Regresja linowa
metoda najmniejszych kwadratów

Tadeusz M. Molenda
Instytut Fizyki US

Wydział Matematyczno-Fizyczny
Uniwersytetu Szczecińskiego

Regresja liniowa

(też: ***metoda najmniejszych kwadratów, metoda wyrównawcza, metoda Gaussa***)

Zagadnienia

- istota metody
- postulat Gaussa
- współczynniki prostej *a* i *b*
- konstrukcja prostej teoretycznej
- przykłady
- transformacja funkcji nieliniowych

Tadeusz M.Molenda, Instytut Fizyki, Uniwersytet Szczeciński

Regresja liniowa – na czym polega?

Jeśli mierzone dwie wielkości ***x*** i ***y*** związane są ze sobą równaniem liniowym $y = ax + b$ to obrazem graficznym jest linia prosta.

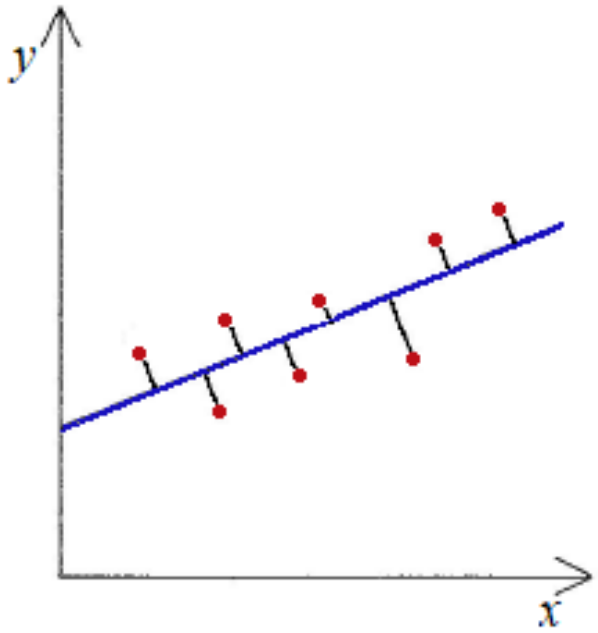
Wykonując ***n*** pomiarów wielkości ***x*** i ***y*** uzyskujemy ***n*** par liczb (x_i, y_i) Punkty o współrzędnych x_i i y_i są rozrzucone na pewnym obszarze.

Regresja liniowa to:

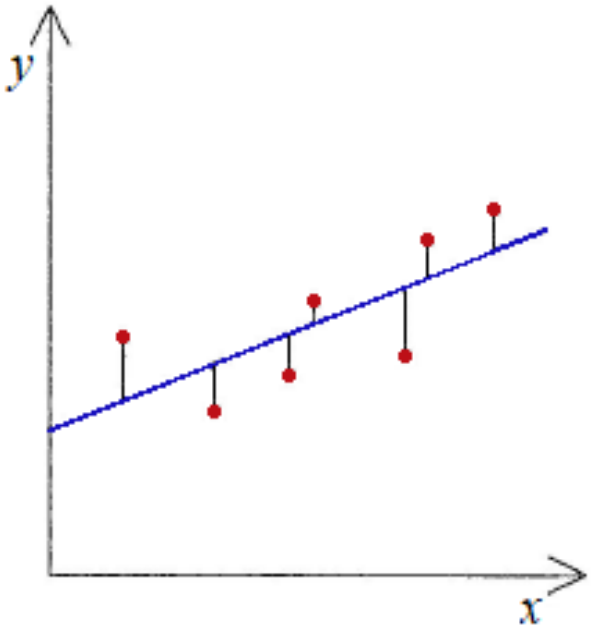
• ustalanie prawidłowości rozrzutu punktów *x* i *y* czyli:

dopasowanie prostej do zbioru punktów doświadczalnych

• szukanie równania linii prostej (tj parametrów ***a*** i ***b***), najlepiej "pasującej" do tych punktów.



a)



b)

Rys. dla najlepszego dopasowania linii prostej do punktów pomiarowych, w ten sposób aby suma kwadratów odległości od linii była minimalna

a) przypadek porównywalnych niepewności pomiaru *x* i *y*;

b) a) przypadek pomijalnie małych niepewności pomiaru wielkości *x.*

Postulat Gaussa

Wykonując ***n*** pomiarów wielkości ***x*** i ***y*** będących w zależności liniowej

$$y = ax + b$$

uzyskujemy ***n*** par liczb (x_i, y_i) i graficznym obrazem są punkty rozrzucone na pewnym obszarze, niekoniecznie na linii prostej!

Rozbieżność **wyniku pomiaru y_i i wartości teoretycznej y**

z równania $y = ax + b$

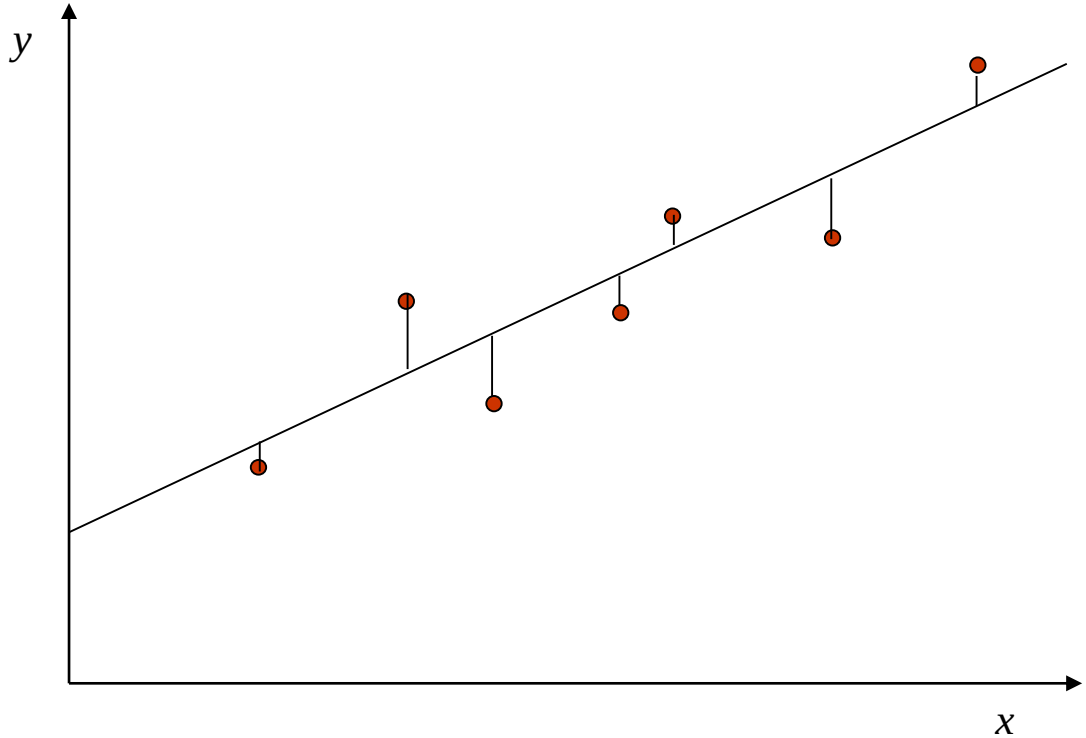
wynika z niepewności pomiarowej i można zapisać w postaci:

$$y_i - y = y_i - (ax_i + b)$$

dopasowanie metodą regresji liniowej oznacza, że

wyrażenie $$\sum_{i=1}^{n}\left(y_i - (\overline{a}x_i + \overline{b})\right)^2 = \min$$

Metoda najmniejszych kwadratów



$$\overset{?}{y} = a \cdot \overset{?}{x} + b$$

$$\sum_{i=1}^{n} \left(y_i - f\left(x_i\right) \right)^2 = \min$$

Tadeusz M.Molenda, Instytut Fizyki, Uniwersytet Szczeciński

Regresja liniowa

polega na znalezieniu parametrów *a* i *b* prostej *y = ax + b*

takich aby spełniały postulat Gaussa

$$\sum_{i=1}^{n}\left(y_i - \bar{a}x_i - \bar{b}\right)^2 = \text{minimum},$$

gdzie a i b współczynniki regresji liniowej

tj. ***aby suma kwadratów różnic między wartościami zmierzonymi y_i i obliczonymi y była jak najmniejsza (przy założeniu, że wszystkie punkty pomiarowe obarczone są jednakowymi niepewnościami przypadkowymi o rozkładzie Gaussa)***

Współczynniki *a* i *b* – wyprowadzenie

Jeśli $$f(a,b) = \sum_{i=1}^{n}\left(y_i - \bar{a}x_i - \bar{b}\right)^2$$ **= min to znaczy że:**

$$\frac{\partial f(a,b)}{\partial a} = 0 \qquad \frac{\partial f(a,b)}{\partial b} = 0,$$

Po zróżniczkowaniu otrzymujemy układ równań:

$$f(a,b) = \sum_{i=1}^{n} \left(y_i - \bar{a} x_i - \bar{b} \right)^2$$

$$\frac{\partial f(a,b)}{\partial a} = 0, \qquad \frac{\partial f(a,b)}{\partial b} = 0$$

$$-2\sum_{i=1}^{n} x_i (y_i - \bar{a} x_i - \bar{b}) = 0$$

$$-2\sum_{i=1}^{n} (y_i - \bar{a} x_i - \bar{b}) = 0$$

Po rozwiązaniu układu równań otrzymuje się wzory na współczynniki ***a*** i ***b***, gdzie i = 1,2,3,..., *n*, (*n jest ilością par punktów* (x_i, y_i)).

$$\boxed{\bar{a} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2}}$$

$$S_{\bar{a}} = \sqrt{\frac{n\left[\sum_{i=1}^{n} y_i^2 - \bar{a}\sum_{i=1}^{n} x_i y_i - \bar{b}\sum_{i=1}^{n} y_i \right]}{(n-2)\left[n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=2}^{n} x_i \right)^2 \right]}}$$

$$\boxed{\bar{b} = \frac{1}{n}\left(\sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i \right)}$$

$$S_{\bar{b}} = S_{\bar{a}} \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}$$

Metoda najmniejszych kwadratów

wzory dla parametrów regresji liniowej

$$\bar{a} = \frac{\left(\sum_{i=1}^{n} x_i\right) \cdot \left(\sum_{i=1}^{n} y_i\right) - n \sum_{i=1}^{n} \left(x_i y_i\right)}{\left(\sum_{i=1}^{n} x_i\right)^2 - n \sum_{i=1}^{n} x_i^2}$$

$$\bar{b} = \frac{\left(\sum_{i=1}^{n} x_i\right) \cdot \left(\sum_{i=1}^{n} x_i y_i\right) - \left(\sum_{i=1}^{n} x_i^2\right) \cdot \left(\sum_{i=1}^{n} y_i\right)}{\left(\sum_{i=1}^{n} x_i\right)^2 - n \sum_{i=1}^{n} x_i^2}$$

Tadeusz M.Molenda, Instytut Fizyki, Uniwersytet Szczeciński

Metoda najmniejszych kwadratów

wzory dla odchylenia standardowego parametrów regresji liniowej

$$S_{\bar{a}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n} d_i^2} \cdot \sqrt{\frac{n}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}}$$

$$S_{\bar{b}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n} d_i^2} \cdot \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}}$$

gdzie: $$d_i = y_i - \left(\bar{a}x_i + b\right)$$

Tadeusz M.Molenda, Instytut Fizyki, Uniwersytet Szczeciński

Współczynnik korelacji liniowej Pearsona

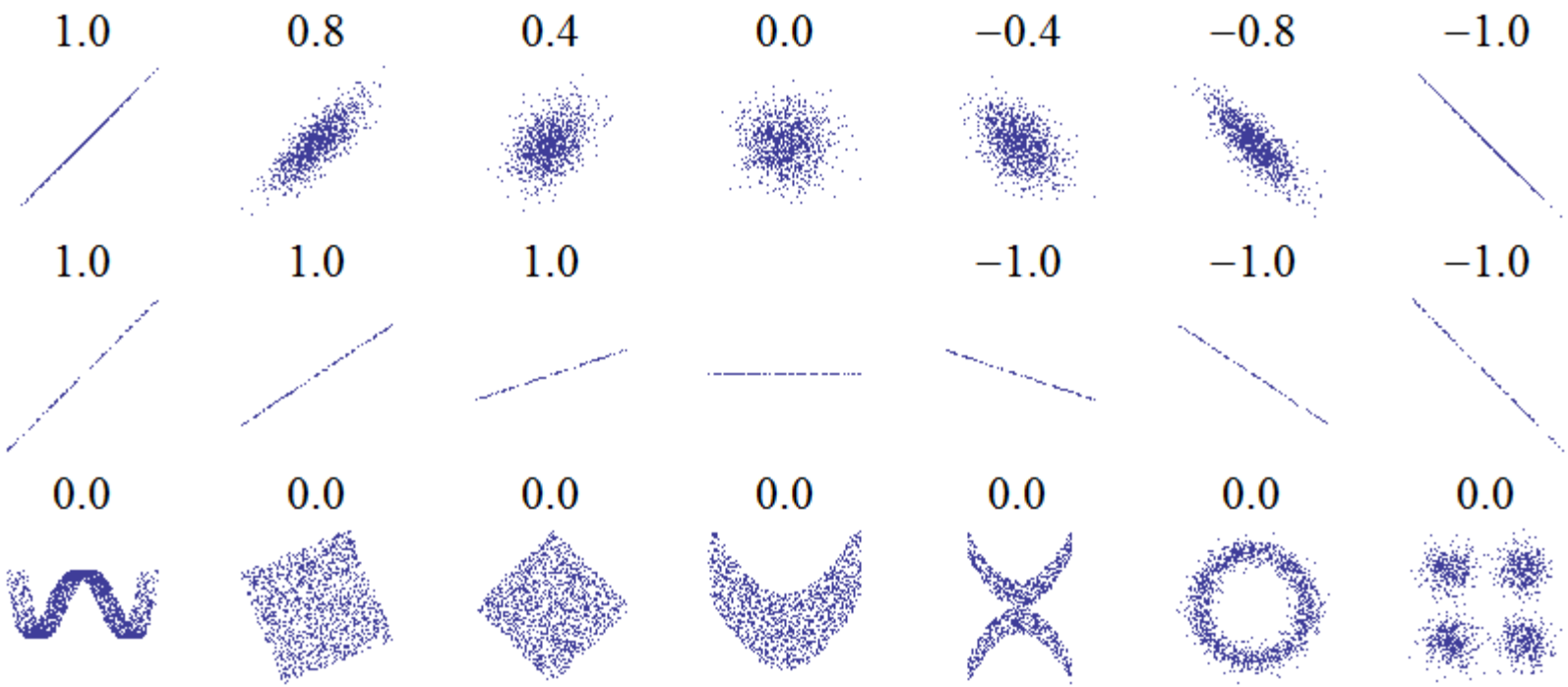
bezwymiarowy wskaźnik z przedziału [–1, 1] określający stopień liniowej zależności między zmiennymi losowymi.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\cdot(y_i - \bar{y})}{\sqrt{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)\cdot\left(\sum_{i=1}^{n}(y_i - \bar{y})^2\right)}}$$

Dla obliczeń komputerowych przydatny jest wzór

$$r = \frac{n\sum_{i=1}^{n}(x_i y_i) - \left(\sum_{i=1}^{n} x_i\right)\cdot\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{\left(n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right)\cdot\left(n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right)}}$$

Warunkiem stosowania regresji liniowej jest aby wartość bezwględna współczynnika *r* była bliska 1.

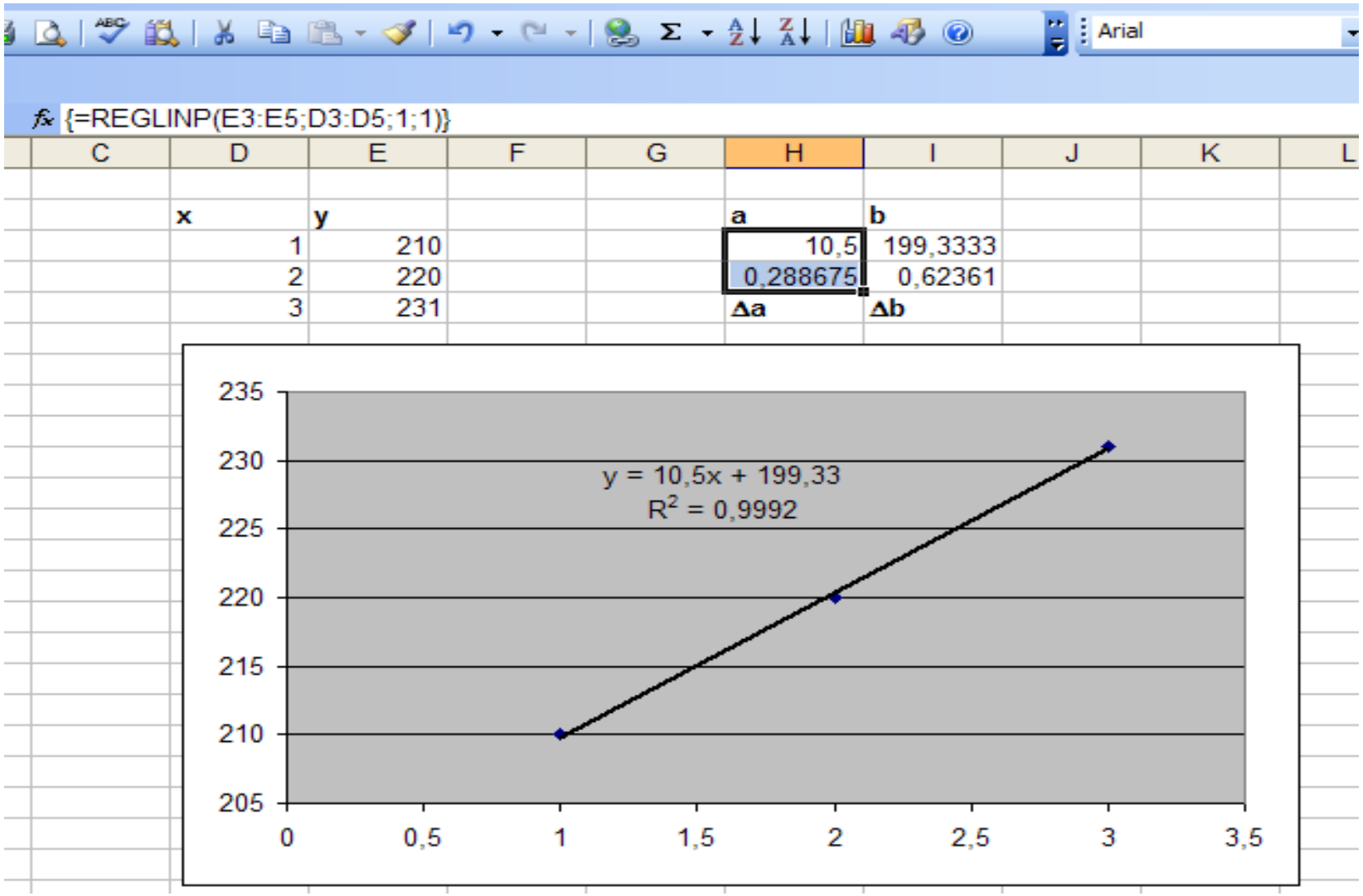


Przykładowe wykresy danych (*x*, *y*) i odpowiadające im wartości współczynnika korelacji liniowej Pearsona

Źródło: Wikipedia: *Współczynnik korelacji Pearsona*,
https://pl.wikipedia.org/wiki/Wsp%C3%B3%C5%82czynnik_korelacji_Pearsona#Poziomy_korelacji_i_ich_interpretacja

Analiza danych pomiarowych

Regresja linowa. Wykorzystanie arkusza kalkulacyjnego.



Regresja liniowa – klasyczna (metoda najmniejszych kwadratów)

Jeżeli pomiędzy dwiema wielkościami fizycznymi występuje zależność liniowa to regresja liniowa jest prostą metodą wyznaczenia parametrów najlepiej dopasowanej prostej. Parametry prostej określonej równaniem $\boldsymbol{y} = \boldsymbol{mx} + \boldsymbol{b}$ wyznaczamy przy użyciu ogólnie dostępnych (dość złożonych) wzorów. Znając współczynniki m i b regresji liniowej oraz współczynnik korelacji (Pearsona) *r* można, korzystając z poniższych wzorów, obliczyć niepewności pomiaru (odchylenia standardowe) typu A (statystyczne)

$$u_{\mathrm{A}}(m) = |m| \sqrt{\frac{1/r^2 - 1}{n-2}}, \qquad u_{\mathrm{A}}(b) = u_{\mathrm{A}}(m) \sqrt{\left(\sum_{i=1}^{n} x_i^2\right) / n}$$

Wartości współczynników charakteryzujących prostą dla regresji liniowej szybko otrzymamy korzystając z funkcji wbudowanych w arkuszu kalkulacyjnym.

Współczynnik korelacji liniowej Pearsona *r* – bezwymiarowy wskaźnik z przedziału [–1, 1] określający stopień liniowej zależności dwóch zestawów danych. Składnia w Excelu: =PEARSON(tablica1;tablica2).

Współczynniki regresji liniowej, składnia w Excelu:

m: =NACHYLENIE(*znane_y*;*znane_x*); ***b***: =ODCIĘTA(*znane_y*;*znane_x*)

Uwaga: zwrócić uwagę, że na pierwszym miejscu jest „*y*” a na drugim „*x*”.

Wartości: m i b, $u_{\mathrm{A}}(m)$ i $u_{\mathrm{A}}(b)$ oraz r^2 i $u(r)$ otrzymamy korzystając z bardziej wszechstronnej funkcji tablicowej REGLINP, która zwraca tablicę wartości.

Składnia: =REGLINP(*znane_y*;*znane_x*;*stała*;*statystyka*).

Stała – argument opcjonalny; domyślna wartość PRAWDA oznacza normalne liczenie wartości wpółczynnnika b; wartość FAŁSZ wymusza , to stała $b = 0$ (wartość *m* jest dopasowana do danych tak, aby spełnić równanie $y = mx$), tak jest w naszym przypadku.

Statystyka – argument opcjonalny. Jeżeli dla wyświetlenia wartości funkcji oznaczymy obszar „2 kolumny na 2 wiersze (3 wiersze)” i wartością jest:

– **PRAWDA**, to funkcja w kolejnych wierszach zwraca kolejno: m i b, $u_{\mathrm{A}}(m)$ i $u_{\mathrm{A}}(b)$ – przy zaznaczeniu obszaru z 2 wierszami (oraz r^2 i $u(r)$ przy zaznaczeniu obszaru z 3 wierszami).

– **FAŁSZ** lub argument został pominięty, to funkcja zwraca jedynie wartości współczynników m i b.

Aby użyć funkcję REGLINP trzeba: (i) zaznaczyć obszar w którym ma się znaleźć wynik; (ii) wpisać nazwę funkcji; (iii) zatwierdzić jej wprowadzanie kombinacją klawiszy *Ctrl*+*Shift*+*Enter*.

Na temat wszystkich statystyk, generowanych przez funkcję REGLINP można przeczytać w Pomocy.

Uwaga. W arkuszu kalkulacyjnym jest wykorzystana tzw. normalna metoda najmniejszych kwadratów, pojawia się pytanie na ile ta metoda, w porównaniu do prostej regresji ortogonalnej z rys. odręcznego, jest zgodna.

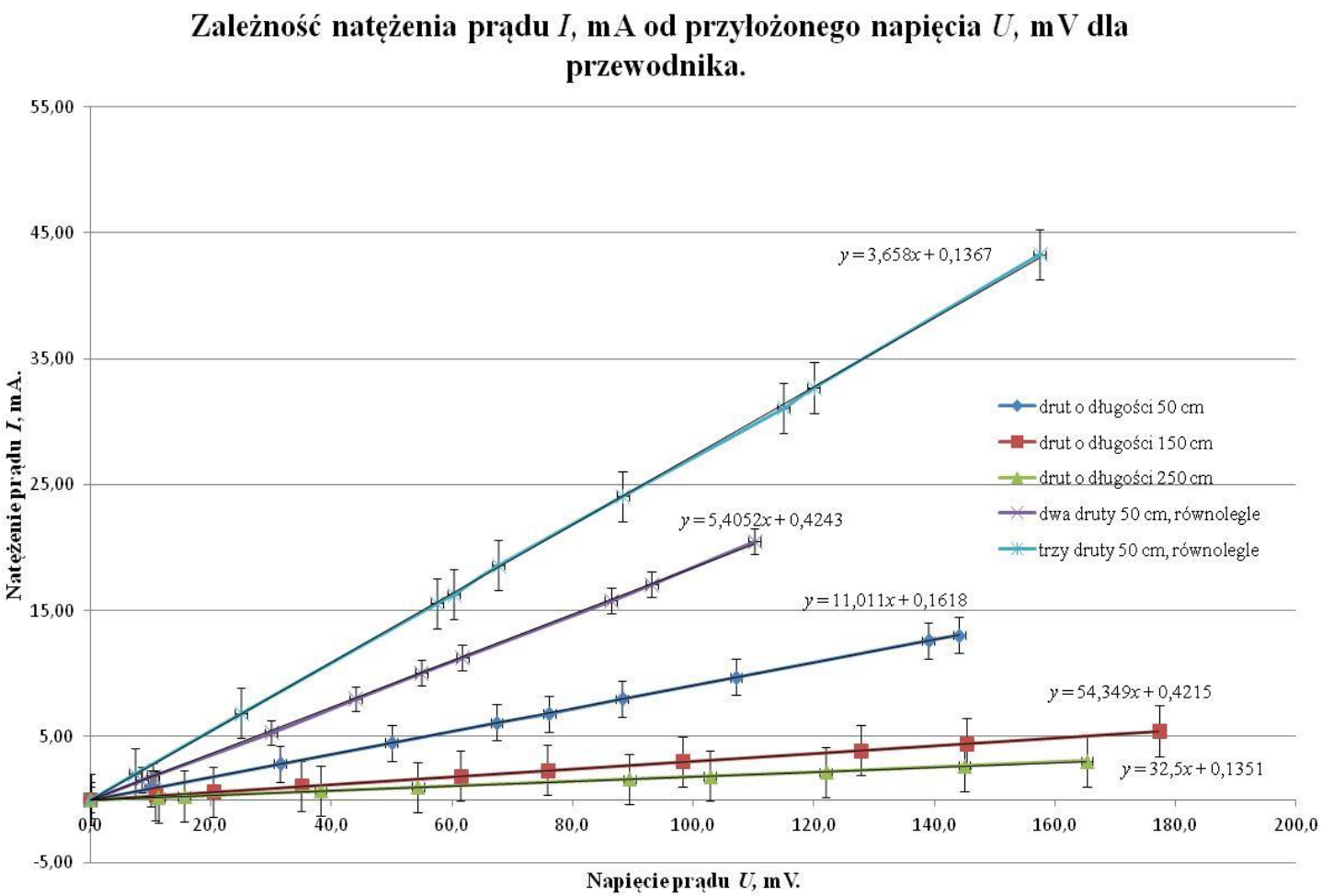
Przykład utworzenia wykresu $I = I\,(U)$ z zaznaczeniem odcinków niepewności na podstawie danych z doświadczenia: Doświadczalne potwierdzenie prawa Ohma

Szczegóły patrz: M. Dyjak, ***Instrukcja właściwego wykonania wykresów na zajęcia dydaktyczne***

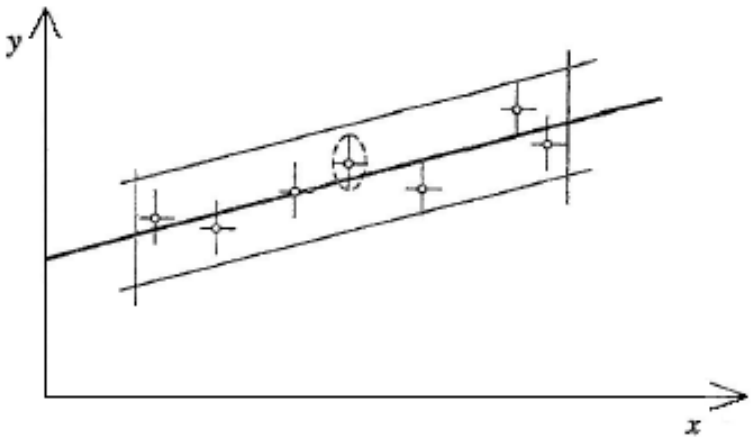
Tabela: Dane pomiarowe U, I z wartościami niepewności granicznych dla mierników cyfrowych

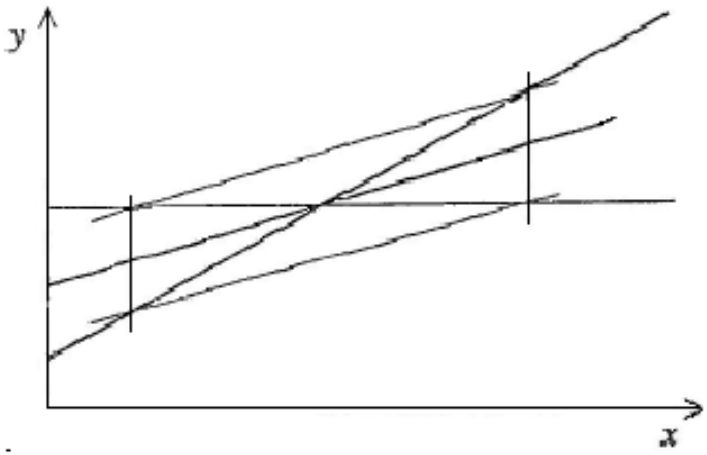
Lp.	U, mV	0,5 % $\{U\}$ + 0,1 mV	I, mA	0,8 % $\{I\}$+ 0,01 mA
1.	0,0	0,10	0,00	0,10
2.	7,4	0,47	2,07	0,20
3.	25,0	1,35	6,86	0,44
4.	57,5	2,98	15,60	0,88
5.	60,3	3,12	16,30	0,92
6.	67,7	3,49	18,60	1,03
7.	88,3	4,52	24,10	1,31
8.	115,0	5,85	31,10	1,66
9.	120,0	6,10	32,70	1,74
10.	157,5	7,98	43,30	2,27

$\{W\}$ – wartość liczbowa wielkości fizycznej W



Rys. **Wykres zależności natężenia prądu *I* od przyłożonego napięcia *U* dla przewodnika**.
Przedstawiono zależności dla kilku długości i połączeń drutów oporowych.
Zaznaczone są „odcinki niepewności”.
Szczegóły patrz: M. Dyjak, ***Instrukcja właściwego wykonania wykresów na zajęcia dydaktyczne***





Graficzna ocena parametrów linii, równoległobok niepewności pomiaru.

Prosta teoretyczna

Prostą o nachyleniu ***a*** przecinającą oś y w punkcie ***b*** nazywamy **prostą teoretyczną**. Ta prosta o wyliczonych parametrach **a** i **b** jest rezultatem **najlepszego uśrednienia** wyników. Wykonując wykres należy nanieść prostą teoretyczną a następnie punkty pomiarowe.

Wada metody:

W wyniku obliczeń otrzymuje się wartości ***a*** **i** ***b*** nawet wtedy, gdy mierzone wartości **nie** są liniowo zależne.

Przykłady:

1. $s = f(t), \quad s = vt \quad \rightarrow \quad y = s, \quad x = t, \quad a = v, \quad b \approx 0$
2. $V = f(t), \quad V = V_0 + at \quad \rightarrow \quad y = V, \quad x = t, \quad a = a, \quad b = V_0$
3. $R = f(t), \quad R = R_0(1 + \alpha t) \quad \rightarrow \quad y = R, \quad x = t, \quad a = R_0\alpha, \quad b = R_0$
4. $\alpha = f(t), \quad \alpha = k_{T,\lambda} dc \quad \rightarrow \quad y = \alpha/d, \quad x = c, \quad a = k_{T,\lambda}$
5. $T = f(R, C), \; T = kRC \quad \rightarrow \quad y = T, \quad x = RC, \; a = k, \; b \approx 0$

Przykład opracowania danych metodą regresji liniowej

$R = R_0 (1 + \alpha \Delta T)$ *zależność rezystancji od temperatury R = f (T)*

ΔT, K	19	38	50	65	80
R, Ω	150	159	170	175	185

Znaleźć równanie prostej najlepiej pasującej do tych danych

1. wzór:

$R = R_0(1 + \alpha \Delta T)$

$R = R_0 + R_0 \alpha \Delta T$

2. znaleźć ***x*** i ***y***

3. sporządzić **tabelę**

lp	x_i ΔT, K	y_i R, Ω	$x_i y_i$	x_i^2	y_i^2
1	19	150	19 x 150	19^2	150^2
2	38	159			
3	50	170			
4	65	175			
5	80	185			
Σ	$\sum x_i$ =252	$\sum y_i$ =839	$\sum x_i y_i$ = 43567	$\sum x_i^2$= 14930	$\sum y_i^2$ = 141531

4. **podstawić wartości:**

$\sum x_i = 252$, $\sum y_i = 839$, $\sum x_i y_i = 43567$, $\sum x_i^2 = 14930$, $\sum y_i^2 = 141531$

do wzorów na a, b, S_a i S_b:

$a = 0{,}5748$; $b = 138{,}83$; $S_a = 0{,}039$; $S_b = 2{,}15$

5. **zapisać wzory końcowe na a i b**

$a = (0{,}57 \pm 0{,}04)$ K, $b = (138{,}8 \pm 2{,}2)\ \Omega$

6. **zapisać równanie regresji liniowej**

$y = 0{,}57x + 138{,}8$

$R = R_0 \alpha \Delta T + R_0$ czyli $R = 0{,}57\,\Delta T + 138{,}8$

$x = T$,

$y = R$

$a = R_0 \alpha$

$b = R_0$

7. **sporządzić wykres $R = f(T)$**

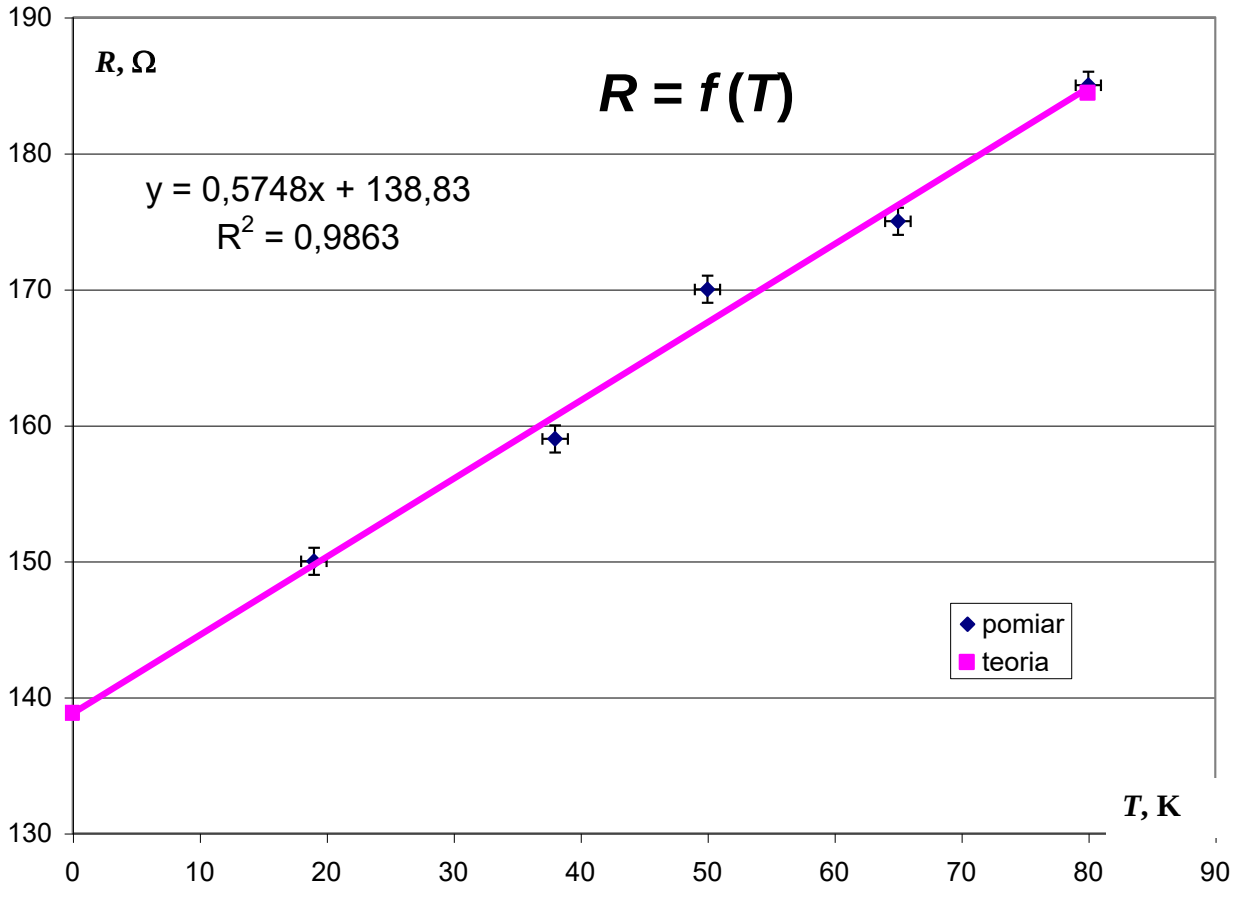
8. **nanieść na wykres niepewności i proste y_{min} i y_{max}**

$y_{min} = (a - S_a)\,x + b - S_b$, $y_{max} = (a + S_a)\,x + b + S_b$

$y = ax + b$,

dla dowolnego x z pomiarów oraz a i b obliczonych metodą regresji wyliczam y.

Mam dwa punkty (0,*b*) i (*x*,*y*) i prowadzę prostą teoretyczną.



Tadeusz M.Molenda, Instytut Fizyki, Uniwersytet Szczeciński

Transformacja funkcji nieliniowych do funkcji liniowych

$$S = \frac{1}{2} a t^2$$

$S = f(t^2)$

$S = y,\ 1/2a = a,\ t^2 = x$

$$E = \delta T^4$$

$E = f(T)$ $\ln E = \ln\delta + 4\ln T$

$\ln E = y,\ \ln\delta = b,\ 4 = a,\ \ln T = x$

$$Q = Q_0 e^{-\frac{t}{RC}}$$

$Q = f(t)$ $\ln Q = \ln Q_0 - t/RC$

$\ln Q = y,\ \ln Q_0 = b,\ -1/RC = a,\ t = x$