Regresní a korelační analýza

Mějme dvojici proměnných, které spolu nějak souvisí.

x je nezávisle (vysvětlující) proměnná

y je závisle (vysvětlovaná) proměnná

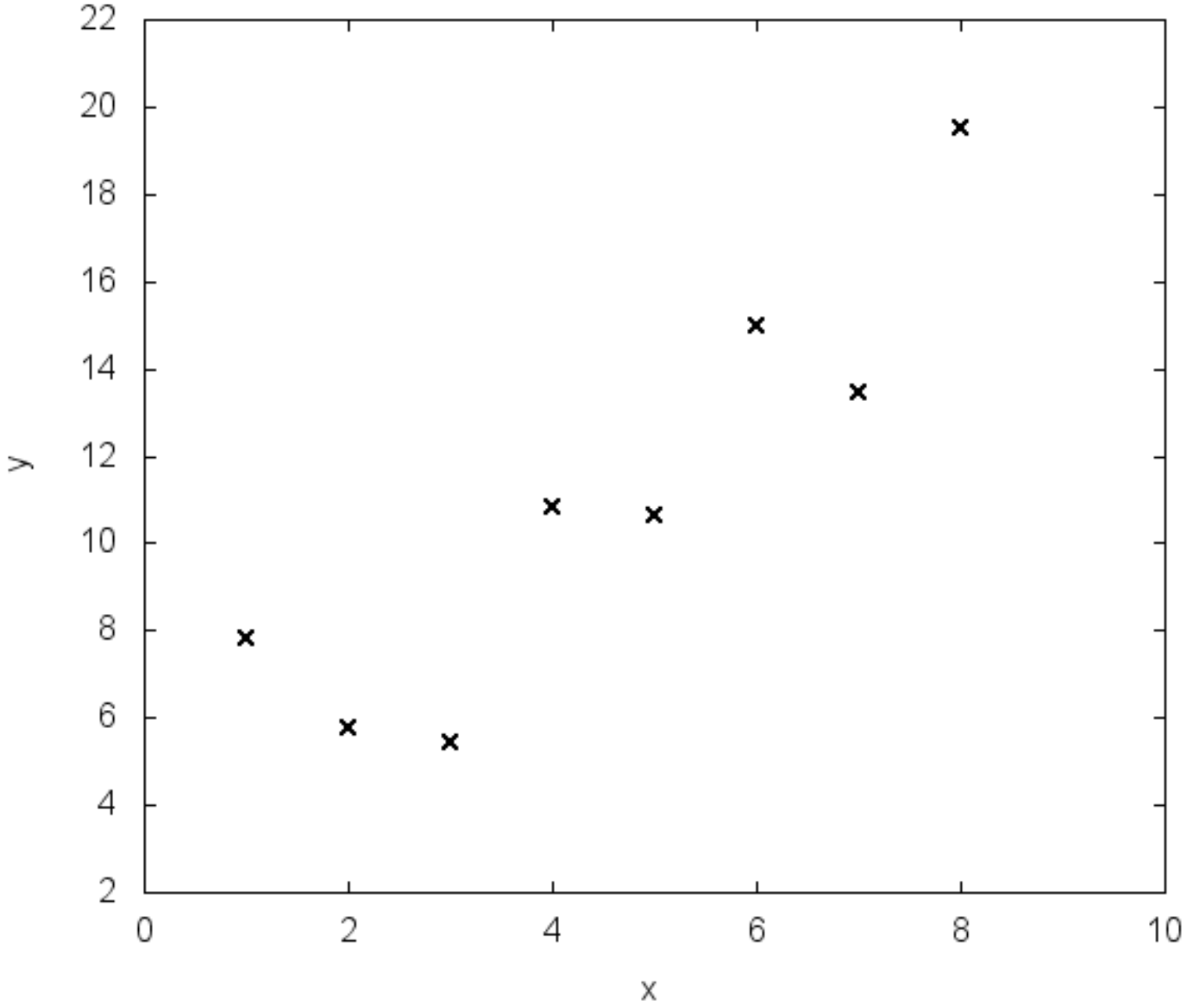
Chceme zjistit funkční závislost $y = f(x)$.

Metoda nejmenších čtverců

Přímým měřením získáme N dvojic veličin $[x_i, y_i]$, které v kartézské soustavě os x, y můžeme znázornit jako bodový graf. Předpokládejme, že mezi x a y existuje funkční vztah $y = f(x)$ známého tvaru. Pokud by při měření nevznikaly náhodné chyby, pak by všechny body $[x_i, y_i]$ ležely na křivce $y = f(x)$. Ve skutečnosti však platí $y_i = f(x_i) + \varepsilon_i$, kde ε_i je náhodná chyba i-tého měření.

Body $[x_i, y_i]$ jsou rozptýleny kolem hledané **regresní křivky**, která má být co nejvěrnějším obrazem funkce $y = f(x)$. Hledáme tedy takové parametry a, b, c, … (tzv. regresní koeficienty) daného typu funkce $y = f(x; a, b, c, \ldots)$, aby se její průběh co nejvíce přimykal k zadaným bodům $[x_i, y_i]$.

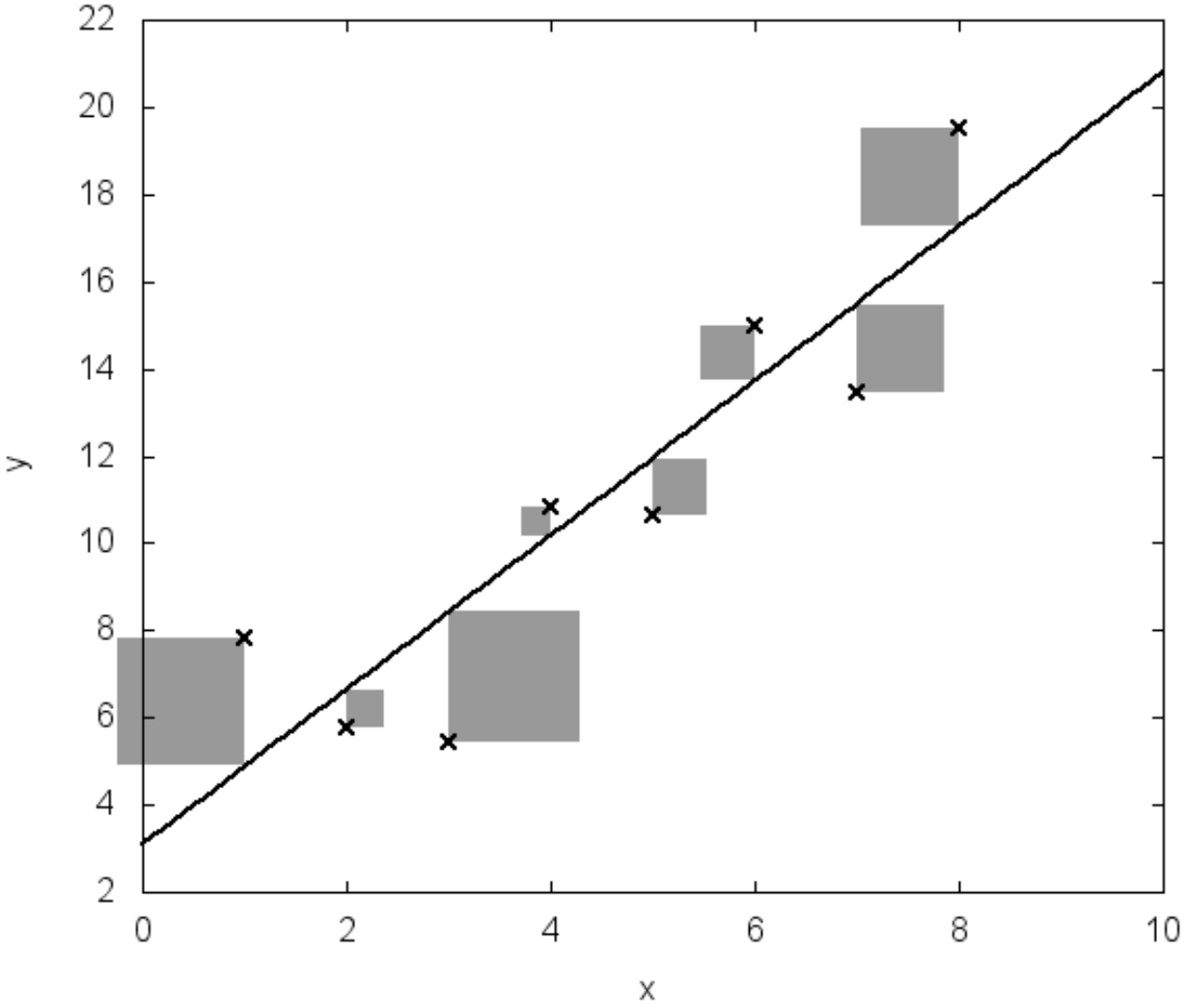
Metoda nejmenších čtverců



Metoda nejmenších čtverců

Hledáme kritérium „přiléhavosti“ regresní křivky k experimentálním bodům.
Nejvěrohodnější je tzv reziduální (zbytkový) součet „čtverců“.

$$S_{\text{resid.}} = \sum_{i=1}^{N} (\ y_i - \ f(x_i) \ )^2$$



=LINREGRESE

OOa Calc i MS Excel nabízí několik funkcí počítajících parametry lineární regrese.
Nejlepší je
=LINREGRESE(DataY;DataX;Typ;Parametr).
Lineární funkce: $y = ax + b$

Výstup funkce LINREGRESE:

(odhad parametru *a*)	(odhad parametru *b*)
(odhad chyby parametru *a*)	(odhad chyby parametru *b*)
(koeficient determinace*)	(chyba odhadu)
*F (F-statistika**)*	*df (počet stupňů volnosti)*
S_t - S_r (rozdíl celkové a reziduální sumy čtverců odchylek*)	S_r *(reziduální suma čtverců odchylek*)*

Kvalita regresního modelu

Kvalita zvoleného regresního modelu (tj. vhodnost vybrané regresní funkce a odhad jejích výběrových regresních koeficientů) se testuje. Výchozí veličinou je při tom reziduální součet čtverců $S_{resid.}$ (červeně), pomocnou celkový součet čtverců S_t (modře).

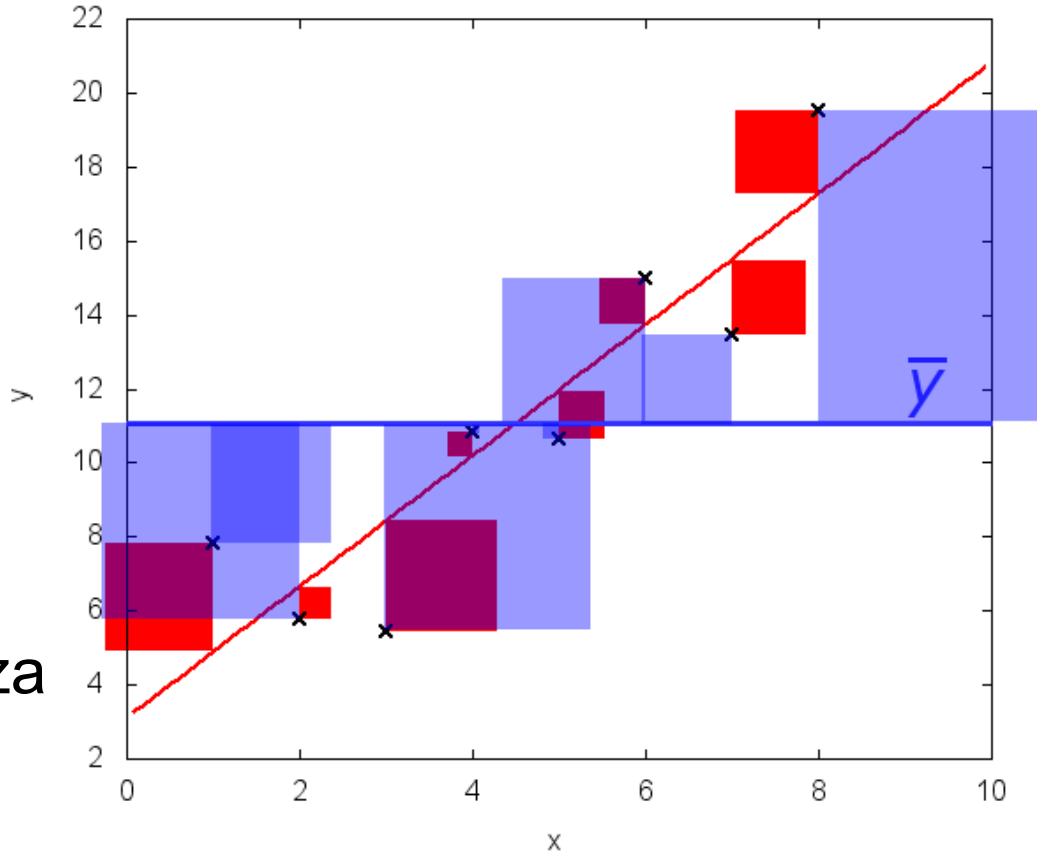
$$S_{\text{resid.}} = \sum_{i=1}^{N} (y_i - f(x_i))^2$$

$$S_t = \sum_{i=1}^{N} \left(y_i - \frac{1}{N} \sum_{i=1}^{N} y_i\right)^2$$

Koeficient determinace

$$r^2 = 1 - \frac{S_{resid.}}{S_t}$$

$r^2 > 0{,}95$ se často považuje za dobré kritérium pro přijetí zvoleného modelu



Korelační analýza

Při korelační analýze prověřujeme existenci závislosti mezi *x* a *y* a těsnost této závislosti.
Budeme předpokládat lineární závislost mezi dvěma veličinami.

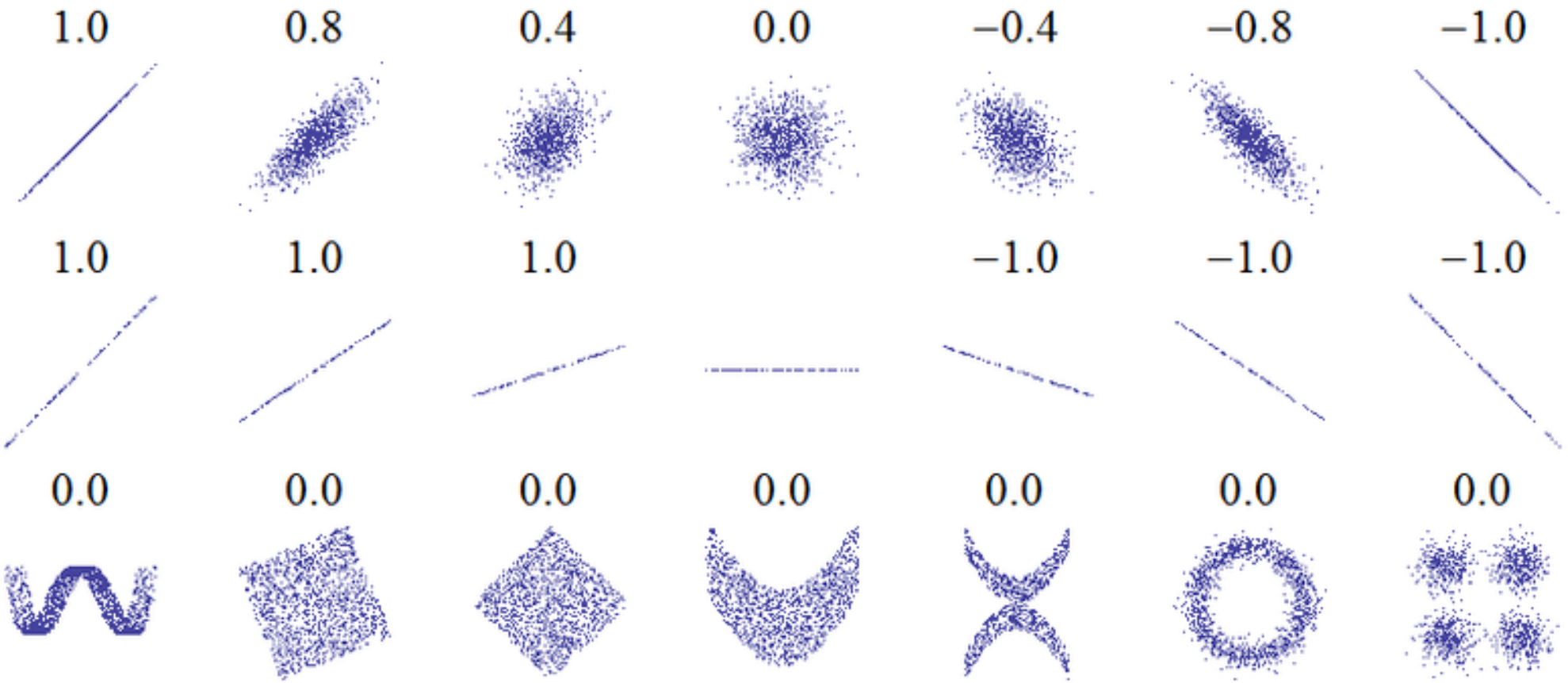
Pearsonův korelační koeficient

Při testování existence lineárního vztahu mezi proměnnými se používá odmocnina z koeficientu determinace r^2, která se nazývá (Pearsonův) korelační koeficient *r*.
Korelační koeficient se počítá pomocí funkce =CORREL() nebo ze vztahu:

$$r = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_1 - \bar{y})}{\sqrt{\sum_{i=1}^{N} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{N} (y_i - \bar{y})^2}}$$

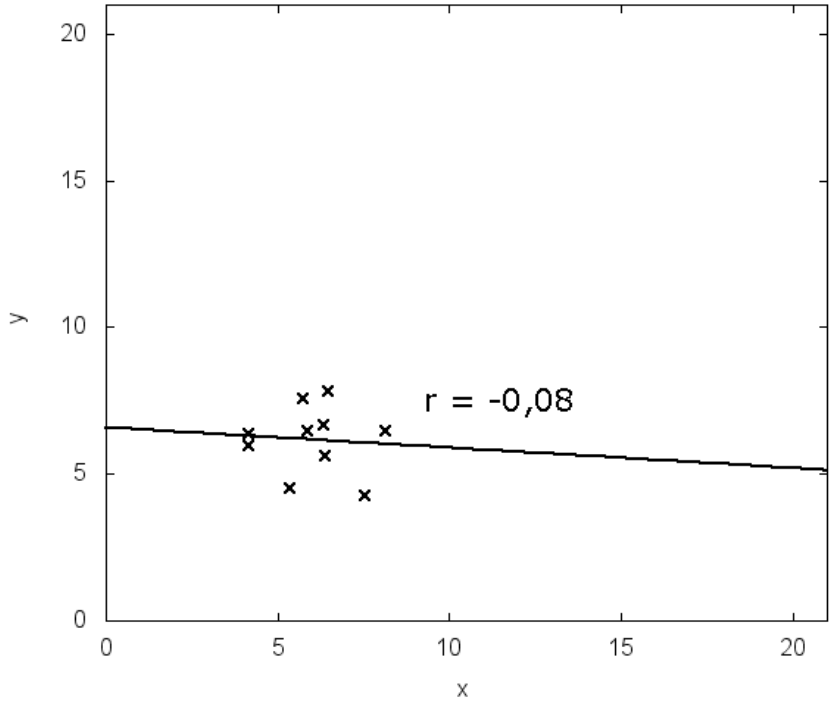
Pearsonův korelační koeficient

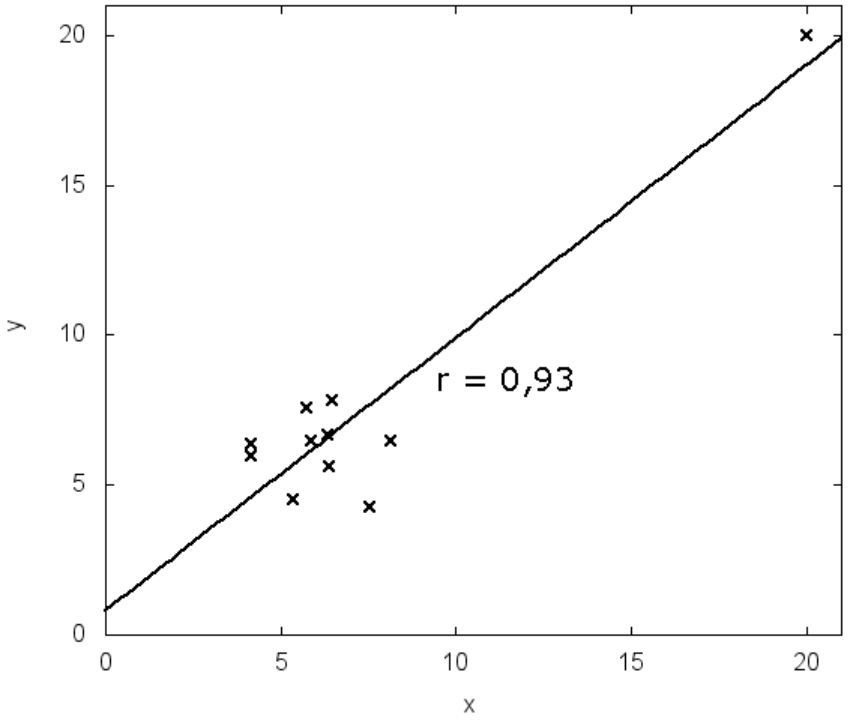
Korelační koeficient může nabývat hodnot od -1 do 1. Čím více se $|r|$ blíží 1, tím těsnější je závislost. Je-li $|r| = 1$, body x,y, leží na přímce, je-li $r = 0$, mezi body není žádný lineární vztah. Je-li $r > 0$, s rostoucím x roste i y, je-li $r < 0$, s rostoucím x naopak y klesá. Zdůrazněme, že korelační koeficient detekuje **lineární** závislost.



Záludnosti korelačního koeficientu

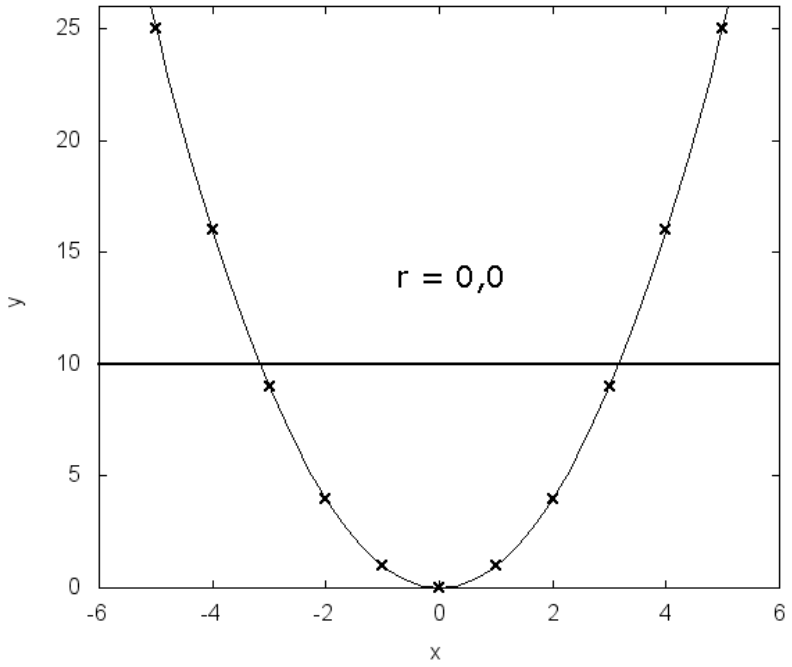
1) Je citlivý na odlehlé hodnoty.

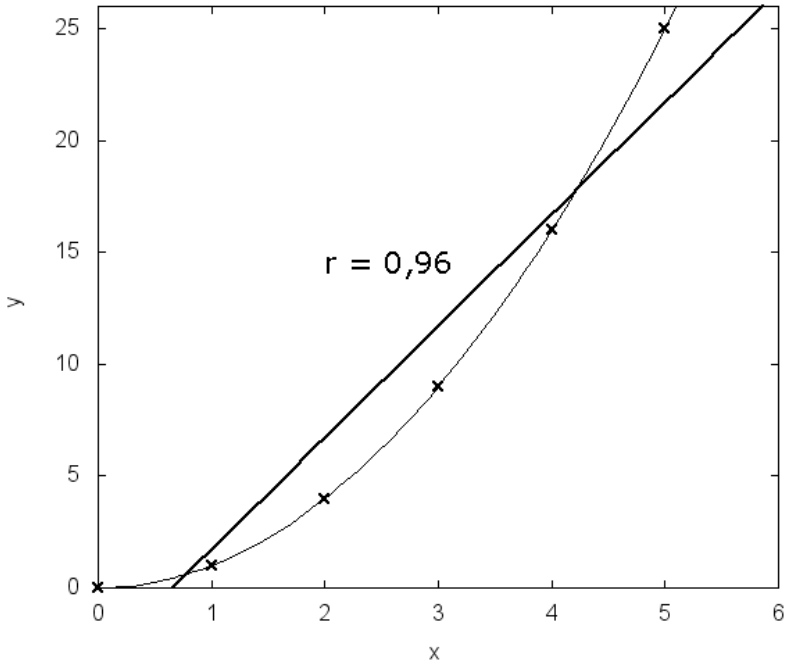




Záludnosti korelačního koeficientu

2) Detekuje pouze lineární závislost.





3) Korelace neznamená příčinnou souvislost.

Test korelačního koeficientu

Ukázali jsme, že korelační koeficient popisuje těsnost korelace mezi proměnnými *x* a *y*. Pokusme se nyní zjistit, jestli mezi proměnnými existuje vůbec nějaká (byť velmi slabá) souvislost.

Na začátku předpokládáme, že je korelační koeficient nulový (lineární závislost mezi x a y neexistuje).

Testovací kritérium: $$t = \frac{|r|}{\sqrt{1-r^2}}\sqrt{N-2}$$

Kritickou hodnotou $t_{1-\alpha}(N-2)$ jsou kvantily Studentova rozdělení s $N-2$ stupni volnosti pro zvolenou hladinu významnosti α, které najdeme ve statistických tabulkách nebo vypočítáme pomocí funkce =T.INV.T2(α,*N*-2).

Vícenásobná lineární regrese

Popisuje závislost více než dvou číselných proměnných z nichž: více je nezávislých (vysvětlující proměnné – značíme je x_1, x_2, ... , x_m) a jen jedna je závislá (vysvětlovaná proměnná y). Předpokládáme lineární závislost typu:

$y = a_1x_1 + a_2x_2 + \ldots + a_mx_m + b + \varepsilon$

deterministická složka náhodná složka (nepopsané vlivy)

Řeší se metodou nejmenších čtverců.
Řešením jsou odhady koeficientů a_1 až a_m a b.

Korelace mezi nezávislými (vysvětlujícími) proměnnými nesmí být příliš silná ($r_{ij} > 0{,}8$). V případě silné korelace se jedna z nezávislých silně korelujících proměnných vyřadí.

Korelační matice

Korelační matice obsahuje párové korelační koeficienty všech dvojic proměnných.
Lze ji vypočítat pomocí funce =correl()

	x_1	x_2	x_3	y
x_1	=CORREL(A$2:A$9;$A$2:$A$9)			
x_2	=CORREL(A$2:A$9;$B$2:$B$9)	rozkopírovat	do sloupců	→
x_3	=CORREL(A$2:A$9;$C$2:$C$9)			
y	=CORREL(A$2:A$9;$D$2:$D$9)			

Předpokládáme, že ve sloupcích A až C jsou nezávisle proměnné a v posledním sloupci D závisle proměnná.

Alternativou je použití:

Nástroje → Doplňky → Analýza → Analýza dat → Korelace

Korelační matice je symetrická podle diagonály a v hlavní diagonále má jedničky.

= LINREGRESE

=LINREGRESE(DataY;DataX;Typ;Parametr).
Lineární funkce: $y = a_1x_1 + a_2x_2 + ... + a_mx_m + b$

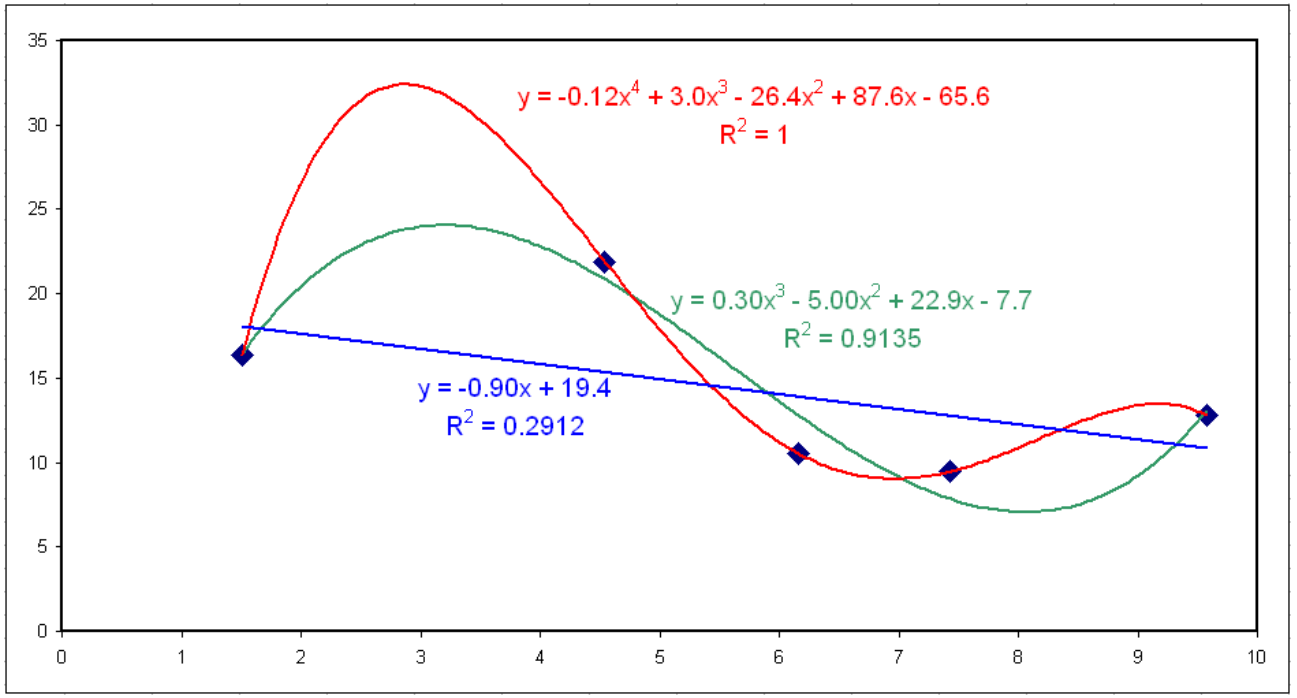
Výstup funkce LINREGRESE:

odhad a_m	...	odhad a_1	odhad b
odhad chyby a_m		odhad chyby a_1	odhad chyby b
koeficient determinace*)	chyba odhadu		
F (F-statistika)	*df* (počet stupňů volnosti)		
$S_t - S_r$ (rozdíl celkové a reziduální sumy čtverců odchylek)	S_r (reziduální suma čtverců odchylek)		

Podrobnější výstup poskytují Analytické nástroje Excelu:
Data → Analýza → Analýza dat → Regrese

Volba vhodného modelu

Platí, že čím více parametrů má regresní rovnice, tím menší je suma čtverců odchylek.
Nelze říct, že čím víc parametrů, tím lepší model - pokud je parametrů modelu stejně jako experimentálních bodů, je suma čtverců odchylek nulová.



Volba vhodného modelu

Kritériem pro rozhodnutí, zda nějaký parametr vylepšil model je test nulové hypotézy:

H_0: složitější model nepřináší zlepšení

Testovací kritérium *F* vypočítáme podle vztahu:

$$F = \frac{\frac{S_r(1) - S_r(2)}{m_2 - m_1}}{\frac{S_r(2)}{n - m_2}}$$

H_0 zamítáme, pokud platí: $F > F_{1-\alpha}(m_2 - m_1; n - m_2)$.

$S_R(1)$ je reziduální součet čtverců jednoduššího modelu, $S_R(2)$ reziduální součet čtverců složitějšího modelu, *n* je počet pozorování, m_1 počet koeficientů jednoduššího modelu a m_2 počet koeficientů složitějšího modelu.

Kritickou hodnotu $F_{1-\alpha}$(volnost1; volnost2) spočítáme v Excelu

=F.INV.RT(α; m_2 - m_1; n - m_2).

Vícenásobný korelační koeficient

Předpokládejme vícerozměrný náhodný vektor ***x***, kde složky x_1, ..., x_m jsou vysvětlující proměnné a složka *y* vysvětlovaná proměnná.

Vícenásobný korelační koeficient $r_{y(1,...,m)}$ definuje míru lineární stochastické závislosti mezi náhodnou veličinou *y* a nejlepší lineární kombinací složek x_1, ..., x_m náhodného vektoru.

Vícenásobný korelační koeficient $r_{y(1,...,m)}$ lze vypočítat za vztahu

$$r_{y(1,..,m)} = \sqrt{1 - \frac{\det R}{\det R_{yy}}}$$

kde ***R*** je korelační matice, ve které r_{ij} jsou párové korelační koeficienty mezi proměnnými *i* a *j* a $\boldsymbol{R}_{ij}$ je matice vzniklá vypuštěním *i*-tého řádku a *j*-tého sloupce z korelační matice ***R***.

Vícenásobný korelační koeficient vrací funkce =LINREGRESE(), ve výstupní matici je jeho druhá mocnina ve třetím řádku a prvním sloupci.

Testování vícenásobného korelačního koeficientu

Předpokládejme, že vektor ***x*** má normální rozdělení a všechny jeho složky mají také normální rozdělení.
Pak platí, že náhodná veličina

$$F_r = \frac{(n-(m+1))\, r^2_{y(1,..,m)}}{m\left(1-r^2_{y(1,..,m)}\right)}$$

má *F*-rozdělení s *m* a *n* - (*m* + 1) stupni volnosti (*m* je počet nezávisle proměnných, *n* je počet změřených (*m* + 1)-tic).

Tabulkovou hodnotu *F*-rozdělení vypočítáme v Excelu pomocí funkce =F.INV.RT(prst;volnost1;volnost2), kde prst je hladina významnosti α (např. 0,05), volnost1 je *m* a volnost2 je *n* - (*m* + 1).
Pokud hodnota F_r větší než krtitická hodnota *F*-rozdělení, můžeme nulovou hypotézu ($r_{y(1,...,m)}$ je nulový) na dané hladině významnosti zamítnout.

Parciální korelační koeficienty

Parciální korelační koeficient umožňuje sledovat vztah mezi dvěma složkami při zkonstantnění ostatních složek.

$$r_{y\,i(1,..,i-1,i+1,..,m)} = \frac{(-1)^i \det R_{y\,i}}{\sqrt{\det R_{yy} \det R_{ii}}}$$

kde $\boldsymbol{R}_{ij}$ je matice vzniklá vypuštěním *i*-tého řádku a *j*-tého sloupce z korelační matice $\boldsymbol{R}$.

Parciální korelační koeficienty
2 nezávisle proměnné

V případě, že máme jen 2 vysvětlující proměnné (1. a 2. proměnná jsou vysvětlující, 3. proměnná je vysvětlovaná), jsou determinanty submatic 2x2 triviální a parciální korelační koeficienty lze spočítat ze vztahů:

$$r_{y1(2)} = \frac{r_{y1} - r_{y2} r_{12}}{\sqrt{(1 - r_{12}^2)(1 - r_{y2}^2)}}$$

$$r_{y2(1)} = \frac{r_{y2} - r_{y1} r_{12}}{\sqrt{(1 - r_{12}^2)(1 - r_{y1}^2)}}$$

Test parciálního korelačního koeficientu

Ukázali jsme, že korelační koeficient popisuje těsnost korelace mezi proměnnými x a y. Pokusme se nyní zjistit, jestli mezi proměnnými existuje vůbec nějaká (byť velmi slabá) souvislost.

H_0: Korelační koeficient je nulový (lineární závislost mezi *x* a *y* neexistuje).

H_1: Korelační koeficient je nenulový (závislost mezi *x* a *y* existuje).

Testovací kritérium:

$$t = \frac{r_{yi(1,..,i-1,i+1,..,m)}}{\sqrt{1 - r^2_{yi(1,..,i-1,i+1,..,m)}}} \sqrt{n-3}$$

Kritický obor: $|t| > t_{1-\alpha}(n-3)$

kde $t_{1-\alpha}(n-3)$ jsou kvantily Studentova rozdělení s $n-3$ stupni volnosti pro zvolenou hladinu významnosti α, které najdeme ve statistických tabulkách nebo vypočítáme pomocí funkce =T.INV.2T(α,*n*-3).

Příklad

Mějme experimentální data podle následující tabulky. x_1 a x_2 jsou vysvětlující proměnné, y je vysvětlovaná proměnná.

x_1	48,8	21,6	76,5	79,0	33,3	59,4	8,9	26,5	76,3	2,2
x_2	14,0	10,6	15,5	17,4	13,2	11,4	4,3	7,8	10,5	10,2
y	222,7	210,9	355,9	314,2	244,1	294,4	142,2	153,5	319,1	91,7

Nejprve vypočítáme korelační matici R - buď pomocí nástroje pro analýzu dat „korelace“ nebo pomocí funkce =CORREL():

	x_1	x_2	y
x_1	1,000	0,702	0,948
x_2	0,702	1,000	0,684
y	0,948	0,684	1,000

Příklad - korelační koeficienty

Z korelační matice $\boldsymbol{R}$ vybereme potřebné submatice a pomocí funkce =DETERMINANT() vypočítáme jejich determinanty.

det $\boldsymbol{R}_{yy}$ = 0,5076; det $\boldsymbol{R}_{y1}$ = 0,4677; det $\boldsymbol{R}_{y2}$ = -0,0189;

det $\boldsymbol{R}_{11}$ = 0,5324; det $\boldsymbol{R}_{22}$ = 0,1021; det $\boldsymbol{R}$ = 0,0515.

$$r_{y1(2)} = \frac{(-1)^2 \det R_{y1}}{\sqrt{\det R_{yy} \det R_{11}}} = \frac{0{,}4677}{\sqrt{0{,}5076 . 0{,}5324}} = 0{,}900$$

$$r_{y2(1)} = \frac{(-1)^3 \det R_{y2}}{\sqrt{\det R_{yy} \det R_{22}}} = \frac{0{,}0189}{\sqrt{0{,}5076 . 0{,}1021}} = 0{,}083$$

$$r_{y(12)} = \sqrt{1 - \frac{\det R}{\det R_{yy}}} = \sqrt{1 - \frac{0{,}0515}{0{,}5076}} = 0{,}948$$

Zatímco r_{y1} i r_{y2} jsou průkazné na hladině α = 0,05.

$r_{y1(2)}$ = 0,900 je průkazný na hladině α = 0,01,

$r_{y2(1)}$ = 0,083 není průkazný na hladině α = 0,05.

$r_{y(12)}$ = 0,948 je průkazný na hladině α = 0,01.

Příklad - vícenásobná regrese

Zpracujeme data pomocí funkce =LINREGRESE().
Výstupní tabulka funkce:

0.84007	2.78178	104.906
3.87411	0.50871	34.5174
0.89871	31.3894	#N/A
31.0554	7	#N/A
61197.3	6897.04	#N/A

Regresní rovnice má tvar: $y = 2.8x_1 + 0.8x_2 + 105$

Porovnáním parametrů rovnice v prvním řádku a jejich směrodatných odchylek ve druhém řádku vidíme, že směrnice závislosti na x_1 je významně různá od nuly, zatímco směrnice u x_2 může být nulová. Zjistili jsme totéž, co z parciálních korelačních koeficientů - totiž že *y* závisí na x_1 a nezávisí na x_2.

Příklad - vícenásobná regrese

Vyzkoušejme, jestli by model, ve kterém *y* závisí na pouze x_1 a nezávisí na x_2 nebyl stejně dobrý jako model se dvěmi vysvětlujícími proměnnými.

Zpracujeme data pomocí funkce =LINREGRESE().
Výstupní tabulka funkce:

2.85908	111.215
0.34062	17.4302
0.89803	29.4605
70.4568	8
61151	6943.37

$$F = \frac{\frac{S_r(1) - S_r(2)}{p_2 - p_1}}{\frac{S_r(2)}{n - p_2}} = \frac{\frac{6943 - 6897}{3 - 2}}{\frac{6897}{10 - 3}} = 0.047$$

$F_{0.95}(3 - 2; 10 - 3) = 5.59$.

Regresní rovnice $y = 2.9x_1 + 111$ tedy není horší než $y = 2.8x_1 + 0.8x_2 + 105$.