POLITECHNIKA OPOLSKA
WYDZIAŁ MECHANICZNY
Katedra Technologii Maszyn i Automatyzacji Produkcji

Laboratorium Podstaw Inżynierii Jakości

Ćwiczenie nr **4**

Temat:

Analiza korelacji i regresji dwóch zmiennych

Zakres ćwiczenia

Na linii produkcyjnej w warunkach produkcji wielkoseryjnej produkowane są przedmioty. Zachodzi podejrzenie, że występuje nadmierne zużywanie się ostrza narzędzia w funkcji czasu powodujące wyraźne zwiększanie się wymiaru w kolejnych przedmiotach schodzących z linii produkcyjnej. Należy sprawdzić prawdziwość tego przypuszczenia metodą analizy korelacji i regresji zmiennych. Należy w tym celu:

1. Wykonać pomiary wymiarów:
   a) 12,7k7 – szerokość płytki skrawającej,
   b) 75js9 – długość elementu ustawczego,
   c) 64h5 – szerokość elementu ustawczego,
   d) Ø30ZA8 – średnica wewnętrzna pierścienia,
   e) Ø44e10 – średnica zewnętrzna pierścienia,

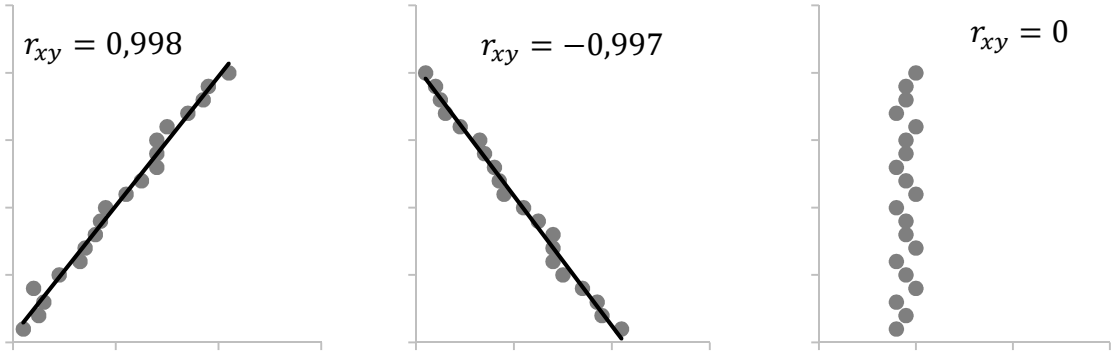
   50 szt. wyrobów i utrwalić wyniki pomiarów w arkuszu kalkulacyjnym.
2. Opracować procedurę obliczeniową oraz wyznaczyć współczynnik korelacji r_{xy}z próby między numerem kolejnego przedmiotu x a jego wymiarem x_i.
3. Określić zależność korelacyjną między badanymi zmiennymi.
4. Wykonać test istotności współczynnika korelacji (sprawdzić hipotezę, że zmienne x i y są skolerowane).
5. Wyznaczyć wartości współczynników a i b równania regresji liniowej $y = ax + b$.
6. Przy współczynniku ufności 0,95 oszacować metodą przedziałową liniową funkcję regresji.
7. Wykonać wykres badanej zależności (regresji) wraz z krzywymi ufności.
8. Wykonać analizę wyników obliczeń.
9. Sformułować wnioski.

I. PODSTAWY TEORETYCZNE

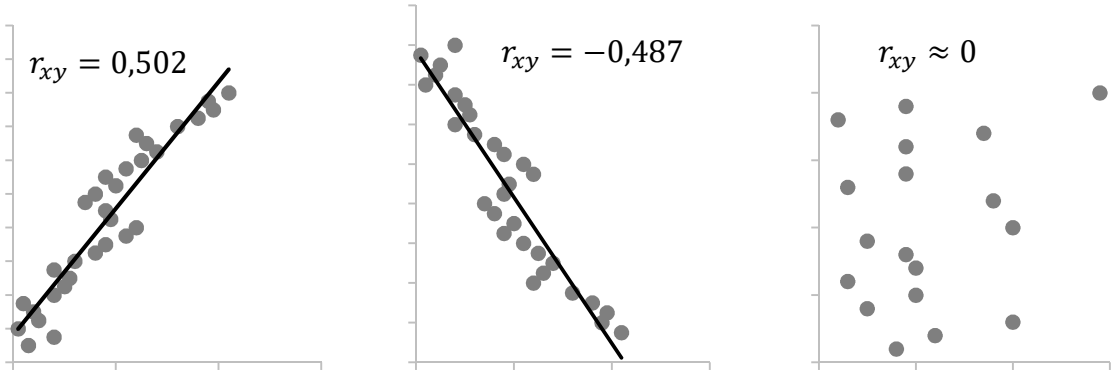
Przy badaniu populacji generalnej równocześnie ze względu na dwie lub więcej cech mierzalnych posługujemy się pojęciami regresji i korelacji. Oba te pojęcia dotyczą zależności miedzy zmiennymi, przy czym korelacja zajmuje się siłą tej zależności, a regresja - jej kształtem, określa rodzaj zależności miedzy cechami (liniowa, krzywoliniowa).
Generalnie, po ustaleniu, że między badanymi cechami istnieje niezbyt słaba (istotna) korelacja, przystępuje się do znalezienia funkcji regresji, która opisując matematycznie związek pomiędzy zmiennymi, pozwala na przewidywanie wartości jednej cechy przy założeniu, że druga cecha przyjęła określoną wartość. Analiza korelacji i analiza regresji są stosowane, gdy konieczne jest zbadanie zależności między dwiema zmiennymi, np. potwierdzenie (lub odrzucenie) zależności przyczynowo-skutkowych wykorzystywanych w sterowaniu jakością.

Diagram korelacji w literaturze przedmiotu występuje także pod innymi nazwami: wykres rozrzutu, wykres zmiennych, wykres korelacji.
Wykresy korelacji są uproszczoną formą graficznej ilustracji związku zachodzącego pomiędzy dwiema zmiennymi (rys. 1.).

a) silna korelacja dodatnia b) silna korelacja ujemna c) korelacja liniowa



d) korelacja dodatnia e) korelacja ujemna f) brak korelacji



Rys. 1. Przykłady zmiennych mocno skorelowanych a) i b), słabo skorelowanych d) i e) oraz nieskorelowanych c) i f).

W celu sporządzenia diagramu korelacyjnego dane przedstawia się we współrzędnych prostokątnych reprezentujących obie zmienne poprzez nanoszenie na wykres punktów P_i, których położenie na wykresie wynika z wartości poszczególnych par wyników (x_i, y_i). Aby uzyskać wiarygodny obraz badanej zależności, wskazane jest przeanalizowanie stosunkowo dużej liczby n par danych (więcej niż $n = 30$).

Sposób grupowania się punktów na wykresie uwidacznia zależności korelacyjne między badanymi zmiennymi:

- jeżeli punkty układają się w pobliżu pewnej krzywej, oznacza to, że pomiędzy badanymi zmiennymi zachodzi znacząca korelacja. Krzywą tę można aproksymować odpowiednią funkcją regresji ze wskazaniem siły związku między tymi wielkościami w postaci współczynnika korelacji. W najprostszym przypadku, punkty układają się w pobliżu prostej o dodatnim lub ujemnym współczynniku pochylenia (regresji). Mówi się wówczas o zachodzeniu pomiędzy nimi odpowiednio korelacji dodatniej lub ujemnej (rysunki 1a i 1b), zmienne są skorelowane.
- jeżeli punkty są na wykresie rozproszone lub ułożone wzdłuż prostej prostopadłej do jednej z osi układu współrzędnych, oznacza to, że badane wielkości nie są skorelowane, są od siebie niezależne (rysunki 1c i 1f).
- jeżeli punkty wykazują pewne skupienie i tworzą „chmurę" rozciągającą się wzdłuż pewnej krzywej (rysunki 1d i 1e), można mówić o istnieniu lub nie istnieniu korelacji zmiennych dopiero po wykonaniu niezbędnych obliczeń, tj. wartości współczynnika korelacji z próby r_{xy}, oraz wartości statystyki t testu istotności hipotezy, że zmienne x i y nie są skorelowane, wobec hipotezy alternatywnej.

Należy podkreślić, iż współczynnik korelacji $r_{x,y} = +1$ lub $r_{x,y} = -1$ (ścisła zależność) nie oznacza, że pomiędzy zmiennymi musi zachodzić związek przyczynowo-skutkowy, np. wartość strzałki ugięcia f belki pod działaniem siły F. Najczęściej rzeczywiście tak jest, ale w każdym przypadku powinno to być wykazane dodatkową analizą fizyczną istoty zależności. Jeżeli $r_{x,y} = \pm 1$, a związek przyczynowo-skutkowy nie występuje, oznacza to, ze związek statystyczny staje się związkiem funkcyjnym.

Współczynnik korelacji ma następującą interpretację:
$r = 0 \rightarrow$ nie ma korelacji, czyli nie ma liniowego związku między dwiema zmiennymi losowymi,
$r = 1 \rightarrow$ zachodzi ścisły dodatni związek między dwiema zmiennymi. Gdy jedna z tych zmiennych przyjmuje większe wartości, druga także przyjmuje większe wartości (i na odwrót),
$r = -1 \rightarrow$ zachodzi ścisły ujemny związek między dwiema zmiennymi. Gdy jedna z tych zmiennych przyjmuje większe wartości, to i druga przyjmuje mniejsze wartości (i na odwrót),
r znajduje się w przedziale $(-1,1) \rightarrow$ wartość współczynnika korelacji jest miarą siły liniowego związku między dwiema zmiennymi.

Klasyfikacja zależności korelacyjnej
$|r| = 0$ - brak korelacji
$0{,}0 < |r| \leq 0{,}1$ - korelacja nikła
$0{,}1 < |r| \leq 0{,}3$ - korelacja słaba
$0{,}3 < |r| \leq 0{,}5$ - korelacja przeciętna
$0{,}5 < |r| \leq 0{,}7$ - korelacja wysoka
$0{,}7 < |r| \leq 0{,}9$ - korelacja bardzo wysoka
$0{,}9 < |r| < 1{,}0$ - korelacja niemal pełna
$|r| = 1$ - korelacja pełna
Korelację pełną można nazwać również zależnością funkcyjną, co oznacza, że pomiędzy *x* i *y* istnieje funkcja, która odwzorowuje *x* w *y* bez występowania jakiejkolwiek reszty, błędu.

Można również spotkać się z następującą klasyfikacją zależności korelacyjnej:
$0{,}0 \leq |r| \leq 0{,}2$ - brak korelacji (brak związku liniowego)
$0{,}2 < |r| \leq 0{,}4$ - korelacja słaba
$0{,}4 < |r| \leq 0{,}7$ - korelacja (umiarkowana) średnia
$0{,}7 < |r| \leq 0{,}9$ - korelacja silna
$0{,}9 < |r| \leq 1{,}0$ - korelacja bardzo silna
Należy pamiętać, że sama interpretacja siły związku jest mniej ważna niż informacja czy dana zależność jest istotna statystycznie. Jeżeli nie jest to oceniamy, że według statystyki (przyjętego poziomu istotności) uzyskana wartość jest dziełem błędu niż prawdziwej zależności. Jeżeli przyjmiemy jedną czy drugą klasyfikację nie popełnimy błędu - jeżeli posługujemy się daną skalą siły korelacji należy na wstępie zaznaczyć, z jakiej korzystamy w przedstawieniu wyników.

Obliczenie współczynnika korelacji $r_{x,y}$ z próby

Estymatorem nieobciążonym i zgodnym współczynnika korelacji miedzy dwiema badanymi cechami x i y w populacji jest współczynnik korelacji z próby (z eksperymentu), zwykle oznaczany symbolem $r_{x,y}$ i obliczany z n par (x_i, y_i) wyników próby według wzoru:

$$r_{xy} = \frac{\sum_{i=1}^{n}[(x_i - \bar{x}) * (y_i - \bar{y})]}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 * \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{\frac{1}{n}\sum x_i y_i - \bar{x}\bar{y}}{\sigma_x \sigma_y} \quad (1)$$

r_{xy} – współczynnik korelacji pomiędzy x i y
$x_i, \bar{x}$ – wartości rozpatrywane i średnia arytmetyczna zmiennej niezależnej,
$y_i, \bar{y}$ - wartości rozpatrywane i średnia arytmetyczna zmiennej zależnej,
σ_y, σ_x – odchylenia standardowe zmiennych x i y
n –ilość obserwacji

gdy $r = +1$ lub $r = -1$ istnieje ścisła zależność w postaci funkcji liniowej $y = ax + b$, gdy $r = 0$ zmienne są nieskorelowane – są niezależne, nie istnieje funkcja liniowa zależności im $|r|$ jest bliższa 1, tym korelacja jest mocniejsza.

Na podstawie wyników tej próby należy sprawdzić hipotezę, że zmienne x i y nie są skorelowane, tzn. hipotezę $H_0 : \rho = 0$, wobec hipotezy alternatywnej $H_1 : \rho \neq 0$.
Test istotności dla tej hipotezy jest następujący:
obliczamy wartość współczynnika korelacji r z próby oraz wartość statystyki

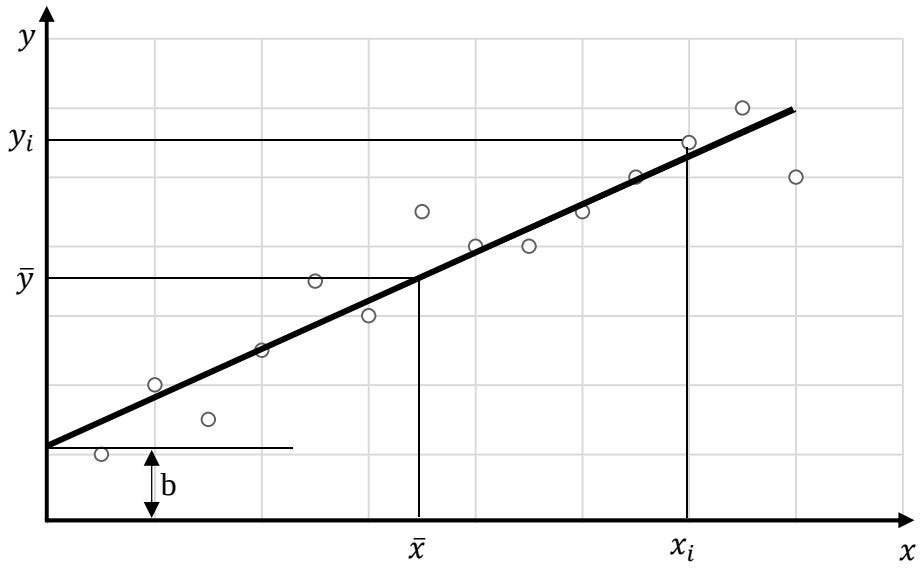
$$t = \frac{r}{\sqrt{1 - r^2}} * \sqrt{n - 2} \tag{2}$$

Statystyka ta ma przy założeniu prawdziwości hipotezy H_0 rozkład *t*-Studenta z $n - 2$ stopniami swobody. Z tablicy rozkładu *t*-Studenta dla ustalonego z góry poziomu istotności α i dla $k = n - 2$ stopni swobody odczytujemy wartość krytyczną $t_{\alpha,k}$ tak by $P\{|t| \geq t_{\alpha,k}\} = \alpha$.
Jeżeli z porównania obliczonej wartości t z wartością krytyczną $t_{\alpha,k}$ otrzymamy nierówność $|t| \geq t_{\alpha,k}$ to hipotezę H_0 o braku korelacji między zmiennymi trzeba odrzucić (czyli, istnieje korelacja/zależność między zmiennymi na przyjętym poziomie istotności α).
Gdy, natomiast $|t| < t_{\alpha,k}$, to nie ma podstaw do odrzucenia hipotezy H_0, że zmienne x i y są nieskorelowane - niezależne (czyli, brak korelacji/zależności między zmiennymi, na przyjętym poziomie istotności α).

Wyznaczenie funkcji regresji liniowej

Analiza regresji liniowej, zwana również regresją prostą określa sposób przyporządkowania jednej zmiennej losowej (zmiennej zależnej y) wartości innej zmiennej (zmiennej niezależnej x), za pomocą funkcji matematycznej i odpowiedniego wykresu (rys. 2). Może to mieć duże znaczenie w przewidywaniu wzajemnego zachowania się obu parametrów.

O ile współczynnik korelacji liniowej mówi nam jak bardzo dane są od siebie zależne o tyle regresja liniowa mówi nam jak bardzo zmieni się y gdy zmienimy x.



Rys. 2. Wyznaczenie linie regresji

Wyznaczenie funkcji regresji liniowej (rys. 2) polega na wyznaczeniu współczynników b i a linii prostej:

$$y = ax + b \tag{3}$$

gdzie:

$$a = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) * (y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{cov(x,y)}{\sigma_x^2} = r_{xy} * \frac{\sigma_y}{\sigma_x} \tag{4}$$

$$b = \bar{y} - a\bar{x} \tag{5}$$

a - estymator współczynnika regresji y względem x, współczynnik nachylenia prostej regresji,
b – stała
$x_i, \bar{x}$ – wartości rozpatrywane i średnia arytmetyczna zmiennej niezależnej,
$y_i, \bar{y}$ – wartości rozpatrywane i średnia arytmetyczna dla zmiennej zależnej.
$cov(x,y)$ – kowariancja zmiennych x i y
σ_y, σ_x – odchylenia standardowe zmiennych x i y
r_{xy} – współczynnik korelacji pomiędzy x i y

Analiza regresji pozwala na:

- ustalenie istotności związku między wielkościami,
- ustalenie wpływu parametrów procesu na cechy wyrobu,
- ustalenie wpływu cech charakteryzujących jakość typu i jakość wykonania wyrobu na jego parametry użytkowe.

Należy dodać, iż związki pomiędzy analizowanymi zmiennymi mogą mieć także charakter nieliniowy np. paraboliczny - krzywoliniowa funkcja regresji.

Wyznaczenie obszaru ufności

Korzystając z założenia o normalności rozkładu, można zbudować tzw. obszar ufności pomiędzy krzywymi ufności dla prostej regresji $y = \alpha x + \beta$ oraz przedział ufności dla współczynnika regresji a korzystając z wzorów matematycznych poniższych modeli.

MODEL 1
Oszacowanie parametrów liniowej regresji $y = \alpha x + \beta$ wraz z jej obszarem ufności

Dwuwymiarowy rozkład badanych dwóch cech mierzalnych x i y w populacji generalnej jest normalny lub zbliżony do normalnego. Z populacji tej wylosowano do próby n elementów i otrzymano dla tych cech wyniki (x_i, y_i) $(i = 1,2, \dots, n)$.
Na podstawie wyników próby należy oszacować parametry liniowej regresji $y = \alpha x + \beta$ wraz z jej obszarem ufności.

Metoda najmniejszych kwadratów daje następujące oszacowanie prostej regresji

$$\hat{y} = ax + b \tag{6}$$

Estymatory a i b są nieobciążonymi i zgodnymi estymatorami parametrów α i β. Obszar ufności dla prostej regresji $y = \alpha x + \beta$ ograniczony tzw. krzywymi ufności, wyznacza się według wzoru:

$$P\{\hat{y}_i - t_\gamma s_{y_i} < \tilde{y}_i < \hat{y}_i + t_\gamma s_{y_i}\} = 1 - \gamma \tag{7}$$

$\hat{y}_i$ – oznacza wartość funkcji $y = ax + b$ wyznaczonej według wzorów (4) i (5),
$\tilde{y}_i$ – oznacza wartość szacowanej funkcji regresji $y = \alpha x + \beta$,
t_γ – jest wartością zmiennej o rozkładzie *t*-Studenta, wyznaczoną z tablicy tego rozkładu dla ustalonego z góry współczynnika ufności $1 - \gamma$ dla $k = n - 2$ stopni swobody

$$S_{\hat{y}_i} = S_r \sqrt{\frac{1}{n} + \frac{\{x_i - \bar{x}\}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \tag{8}$$

gdzie S_r jest odchyleniem przeciętnym od prostej regresji, obliczanym ze wzoru:

$$S_r = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{9}$$

Określony w ten sposób obszar ufności z prawdopodobieństwem $1 - \gamma$ (niepewnością γ) pokrywa prawdziwą funkcję regresji regresji $y = \alpha x + \beta$ w populacji generalnej.

MODEL II

Oszacowanie parametrów liniowej regresji $y = \alpha x + \beta$ wraz z przedziałem ufności dla współczynnika regresji regresji α

Dwuwymiarowy rozkład badanych dwóch cech mierzalnych x i y w populacji generalnej jest normalny lub zbliżony do normalnego. Z populacji tej wylosowano do próby n elementów i otrzymano dla tych cech wyniki (x_i, y_i) $(i = 1,2, \dots, n)$.
Na podstawie wyników próby należy oszacować parametry liniowej regresji $y = \alpha x + \beta$ wraz z przedziałem ufności dla współczynnika regresji α.

Przedział ufności dla współczynnika regresji α funkcji regresji $y = \alpha x + \beta$ w populacji wyznacza się według wzoru:

$$P\left\{\alpha - t_\gamma \frac{S_r}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} < \alpha < \alpha + t_\gamma \frac{S_r}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}\right\} = 1 - \gamma \qquad (10)$$

gdzie:
α – jest współczynnikiem regresji uzyskanym metodą najmniejszych kwadratów dla funkcji $\hat{y} = ax + b$ wyznaczonym z próby wg wzoru (4),
S_r – jest odchyleniem przeciętnym od prostej regresji, obliczanym ze wzoru (9),
t_γ – jest wartością zmiennej o rozkładzie *t*-Studenta, wyznaczoną z tablicy tego rozkładu dla ustalonego z góry współczynnika ufności $1 - \gamma$ dla $k = n - 2$ stopni swobody.

Dla wyznaczenia przedziału ufności dla współczynnika regresji α trzeba znaleźć $\hat{y} = ax + b$, tj. oszacowanie całej liniowej funkcji regresji.

PRZYKŁAD

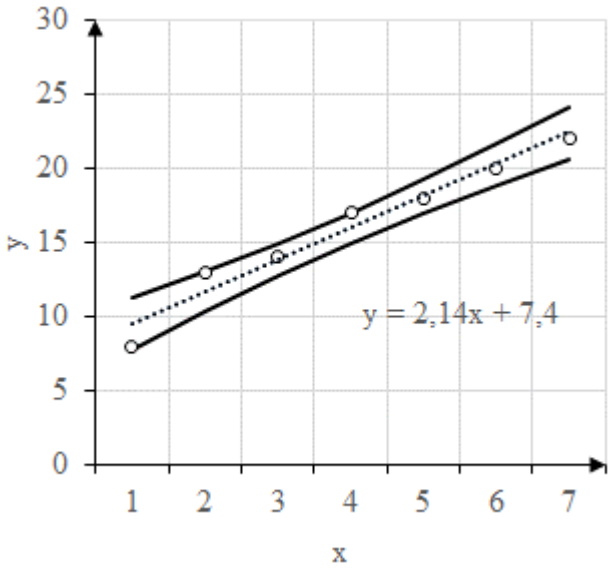
Badając zależność między wielkością x pewnego wyrobu a zużyciem y pewnego surowca zużywanego w produkcji tego wyrobu otrzymano dla losowej próby $n = 7$ obserwacji następujące wyniki (x_i w tys. sztuk, y_i w tonach).

Tabela 1. Zależność x wyrobu między zużycie surowca y

x_i	1	2	3	4	5	6	7
y_i	8	13	14	17	18	20	22

Należy przy współczynniku ufności 0,95 (95%) oszacować metodą przedziałową zarówno całą liniową funkcję regresji, jak i sam współczynnik regresji zużycia surowca względem wielkości produkcji.

Nanosząc otrzymane punkty empiryczne $P_i(x_i, y_i)$ na wykres (rys. 3) można stwierdzić, że badaną regresję można przyjąć za liniową.



Rys. 3. Liniowa funkcja regresji wraz z krzywymi ufności

Estymację liniowej funkcji regresji przeprowadzono według wzorów z MODELU I. Wartości estymatorów a i b wyznaczono metodą najmniejszych kwadratów stosując wzory (4) i (5). Odpowiednie obliczenia przeprowadzono tabelarycznie w programie Microsoft Excel.

Tabela 2. Obliczenia

	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$\hat{y}_i$	$(y_i - \hat{y}_i)^2$	$\frac{(x_i - \bar{x})^2}{\sum(x_i - \bar{x})^2}$	$\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum(x_i - \bar{x})^2}$	$\sqrt{\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$	$S_{\hat{y}_i}$	$t_\gamma S_{\hat{y}_i}$	$\hat{y}_i - t_\gamma S_{y_i}$	$\hat{y}_i + t_\gamma S_{y_i}$
	1	8	-3	-8	24	9	64	9,5	2,25	0,321	0,464	0,681	0,694	1,8	7,8	11,3
	2	13	-2	-3	6	4	9	11,7	1,69	0,143	0,286	0,535	0,545	1,4	10,3	13,1
	3	14	-1	-2	2	1	4	13,8	0,04	0,036	0,179	0,423	0,431	1,1	12,7	14,9
	4	17	0	1	0	0	1	16,0	1,00	0,000	0,143	0,378	0,385	1,0	15,0	17,0
	5	18	1	2	2	1	4	18,1	0,01	0,036	0,179	0,423	0,431	1,1	17,0	19,2
	6	20	2	4	8	4	16	20,2	0,04	0,143	0,286	0,535	0,545	1,4	18,8	21,6
	7	22	3	6	18	9	36	22,4	0,16	0,321	0,464	0,681	0,694	1,8	20,6	24,2
∑	28	112			60	28	134		5,19		0,143					

$$\bar{x} = \frac{28}{7} = 4$$

$$\bar{y} = \frac{112}{7} = 16$$

stąd

$$a = \frac{60}{28} = 2{,}14$$
$$b = 16 - (2{,}14 * 4) = 7{,}4$$

Otrzymano zatem oszacowanie prostej regresji:

$$\hat{y} = 2{,}14x + 7{,}4$$

Wartości $\hat{y}_i$ oraz kwadraty ich odchyleń empirycznych wartości y_i obliczono w tabeli 2.

$$S_r = \sqrt{\frac{1}{5} * 5{,}19} = \sqrt{1{,}038} = 1{,}02$$

Dla $k = n - 2$ i dla przyjętego współczynnika ufności $\gamma = 0{,}95$ otrzymano z tablic rozkładu t-Studenta wartość $t_\gamma = 2{,}571$. Ponadto $\frac{1}{n} = \frac{1}{7} = 0{,}143$.

Wartości $S_{\hat{y}_i}$ oraz rzędne punktów leżących na krzywych ufności przedstawiono w tabeli 2.
W ostatnich dwóch punktach tabeli otrzymano dla odpowiednich odciętych x_i rzędne punktów leżących na dolnej i górnej krzywej ufności. Krzywe te naniesiono na rys. 3.
Obszar między tymi krzywymi z prawdopodobieństwem 95% pokrywa nieznaną funkcję regresji $y = \alpha x + \beta$ w populacji generalnej. Przedział ufności dla współczynnika regresji α otrzymano ze wzoru podanego w MODELU II.

$$a = 2{,}14$$
$$t_\gamma = 2{,}571$$
$$S_r = 1{,}02$$
$$\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \sqrt{28} = 5{,}29$$

Przedział ufności dla współczynnika regresji α jest następujący:

$$P\left\{\alpha - t_\gamma \frac{S_r}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} < \alpha < \alpha + t_\gamma \frac{S_r}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}\right\} = 1 - \gamma$$

$$2{,}14 - 2{,}571 * \frac{1{,}02}{5{,}29} < \alpha < 2{,}14 + 2{,}571 * \frac{1{,}02}{5{,}29}$$

Czyli

$$2{,}14 - 0{,}50 < \alpha < 2{,}14 + 0{,}50$$
$$1{,}64 < \alpha < 2{,}64$$

Korelacja

$$r_{xy} = \frac{\sum_{i=1}^{n}[(x_i - \bar{x}) * (y_i - \bar{y})]}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 * \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{60}{\sqrt{28 * 134}}$$

$$r_{xy} = 0{,}979$$