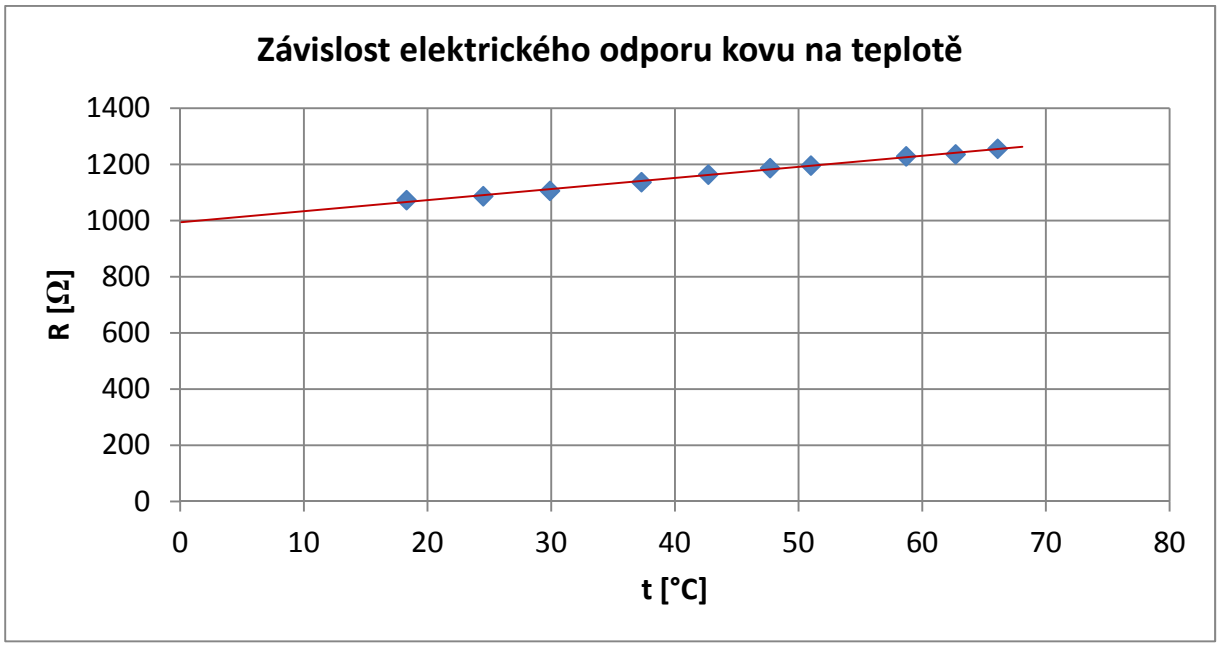
Lineární regrese

Častým úkolem je stanovení vzájemné závislosti dvou (či více) fyzikálních veličin a její matematické vyjádření. K tomuto účelu se používají různé regresní metody, pomocí nichž hledáme vhodnou funkci *f (x)*, aproximující závislost mezi naměřenými veličinami. Jedna z nejčastějších metod je *metoda nejmenších čtverců*.

Mějme *n* naměřených dvojic $[x_i;y_i]$, kterými prokládáme křivku určenou rovnicí $y = f(x)$. Hledáme takovou funkci $f(x)$, která má minimální součet druhých mocnin rozdílů ypsilonových souřadnic naměřených bodů a bodů ležících na proložené křivce:

$$S = \sum_{i=1}^{n} \left(y_i - f(x_i)\right)^2 \text{ je minimální}$$

Z matematiky víte, že k takovému výpočtu slouží parciální derivace, které položíme rovny nule. Obecně lze tento postup aplikovat na řadu funkcí $f(x)$, ale nejčastěji se používá pro aproximaci dat přímkou $y = f(x) = k \cdot x + q$, čili tzv. ***lineární regresi***.



Lineární regrese

Přesné odvození regresních koeficientů *k* a *q* lze nalézt v literatuře, zde uvádím až výsledné vztahy:

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

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

Vhodnost použití lineární regrese ověříme pomocí *korelačního koeficientu* r_{xy}, jehož hodnota leží v intervalu <-1; 1>. Aproximace přímkou je oprávněná, je-li $|r_{xy}| > 0{,}99$ (tzv. *pravidlo dvou devítek*). Pro výpočet platí vztah:

$$r_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}}, \text{ kde } \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \text{ a } \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

Pro odchylky nalezených regresních koeficientů platí vztahy:

$$\sigma_k = \sqrt{\frac{S_0}{(n-2) \cdot \left[\sum_{i=1}^{n} x_i^2 - \frac{1}{n} \cdot \left(\sum_{i=1}^{n} x_i\right)^2\right]}} \quad \text{a} \quad \sigma_q = \sqrt{\frac{S_0 \cdot \frac{1}{n} \cdot \sum_{i=1}^{n} x_i^2}{(n-2) \cdot \left[\sum_{i=1}^{n} x_i^2 - \frac{1}{n} \cdot \left(\sum_{i=1}^{n} x_i\right)^2\right]}}$$

$$\text{kde } S_0 = \left(\sum_{i=1}^{n} y_i^2\right) - \frac{1}{n} \cdot \left(\sum_{i=1}^{n} y_i\right)^2 - k \cdot \left[\sum_{i=1}^{n} x_i y_i - \frac{1}{n} \cdot \left(\sum_{i=1}^{n} x_i\right) \cdot \left(\sum_{i=1}^{n} y_i\right)\right]$$

Interval spolehlivosti stanovení regresních koeficientů, a tedy jejich přesnost, závisí na těchto odchylkách a zvolené pravděpodobnosti *P*. Studentův součinitel $t_{P,(n-1)}$ má parametry *n-1* a *P* = 95 %.

Výpočty regresních koeficientů a jejich chyb nemusíte provádět ručně, je výhodné použít výpočetní techniku (kalkulačky, programy pro PC …). Například v programu EXCEL slouží k těmto výpočtům funkce LINREGRESE() používaná jako maticový vzorec.

Lineární regrese

Příklad: Bylo provedeno měření závislosti elektrického odporu kovu na teplotě (naměřená data viz. tabulka níže). Je možné získaná data proložit přímkou? Určete parametry přímky nejlépe vystihující získanou závislost včetně jejich chyb.

t [°C]	18,3	24,5	29,9	37,3	42,7	47,7	51,0	58,7	62,7	66,1
R [Ω]	1073	1087	1106	1137	1164	1187	1196	1229	1236	1256

Nejprve spočteme regresní koeficient: $r_{xy} = \dfrac{\sum_{i=1}^{10}\left(t_i - \bar{t}\right)\left(R_i - \bar{R}\right)}{\sqrt{\sum_{i=1}^{10}\left(t_i - \bar{t}\right)^2 \sum_{i=1}^{10}\left(R_i - \bar{R}\right)^2}} \approx 0{,}997598$

Z jeho velikosti vyplívá, že naměřená data lze oprávněně proložit přímkou $y = k \cdot x + q$.

$$k = \frac{10 \cdot \left(\sum_{i=1}^{10} t_i R_i\right) - \left(\sum_{i=1}^{10} t_i\right) \cdot \left(\sum_{i=1}^{10} R_i\right)}{10 \cdot \left(\sum_{i=1}^{10} t_i^2\right) - \left(\sum_{i=1}^{10} t_i\right)^2} = \frac{10 \cdot 521\,726{,}7 - 438{,}9 \cdot 11\,671}{10 \cdot 21\,666{,}21 - (438{,}9)^2} \approx 3{,}94796$$

$$q = \frac{\left(\sum_{i=1}^{10} t_i^2\right) \cdot \left(\sum_{i=1}^{10} R_i\right) - \left(\sum_{i=1}^{10} t_i\right) \cdot \left(\sum_{i=1}^{10} t_i R_i\right)}{10 \cdot \left(\sum_{i=1}^{10} t_i^2\right) - \left(\sum_{i=1}^{10} t_i\right)^2} = \frac{21\,666{,}21 \cdot 11\,671 - 438{,}9 \cdot 521\,726{,}7}{10 \cdot 21\,666{,}21 - (438{,}9)^2} \approx 993{,}8$$

$$\sigma_k = \sqrt{\frac{S_0}{(10-2) \cdot \left[\sum_{i=1}^{10} t_i^2 - \frac{1}{10} \cdot \left(\sum_{i=1}^{10} t_i\right)^2\right]}} = \sqrt{\frac{180{,}53662}{8 \cdot \left[21\,666{,}21 - 0{,}1 \cdot (438{,}9)^2\right]}} \approx 0{,}09691$$

$$\sigma_q = \sqrt{\frac{S_0 \cdot \frac{1}{10} \cdot \sum_{i=1}^{10} t_i^2}{(10-2) \cdot \left[\sum_{i=1}^{10} t_i^2 - \frac{1}{10} \cdot \left(\sum_{i=1}^{10} t_i\right)^2\right]}} = \sqrt{\frac{180{,}53662 \cdot 0{,}1 \cdot 21\,666{,}21}{8 \cdot \left[21\,666{,}21 - 0{,}1 \cdot (438{,}9)^2\right]}} \approx 4{,}51089$$

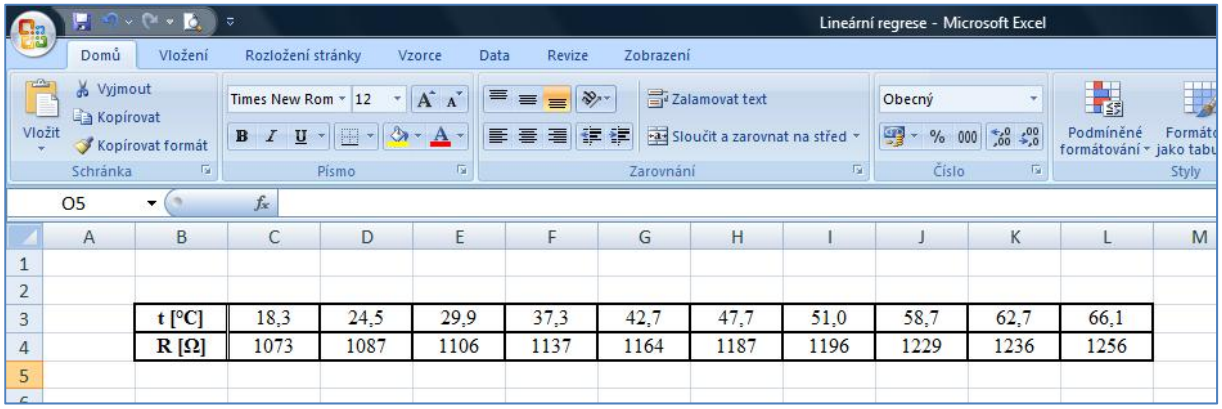
Studentův součinitel $t_{0{,}95;\,9} = 2{,}306$.

Naměřená data lze proložit přímkou s parametry: $k = (3{,}9 \pm 0{,}2)$ a $q = (990 \pm 10)$

Zpracování dat pomocí programu EXCEL

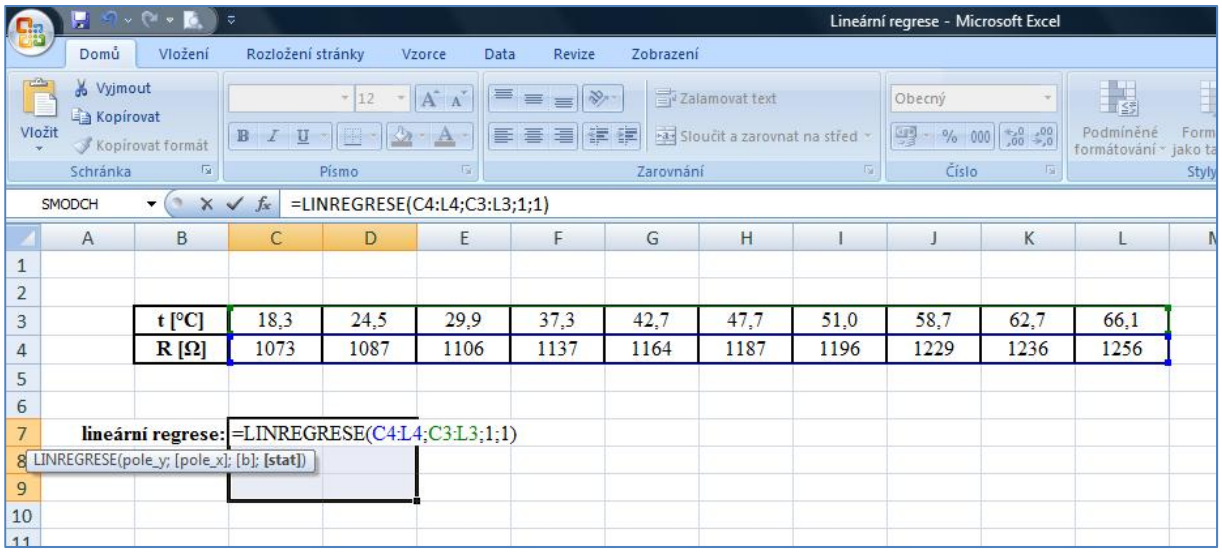
K provedení lineární regrese slouží funkce LINREGRESE(…). Je však nutné ji použít jako tzv. maticový vzorec.

Postup si ukážeme na řešení předchozího příkladu:



3	t [°C]	18,3	24,5	29,9	37,3	42,7	47,7	51,0	58,7	62,7	66,1
4	R [Ω]	1073	1087	1106	1137	1164	1187	1196	1229	1236	1256

Vyznačíme oblast 3 řádky x 2 sloupce a napíšeme vzorec s funkcí pro výpočet lineární regrese:

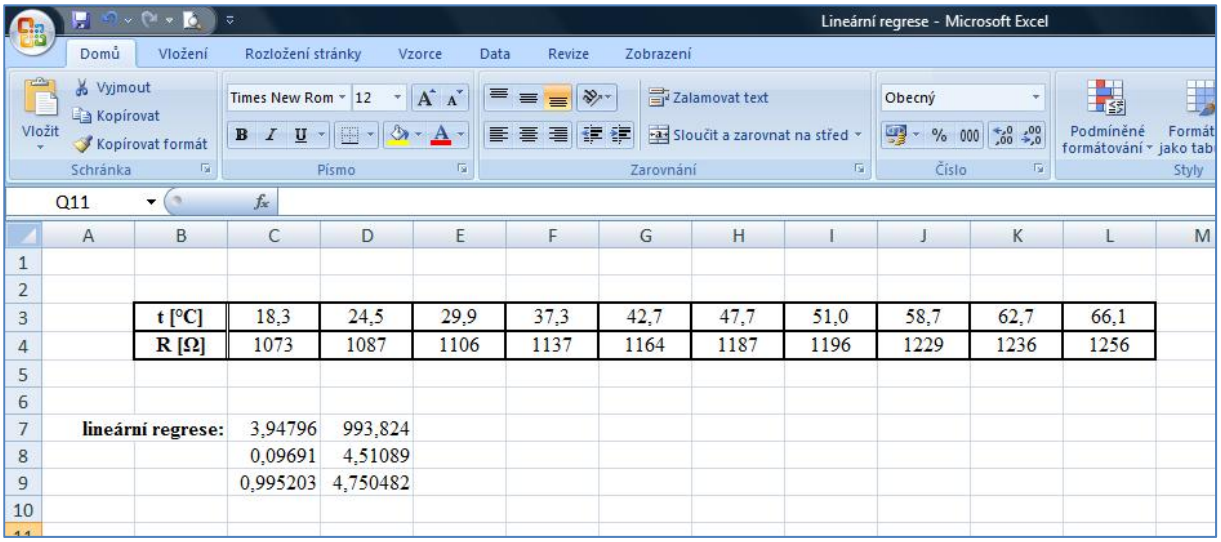


3	t [°C]	18,3	24,5	29,9	37,3	42,7	47,7	51,0	58,7	62,7	66,1
4	R [Ω]	1073	1087	1106	1137	1164	1187	1196	1229	1236	1256

Funkce LINREGRESE(…) má čtyři parametry, které oddělujeme středníkem. První parametr je oblast *y*-ových hodnot, druhý parametr je oblast *x*-ových hodnot. Třetí parametr udává, zda má být regresní koeficient *q* roven nule (parametr nastaven na 0), nebo se jeho hodnota má spočítat (parametr nastaven na 1 nebo není uveden). Čtvrtý parametr nastaven na 1 znamená, že chceme zjistit další regresní statistiky (např. odchylky koeficientů).

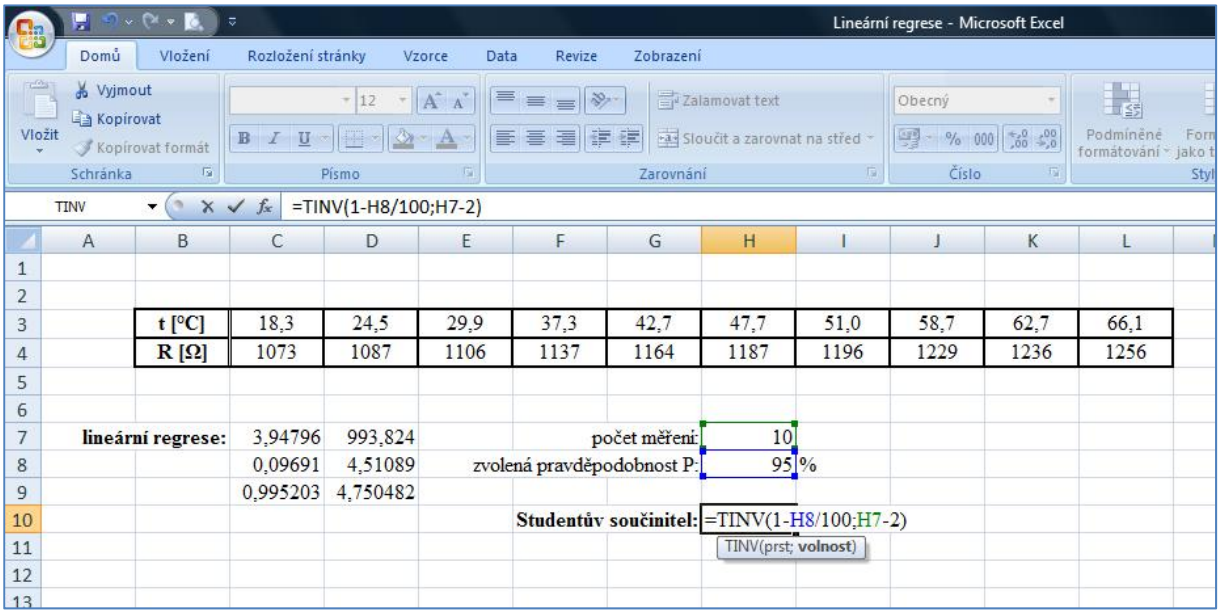
Místo klávesy *ENTER* stiskneme trojkombinaci kláves *CTRL + SHIFT + ENTER*. Ve vyznačené oblasti 3x2 se pak nachází příslušné regresní koeficienty, jejich odchylky a druhá mocnina korelačního koeficientu. Rozmístění je znázorněno v následující tabulce (údaj v bílém políčku nás nezajímá):

k	q
σ_k	σ_q
$r2=(r_{xy})^2$	



	A	B	C	D	E	F	G	H	I	J	K	L	M
1													
2													
3		**t [°C]**	18,3	24,5	29,9	37,3	42,7	47,7	51,0	58,7	62,7	66,1	
4		**R [Ω]**	1073	1087	1106	1137	1164	1187	1196	1229	1236	1256	
5													
6													
7		**lineární regrese:**	3,94796	993,824									
8			0,09691	4,51089									
9			0,995203	4,750482									
10													

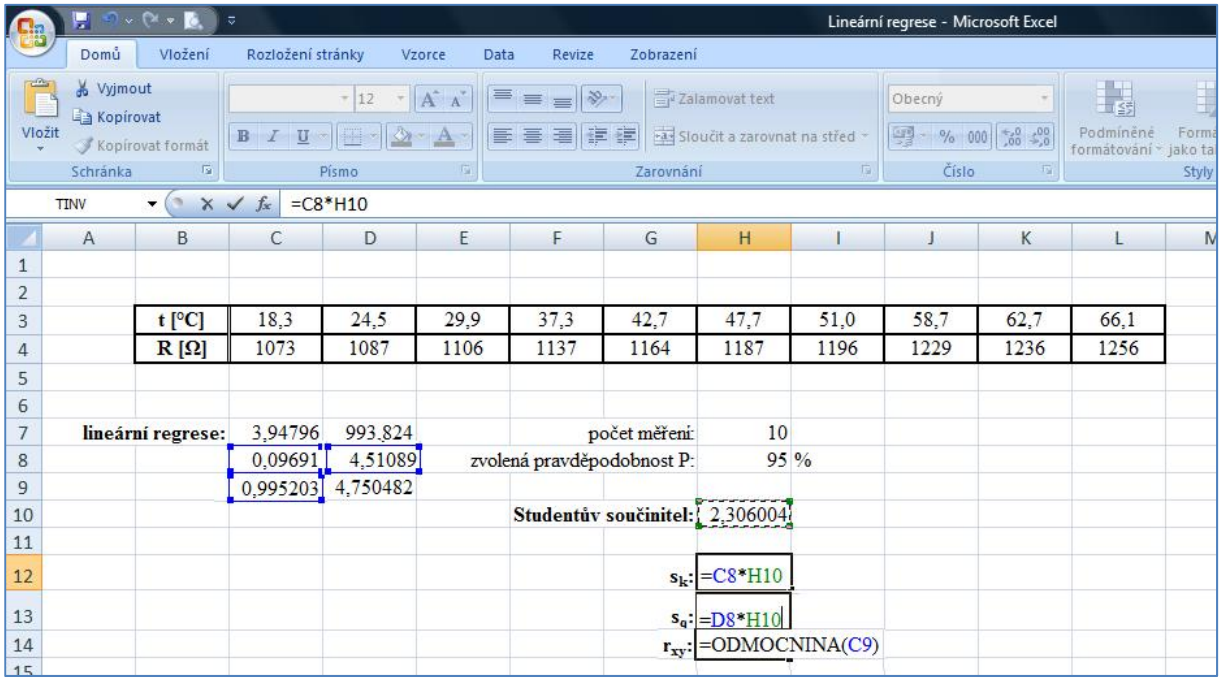
Určíme hodnotu Studentova součinitele $t_{P,(n-1)}$ pomocí funkce TINV(…). Tato funkce má dva parametry – první je pravděpodobnost, že výsledek bude ležet mimo interval spolehlivosti (pro zvolenou pravděpodobnost P je to 1- P/100) a druhým parametrem je počet stupňů volnosti (pro n měření je to n-2).



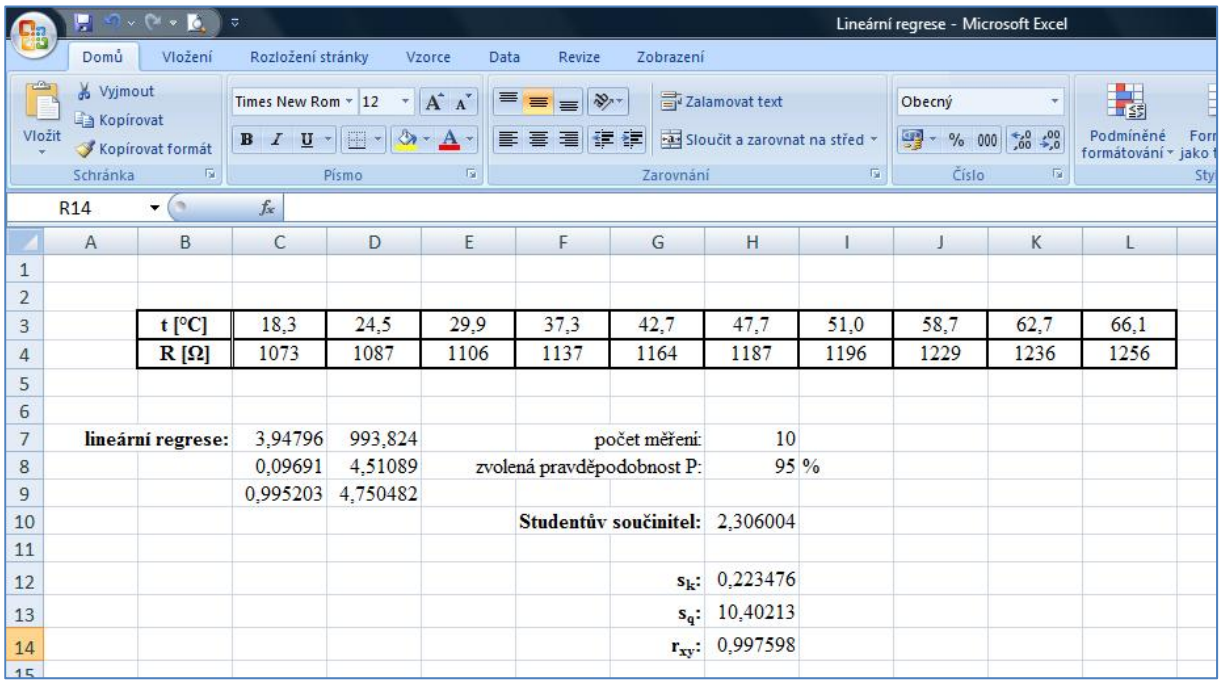
	A	B	C	D	E	F	G	H	I	J	K	L
1												
2												
3		**t [°C]**	18,3	24,5	29,9	37,3	42,7	47,7	51,0	58,7	62,7	66,1
4		**R [Ω]**	1073	1087	1106	1137	1164	1187	1196	1229	1236	1256
5												
6												
7		**lineární regrese:**	3,94796	993,824			počet měření:	10				
8			0,09691	4,51089			zvolená pravděpodobnost P:	95 %				
9			0,995203	4,750482								
10							**Studentův součinitel:**	=TINV(1-H8/100;H7-2)				
11								TINV(prst; volnost)				
12												
13												

Lineární regrese

Určíme intervaly spolehlivosti regresních koeficientů a korelační koeficient.



	A	B	C	D	E	F	G	H	I	J	K	L
3		t [°C]	18,3	24,5	29,9	37,3	42,7	47,7	51,0	58,7	62,7	66,1
4		R [Ω]	1073	1087	1106	1137	1164	1187	1196	1229	1236	1256
7		lineární regrese:	3,94796	993,824		počet měření:		10				
8			0,09691	4,51089		zvolená pravděpodobnost P:		95 %				
9			0,995203	4,750482								
10						Studentův součinitel:		2,306004				
12						s_k:		=C8*H10				
13						s_q:		=D8*H10				
14						r_{xy}:		=ODMOCNINA(C9)				



	A	B	C	D	E	F	G	H	I	J	K	L
3		t [°C]	18,3	24,5	29,9	37,3	42,7	47,7	51,0	58,7	62,7	66,1
4		R [Ω]	1073	1087	1106	1137	1164	1187	1196	1229	1236	1256
7		lineární regrese:	3,94796	993,824		počet měření:		10				
8			0,09691	4,51089		zvolená pravděpodobnost P:		95 %				
9			0,995203	4,750482								
10						Studentův součinitel:		2,306004				
12						s_k:		0,223476				
13						s_q:		10,40213				
14						r_{xy}:		0,997598				

Výsledkem je tedy proložení dat přímkou $y = k \cdot x + q$ s regresními koeficienty $k = (3,9 \pm 0,2)$ a $q = (990 \pm 10)$. Korelační koeficient $r_{xy} = 0,9976$, lineární regresi je tedy možné použít.

Poznámka: Jak souvisí získané regresní koeficienty s materiálovými konstantami zkoumaného kovu?

Pro teplotní závislost kovu platí vztah:

$$R = R_0 \cdot (1 + \alpha \cdot t),$$

kde α je teplotní součinitel elektrického odporu a R_0 je odpor kovu při teplotě 0°C.

Vztah roznásobíme a porovnáme s rovnicí přímky $y = q + k \cdot x$:

$$y \rightsquigarrow R,\ x \rightsquigarrow t,\ k = R_0 \cdot \alpha,\ q = R_0$$

Materiálové konstanty α a R_0 určíme tedy jako:

$$R_0 = q, \quad \alpha = \frac{k}{q}$$

Pro chyby pak můžeme odvodit vztahy:

$$\sigma_{R_0} = \sqrt{\left(\frac{\partial R_0}{\partial q} \cdot \sigma_q\right)^2} = \sigma_q$$

$$\sigma_\alpha = \sqrt{\left(\frac{\partial \alpha}{\partial k} \cdot \sigma_k\right)^2 + \left(\frac{\partial \alpha}{\partial q} \cdot \sigma_q\right)^2} = \sqrt{\left(\frac{\sigma_k}{q}\right)^2 + \left(\frac{-k}{q^2} \cdot \sigma_q\right)^2} = \alpha \cdot \sqrt{\left(\frac{\sigma_k}{k}\right)^2 + \left(\frac{\sigma_q}{q}\right)^2}$$

V předchozím příkladu byla naměřená data proložena přímkou $y = k \cdot x + q$ s regresními koeficienty $k = (3{,}9 \pm 0{,}2)$ a $q = (990 \pm 10)$. Materiálové konstanty α a R_0 jsou tedy:

$$R_0 = (990 \pm 10)\ \Omega$$

$$\alpha = (3{,}9 \pm 0{,}2) \cdot 10^{-3}\ K^{-1}$$