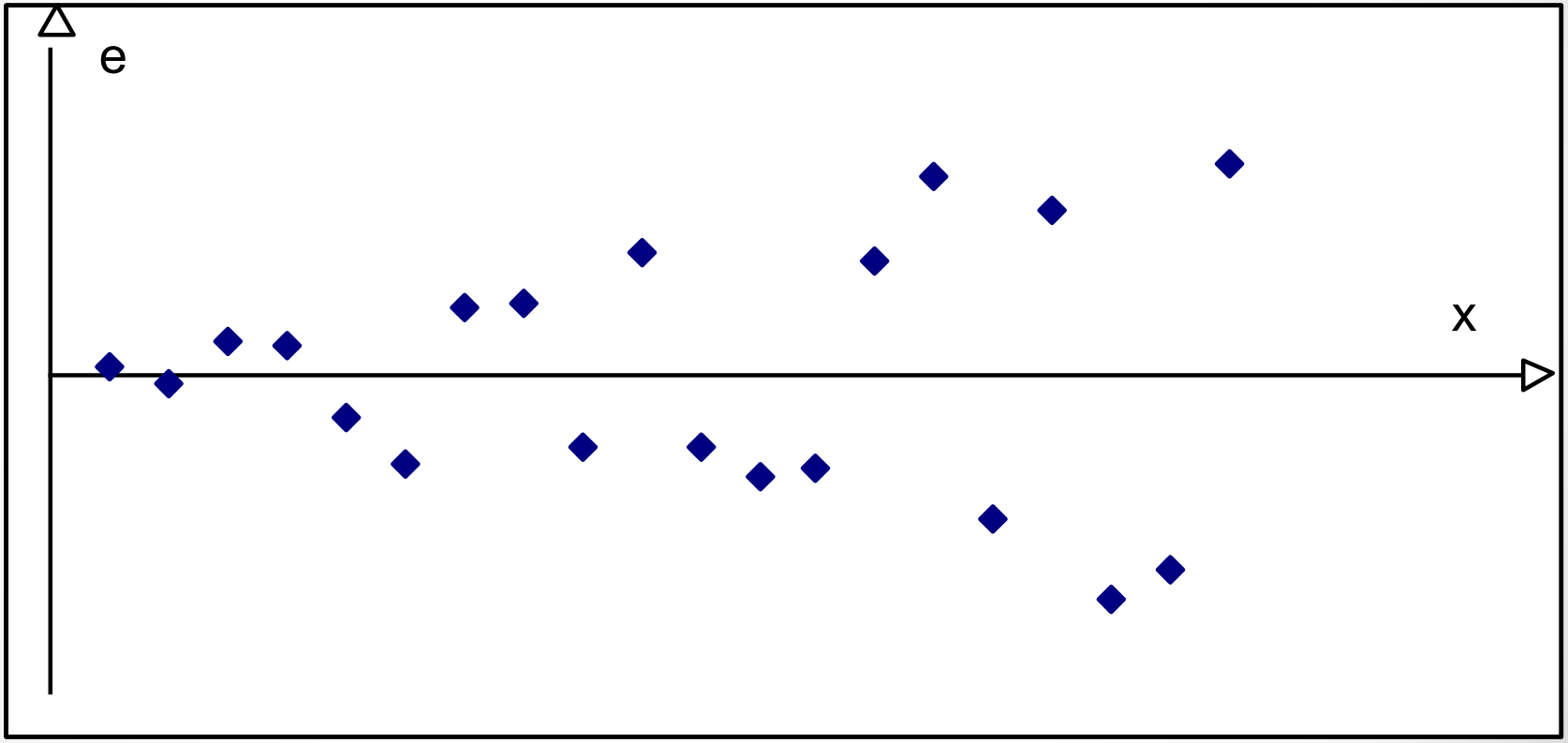
Analiza reszt



Brak losowości reszt

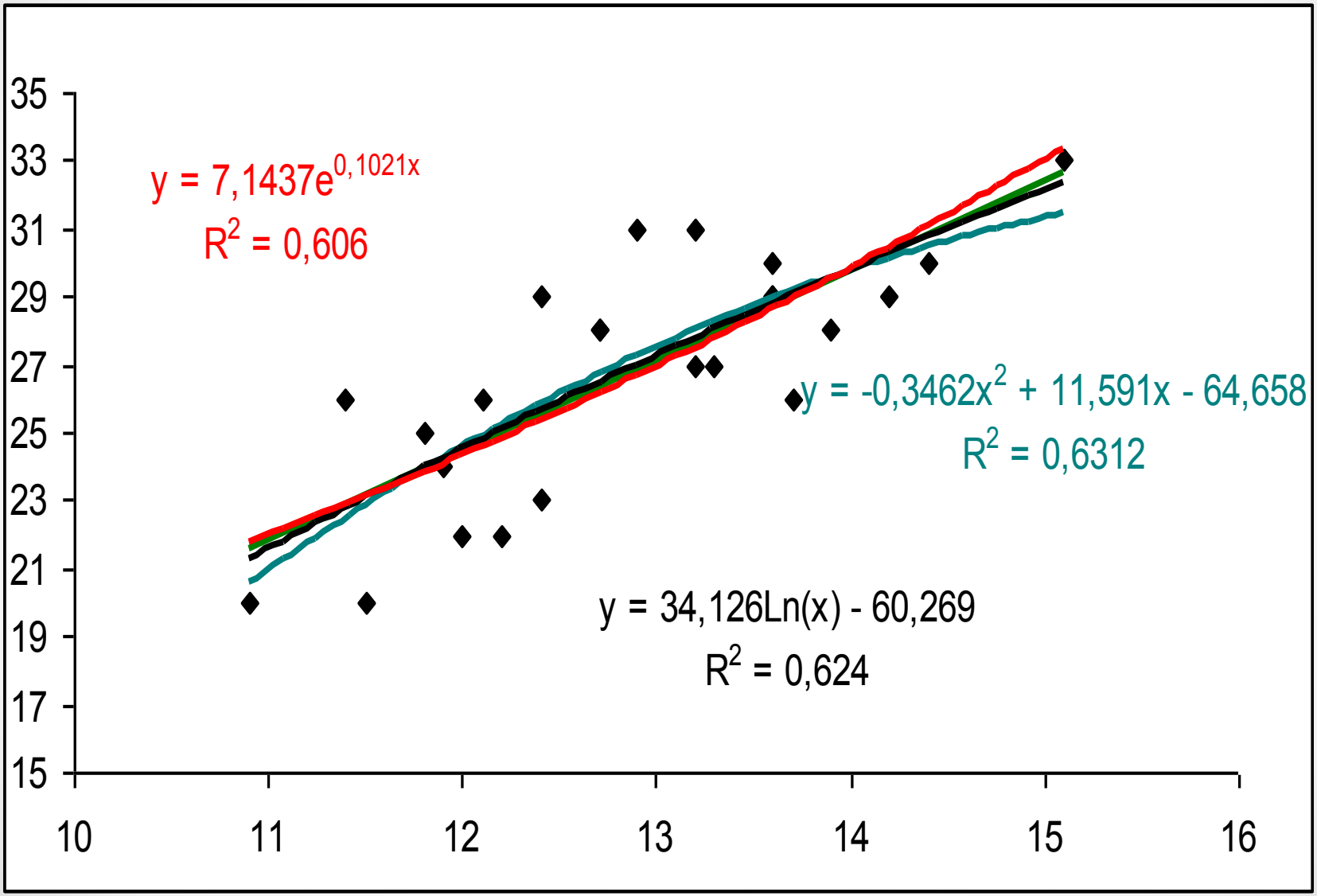


Brak stabilnej wariancji reszt



Inne funkcje regresji

Nazwa funkcji	Wzór	Liniowe przekształcenie
Liniowa	$\hat{y} = a + b \cdot x$	$\hat{y} = a + b \cdot x$
Logarytmiczna	$\hat{y} = a + b \cdot \ln x$	$\hat{y} = a + b \cdot \ln x$
Pierwiastkowa	$\hat{y} = a + b \cdot \sqrt{x}$	$\hat{y} = a + b \cdot \sqrt{x}$
Potęgowa	$\hat{y} = a \cdot x^b$	$\ln \hat{y} = \ln a + b \cdot \ln x$
Wykładnicza	$\hat{y} = e^{a+b \cdot x}$	$\ln \hat{y} = a + b \cdot x$
Typu S	$\hat{y} = e^{a+\frac{b}{x}}$	$\ln \hat{y} = a + b \cdot \frac{1}{x}$
Hiperboliczna	$\hat{y} = \frac{1}{a + b \cdot x}$	$\frac{1}{\hat{y}} = a + b \cdot x$


y = 7,1437e^0,1021x
R^2 = 0,606
y = -0,3462x^2 + 11,591x - 64,658
R^2 = 0,6312
y = 34,126Ln(x) - 60,269
R^2 = 0,624
35
33
31
29
27
25
23
21
19
17
15
10
11
12
13
14
15
16

Gulo gulo