AKADEMIA GÓRNICZO-HUTNICZA
IM. STANISŁAWA STASZICA W KRAKOWIE

Regresja i korelacja

Statystyka

Dr inż. Janusz Majewski
Katedra Informatyki

Regresja i korelacja

Dla każdej jednostki (obiektu) mamy wartości dwóch zmiennych losowych: x i y. Chcemy zbadać związek między tymi dwoma zmiennymi dla:

- -uzyskania liczbowych miar pewnych podstawowych cech zależności,
- -dostarczenia możliwości prognozowania (predykcji) wartości jednej ze zmiennych, gdy druga jest znana,
- - stwierdzenia, czy obserwowany kierunek trendu jest istotny.

Mówimy o związku między dwiema zmiennymi, gdy rozkład jednej zmiennej związany jest z wartościami drugiej. Nie znaczy to, że jedna zmienna jest przyczyną drugiej, nie mówimy więc o związku przyczynowo-skutkowym (por. liczba rozwodów versus produkcja papierosów)

Regresja prostoliniowa

Obserwowano zmienne x i y dla dużej liczby obiektów. Interesuje nas, jakiej przeciętnej zmianie ulega y gdy x przyjmuje różne wartości.

Zależność $E(y|x)$ od x nazywamy funkcją regresji. Mówimy o regresji prostoliniowej, jeżeli zmienna y przyjmuje rozkład normalny ze średnią

$$E(y|x) = A + Bx$$

oraz stałą (niezależną od x) wariancją równą σ^2.

Mamy n par obserwacji (x_i, y_i). Należy znaleźć liczby a i b będące estymatorami A i B tak, aby zminimalizować sumę kwadratów

$$\sum_i (y_i - Y_i)^2$$

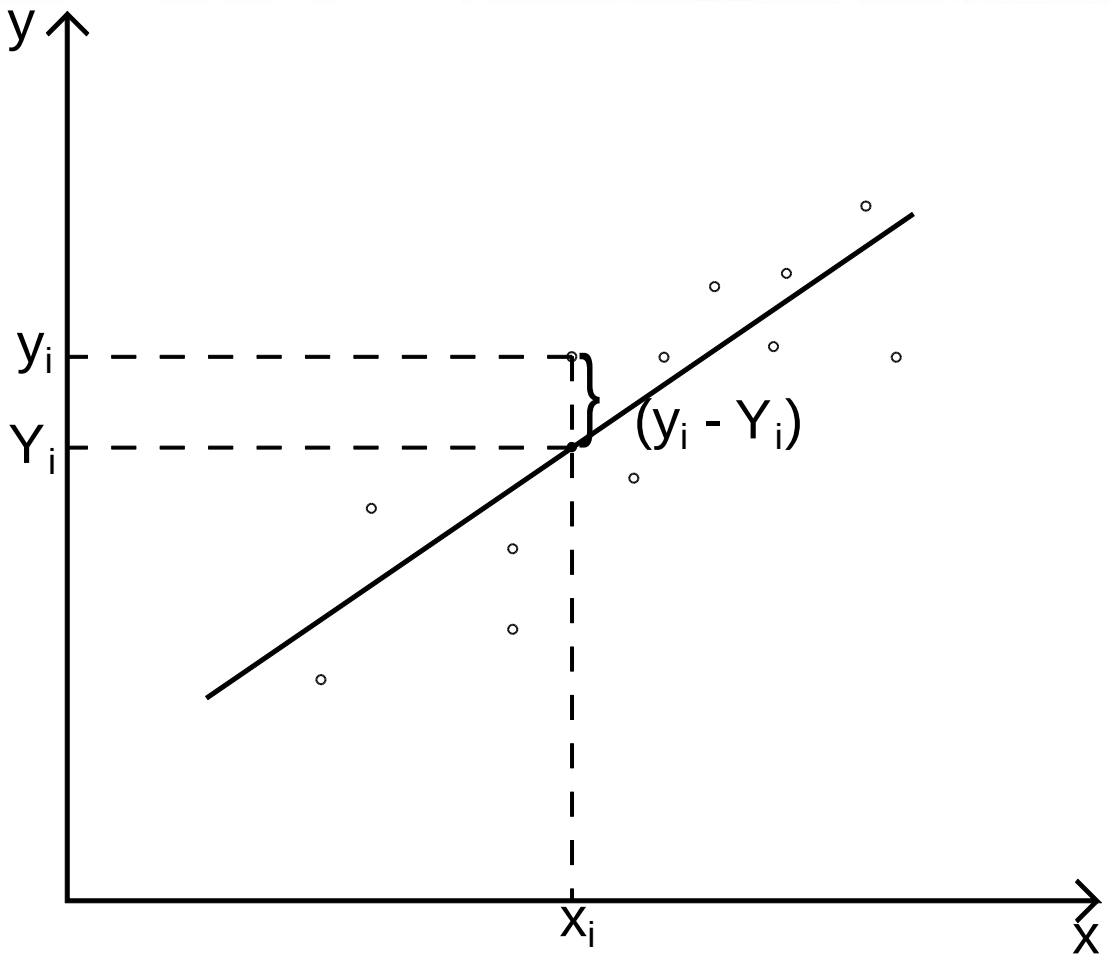
gdzie Y_i są wartościami wyznaczonymi przez szacowane równanie regresji:

$$Y_i = a + b \cdot x_i$$

Regresja prostoliniowa

y_i- wartość obserwowana

Y_i- wartość teoretyczna



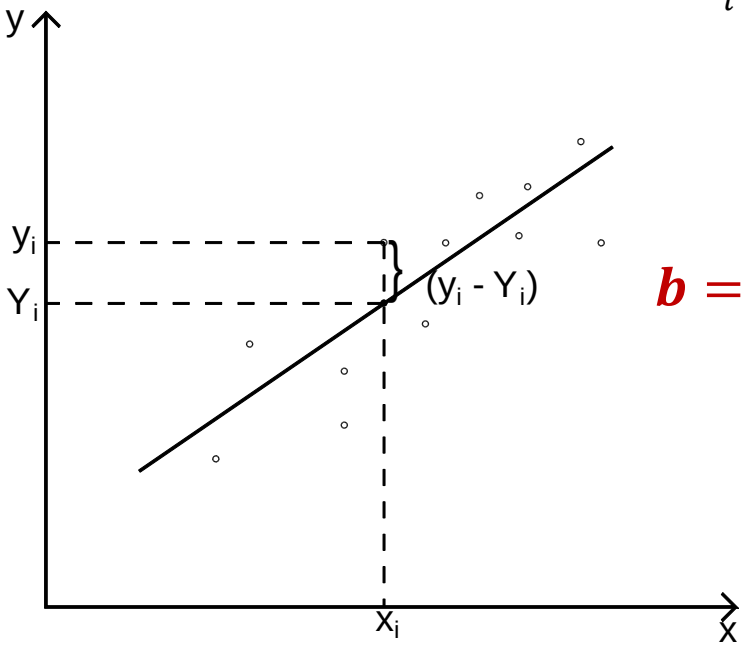
Regresja prostoliniowa

Mamy n par obserwacji (x_i, y_i). Należy znaleźć liczby a i b będące estymatorami A i B tak, aby zminimalizować sumę kwadratów

$$\sum_i (y_i - Y_i)^2$$

gdzie Y_i są wartościami wyznaczonymi przez szacowane równanie regresji:

$$Y_i = a + b \cdot x_i$$



y_i- wartość obserwowana
Y_i- wartość teoretyczna

$$\boldsymbol{b = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = \frac{\sum_i x_i y_i - n\bar{x}\bar{y}}{\sum_i x_i^2 - n\bar{x}^2}}$$

$$\boldsymbol{a = \bar{y} - b\bar{x}}$$

Regresja prostoliniowa

Pojęcie regresji związane jest z kształtem zależności $E(y|x)$ od x, pojęcie korelacji związane jest zaś z siłą tej zależności

Współczynnik korelacji

$$\rho = \frac{cov(x,y)}{\sigma_x \sigma_y}$$

$cov(x,y)$ - kowariancja zmiennych x i y

$$cov(x,y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} [x - E(x)]\,[y - E(Y)]\, f_{xy}(x,y)\, dx\, dy$$

σ_x, σ_y - odchylenie standardowe zmiennych x i y

Regresja prostoliniowa

Współczynnik korelacji

$$\rho = \frac{cov(x, y)}{\sigma_x \sigma_y}$$

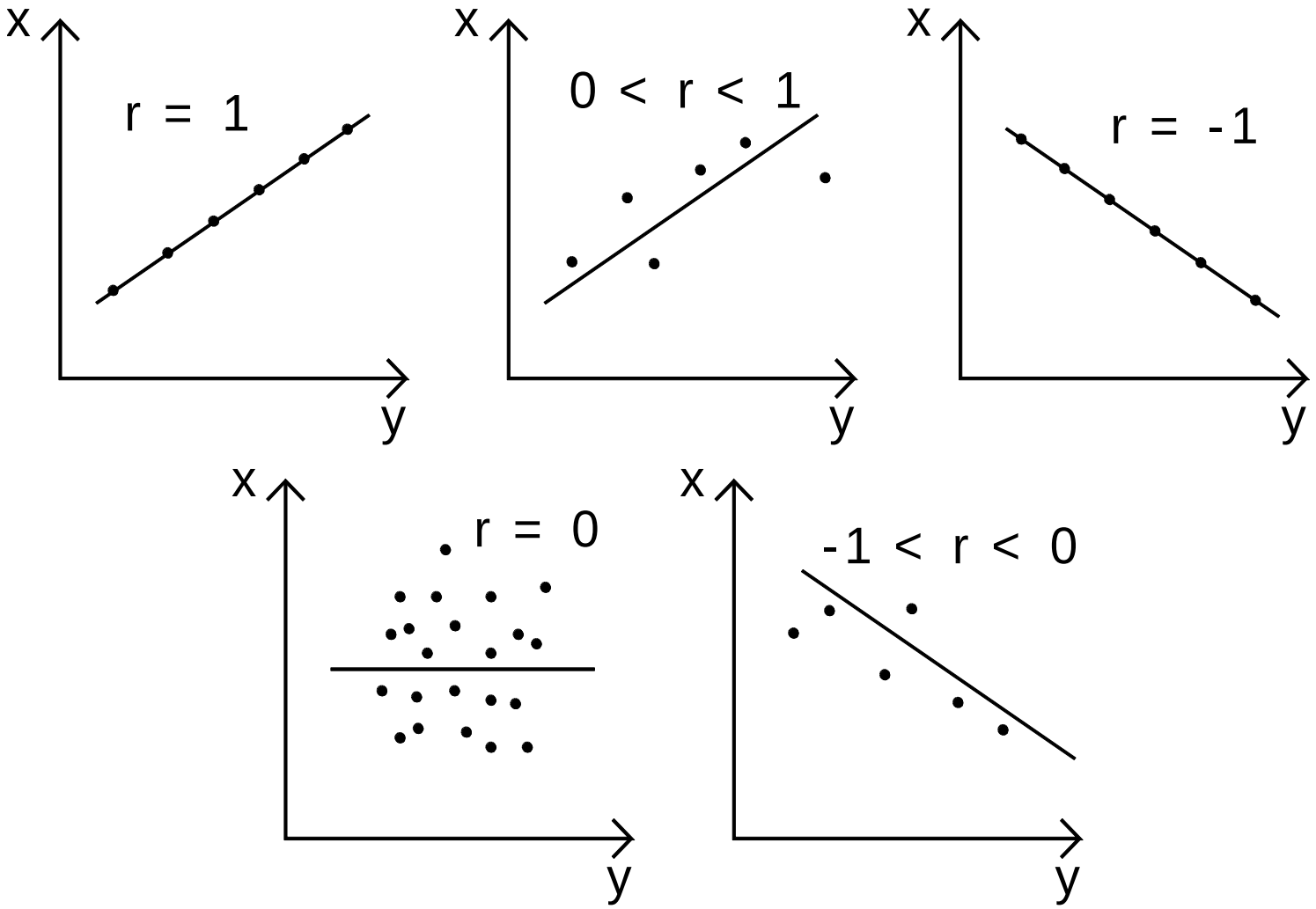
Estymatorem ρ jest współczynnik korelacji Pearsona r

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

Poza tym $b = r \frac{S_y}{S_x}$

S_x, S_y-estymatory odchyleń standardowych

Regresja prostoliniowa



Testowanie parametrów regresji i współczynnika korelacji

$$S^2(b) = \frac{S_o^2}{\sum_i (x_i - \bar{x})^2}$$

$$S^2(a) = S_o^2 \left[\frac{1}{n} + \frac{\bar{x}}{\sum_i (x_i - \bar{x})^2}\right]$$

$$S_o^2 = \frac{\sum_i (y_i - Y_i)^2}{n-2} = \frac{\sum_i (y_i - \bar{y})^2 (1-r)^2}{n-2}$$

$H_o: A = 0$ $H_1: A \neq 0$ $t = \frac{a}{\sqrt{S^2(a)}}$ Odrzucamy H_o, gdy $\lvert t \rvert \geq {}_{\alpha}t_{(n-2)}$	$H_o: B = 0$ $H_1: B \neq 0$ $t = \frac{b}{\sqrt{S^2(b)}}$ Odrzucamy H_o, gdy $\lvert t \rvert \geq {}_{\alpha}t_{(n-2)}$	$H_o: \rho = 0$ $H_1: \rho \neq 0$ $t = \frac{r}{\sqrt{S^2(r)}}$ Odrzucamy H_o, gdy $\lvert t \rvert \geq {}_{\alpha}t_{(n-2)}$

$$S^2(r) = \frac{1 - r^2}{n-2}$$

Nie wolno wykorzystać tak obliczonego $S^2(r)$ dla estymacji przedziałowej ρ !

Zadanie predykcji

Dana jest pewna wartość x_o. Chodzi o znalezienie przewidywanej wartości Y zmiennej y odpowiadającej danemu x_o. Najlepszym oszacowaniem wartości przewidywanej jest wartość Y_o. wynikająca z prostej regresji, czyli

$$Y_o = a + b \cdot x_o$$

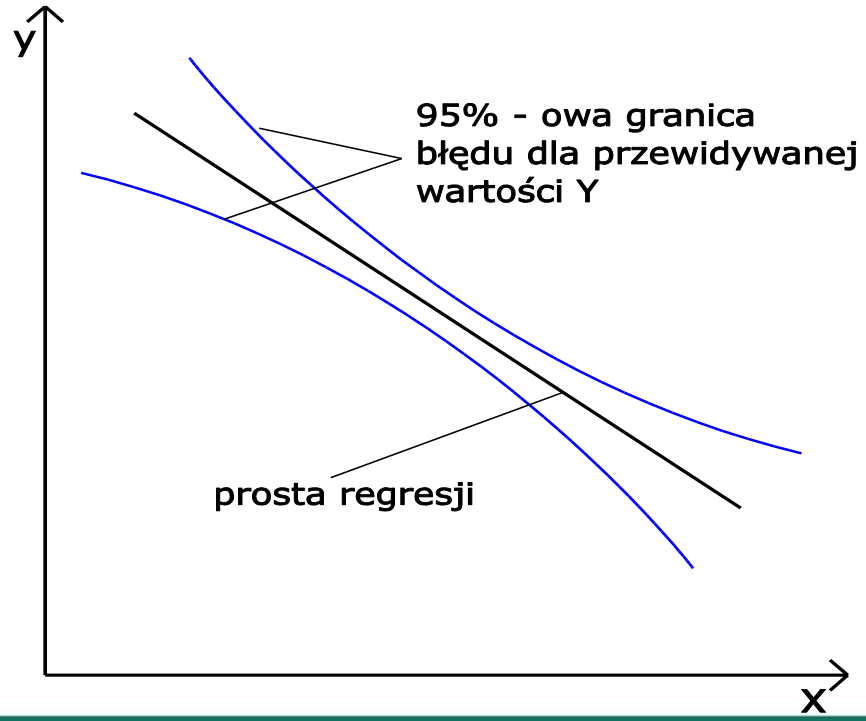
Granice błędu dla Y_o związane z losowością samego y oraz niedokładnością określenia parametrów prostej regresji wyrażają się wzorem:

$$Y_o \pm {}_{\alpha}t_{(n-2)} \cdot \sqrt{S_o^2} \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_o - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}}$$

Zadanie predykcji

Granice błędu dla Y_o związane z losowością samego y oraz niedokładnością określenia parametrów prostej regresji wyrażają się wzorem:

$$Y_o \pm {}_{\alpha}t_{(n-2)} \cdot \sqrt{S_o^2} \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_o - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}}$$



Analiza regresji - przykład

Przykład: Przez dwa tygodnie karmiono szczury dietą ubogą w witaminę D dla wywołania krzywicy. Następnie przez dalsze dwa tygodnie podawano szczurom preparat zawierający witaminę D. Po upływie tego czasu określono stopień wyleczenia przez radiografię prawego kolana każdego zwierzęcia doświadczalnego. Porównano analizowane zdjęcie radiologiczne ze standardowym zestawem fotografii opatrzonych numerami od 0 do 12 (stopniowanie w kierunku wzrastającego wyleczenia). Każdą dawkę preparatu podawano grupie kilku szczurów i późniejsze zdjęcia analizowało kilku radiologów. Zbadać regresję między **logarytmem** dawki (x_i) preparatu a średnim efektem dla każdej dawki (y_i)

dawka	2.5	5	10	20	40	80	160	320
x_i	0,398	0,699	1,000	1,301	1,602	1,903	2,204	2,505
y_i	0,250	1,0833	1,6667	2,8333	3,5833	4,3333	4,9167	5,5555

Analiza regresji - przykład

dawka	2.5	5	10	20	40	80	160	320
x_i	0,398	0,699	1,000	1,301	1,602	1,903	2,204	2,505
y_i	0,250	1,0833	1,6667	2,8333	3,5833	4,3333	4,9167	5,5555

$\bar{x} = 1{,}4515$ $\quad \bar{y} = 3{,}00$ $\quad n = 8$

$b = 2{,}5115$ $\quad a = -0{,}6454$ $\quad r = 0.9943$

Przykładowo przeprowadzimy test dla r:

$H_o: \rho = 0$

$H_o: \rho \neq 0$

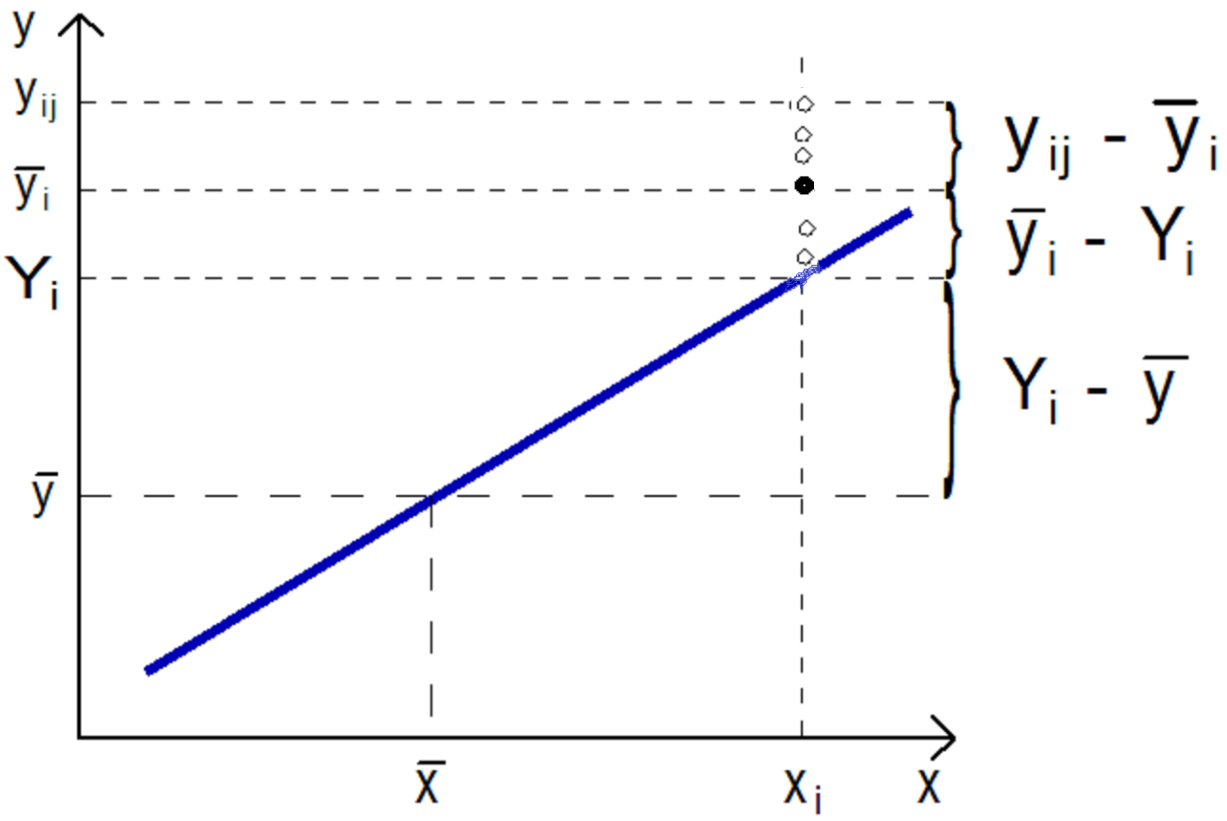
$$t = r\sqrt{\frac{n-2}{1-r^2}} = 22{,}8296$$

$t > t_{kryt}$

$_{0{,}05}t_{(6)} = 2{,}447$

H_o – odrzucamy przyjmując $\rho \neq 0$

Test na liniowość



$$\sum_{ij}(y_{ij} - \bar{y})^2 = \sum_{ij}(y_{ij} - y_i)^2 + \sum_{ij}(y_i - Y_i)^2 + \sum_{ij}(Y_i - \bar{y})^2$$

Test na liniowość

x_i	x_1	x_2	…	x_i	…	x_k	
y_{ij}	y_{11} y_{12} … y_{1n_1}	y_{21} y_{22} … y_{2n_2}		y_{i1} y_{i2} … y_{in_i}		y_{k1} y_{k2} … y_{kn_k}	
	T_1	T_2	…	$T_i = \sum_{j=1}^{n_i} y_{ij}$	…	T_k	$T = \sum_{i=1}^{k} T_k$
	s_1	s_2	…	$s_i = \sum_{j=1}^{n_i} y_{ij}^2$	…	s_k	$S = \sum_{i=1}^{k} s_i$
	$\overline{y_1}$	$\overline{y_2}$	…	$\bar{y}_i = \frac{T_i}{n_i}$	…	$\overline{y_k}$	$\bar{y} = \frac{T}{N}$
	n_1	n_2	…	n_i	…	n_k	$N = \sum_{i=1}^{k} n_i$

Test na liniowość

$$\sum (y_{ij} - \bar{y})^2 = \sum (y_{ij} - \bar{y}_i)^2 + \sum (\bar{y}_i - Y_i)^2 + \sum (Y_i - \bar{y})^2$$

SK	=	**SKR**	+	**SKL**	+	**SKB**
całkowita suma kwadratów odchyleń wartości obserwacji od średniej		suma kwadratów odchyleń wartości obserwacji od średniej serii (zmienność resztowa)		suma kwadratów odchyleń średnich serii od wartości teoretycznych wynikających z prostej regresji (świadczy o dopasowaniu obserwacji do prostej regresji)		suma kwadratów odchyleń wartości teoretycznych od średniej (świadczy o nachyleniu prostej regresji- por. współczynnik B)

Test na liniowość

$$SKR = S - \sum_{i=1}^{k} \frac{T_1^2}{n_i}$$

$$SK = S - \frac{T^2}{N}$$

$$SKB = \frac{\sum_{i=i}^{k} x_i T_i - \frac{T_i \sum_{i=1}^{k} n_i x_i}{N}}{\sum_{i=1}^{k} n_i x_i^2 - \frac{(\sum_{i=1}^{k} n_i x_i)^2}{N}}$$

$$SKL = SK - SKR - SKB$$

Test na liniowość

Źródło zmienności	Suma kwadratów	Liczba st. swobody	Średni kwadrat	Stosunek wariancji
Odchylenie wartości teoret. od średniej	*SKB*	*1*	$S_1^2 = \frac{SKB}{1}$	$F_1 = \frac{S_1^2}{S_o^2}$
Odchylenie średnich serii od prostej regr.	*SKL*	*k-2*	$S_2^2 = \frac{SKL}{k-2}$	$F_2 = \frac{S_2^2}{S_o^2}$
Reszta wewnątrz serii	*SKR*	*N-k*	$S_o^2 = \frac{SKR}{N-k}$	
Ogółem	*SK*	*N-1*		

Testowane hipotezy:

1) $H_o: B = 0$
$H_1: B \neq 0$
(stosunek F_1)

2) H_o: Funkcja regresji jest liniowa
H_1: Funkcja regresji nie jest liniowa
(stosunek F_2)

Test na liniowość

Przykład: W poprzednim przykładzie dopasowaliśmy linię regresji dla zależności wzajemnej logarytmu dawki leku (x_i) i średniego efektu terapeutycznego (y_i). Tutaj zamiast efektu średniego uwzględniamy, że każdą dawkę leku (x_i) podawano sześciu szczurom. Wobec tego dysponujemy sześcioma ocenami efektu leczenia dla każdego x_i

dawka	2,5	5	10	20	40
x_i	0,398	0,699	1,000	1,301	1,602
y_{ij}	0	1,0	1,5	3,0	6,5
	0	1,5	1,0	3,0	3,5
	0	1,5	2,0	5,5	4,5
	0	1,0	3,5	2,5	3,5
	0	1,0	2,0	1,0	3,5
	0,5	0,5	0	2,0	3,0

Test na liniowość

Źródło zmienności	Suma kwadratów	Liczba st. swobody	Średni kwadrat	Stosunek wariancji	Istotność
Odchyl. wart. teoretycznych od średniej	57,0375	1	57,0375	52,248	**P<0,005**
Odchyl. śr. efektu dawki od prostej	0,8458	3	0,2819	0,258	nieistotne
Reszta wewnątrz dawki	27,2917	25	1,0971		
Ogółem	85,1750	29			