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***).

Hendl, J. (2014). *Statistika v aplikacích.* Praha: Portál.

Metody zpracování naměřených závislostí

1. **Funkce, která v uzlových bodech x_i nabývá zadané hodnoty y_i.**

   - **Interpolace** – pro x z intervalu (x_{min}, x_{max})
   - **Extrapolace** – pro x mimo interval (x_{min}, x_{max})

2. **Funkce, která uzlovými body obecně neprochází.**

   - **Aproximace** - neznáme typ funkční závislosti $y = f(x)$
   - **Regrese** - známe typ funkční závislosti $y = f(x)$
     - cílem je určit z naměřených dat parametry $f(x)$ tak, aby funkce co nejlépe vystihla naměřenou závislost

   Při regresi je funkční závislost předem daná např. teoretickým vztahem, kdežto aproximační funkce bývá volena bez hlubšího fyzikálního významu.

Hendl, J. (2014). *Statistika v aplikacích.* Praha: Portál.

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.

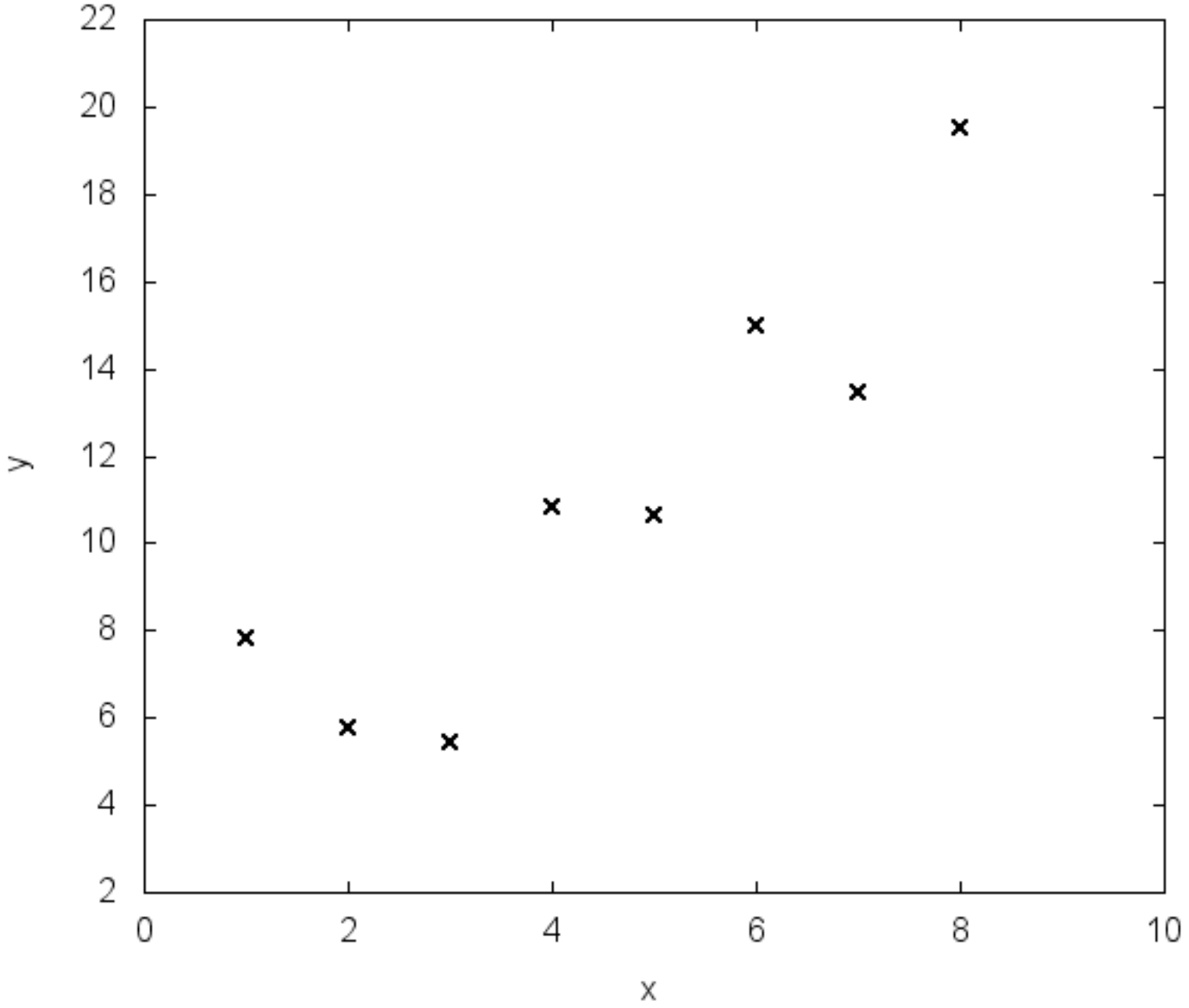
- Pokud by při měření nevznikaly náhodné chyby, pak by všechny body $[x_i, y_i]$ ležely na křivce $y = f(x)$.
- Ve skutečnosti však platí $y_i = f(x_i) + \varepsilon_i$, kde ε_i je náhodná chyba i-tého měření.

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 $[x_i, y_i]$.

(Rektorys, 1981)

Rektorys, K. Přehled užité matematiky. 4., nezm. vyd. (1981) Praha: SNTL .

Metoda nejmenších čtverců



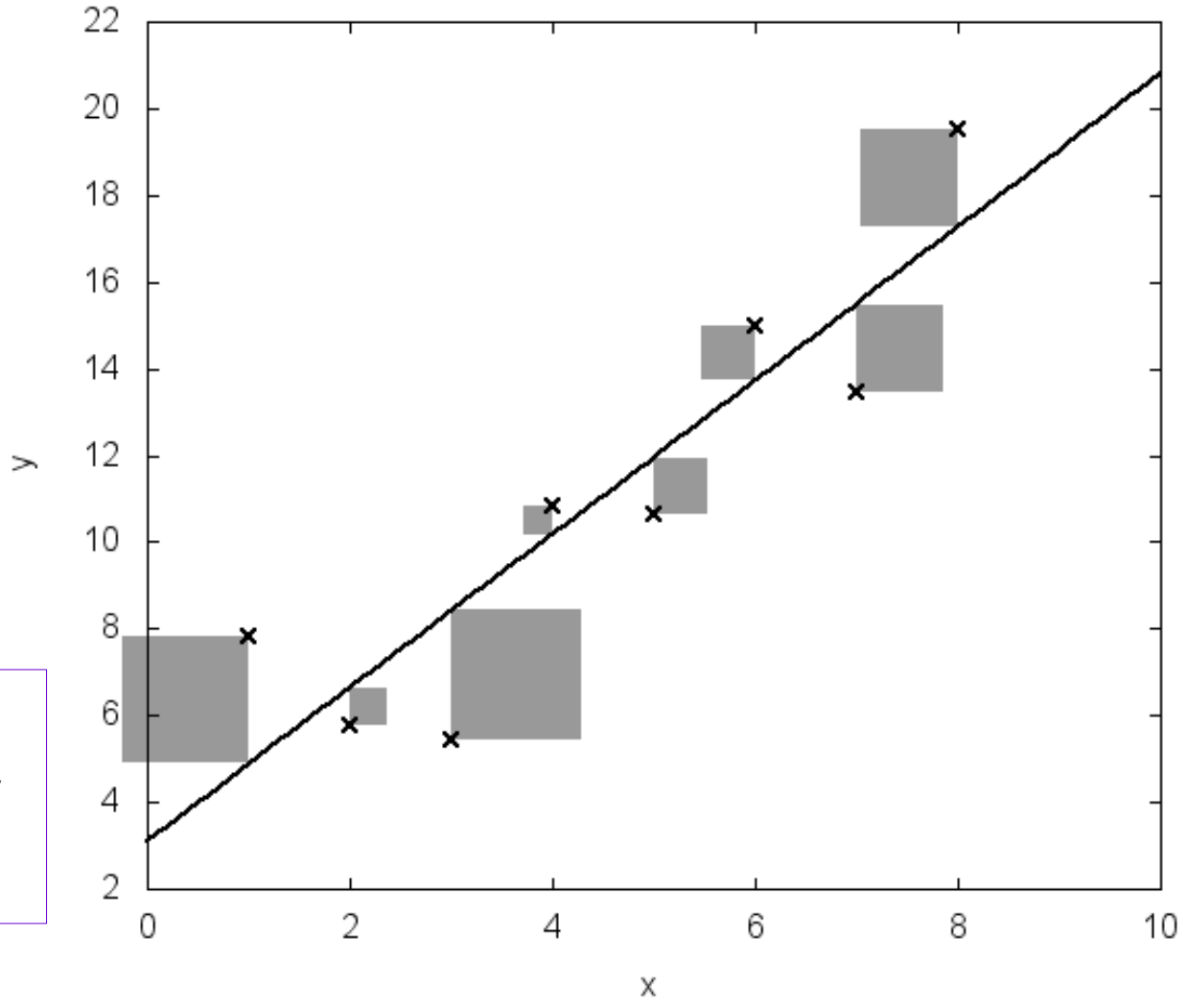
Metoda nejmenších čtverců

Hledáme kritérium „přiléhavosti“ regresní křivky k experimentálním bodům.

Nejvěrohodnější je tzv. **reziduální (zbytkový) součet „čtverců“**.

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

(Rektorys, 1981)



Rektorys, K. Přehled užité matematiky. 4., nezm. vyd. (1981) Praha: SNTL .

Metoda nejmenších čtverců

Hledáme 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 $[x_i, y_i]$.

Tzn., že hledáme tedy takové parametry *a*, *b*, *c*, které by **minimalizovaly zbytkový součet čtverců**.

$$\frac{\partial S_{\text{resid.}}}{\partial a} = 0, \qquad \frac{\partial S_{\text{resid.}}}{\partial b} = 0, \qquad \frac{\partial S_{\text{resid.}}}{\partial c} = 0, \ldots$$

Nejběžnější regresní metoda se proto nazývá

metoda nejmenších čtverců.

(Rektorys, 1981)

Rektorys, K. Přehled užité matematiky. 4., nezm. vyd. (1981) Praha: SNTL .

Lineární regrese 1 proměnné

Ukažme si proložení přímky $y = a + bx$ naměřenými body $[x_i, y_i]$ metodou nejmenších čtverců.

Reziduální součet čtverců odchylek je

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

Hledáme takové $\hat{a}$ a $\hat{b}$, aby rez. součet čtverců byl minimální.

$$\frac{\partial S_{\text{resid.}}}{\partial \hat{a}} = -2 \sum_{i=1}^{N} (y_i - \hat{a} - \hat{b}\, x_i) = -2 \left(\sum_{i=1}^{N} y_i - N\, \hat{a} - \hat{b} \sum_{i=1}^{N} x_i \right) = 0$$

$$\frac{\partial S_{\text{resid.}}}{\partial \hat{a}} = -2 \sum_{i=1}^{N} (y_i - \hat{a}\, x_i - \hat{b})\, x_i = -2 \left(\sum_{i=1}^{N} x_i\, y_i - \hat{a} \sum_{i=1}^{N} x_i - \hat{b} \sum_{i=1}^{N} x_i^2 \right) = 0$$

(Rektorys, 1981)

Rektorys, K. Přehled užité matematiky. 4., nezm. vyd. (1981) Praha: SNTL .

=LINREGRESE

OOa Calc i MS Excel nabízí několik funkcí počítajících parametry lineární regrese.

Nejlepší je `=LINREGRESE(DataY;DataX;Typ;Parametr)`

Výstup funkce LINREGRESE:

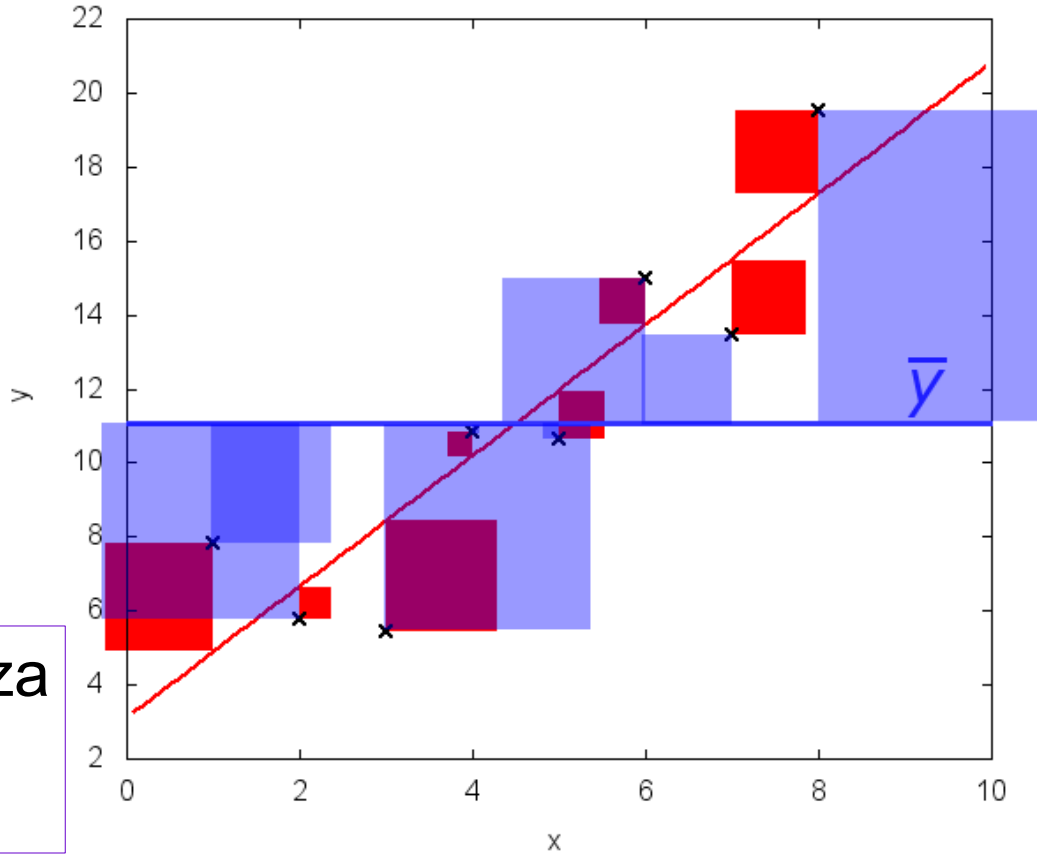
(odhad parametru *b*)	(odhad parametru *a*)
(odhad chyby parametru *b*)	(odhad chyby parametru *a*)
(koeficient determinace*)	(chyba odhadu)
*F (F-statistika**)*	*df (počet stupňů volnosti)*
S_t - S_r (rozdíl celkové a reziduální sumy čtverců odchylek*)	S_r *(reziduální suma čtverců odchylek*)*

Regresní model

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

(Rektorys, 1981)

Rektorys, K. Přehled užité matematiky. 4., nezm. vyd. (1981) Praha: SNTL .

Korelační analýza

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

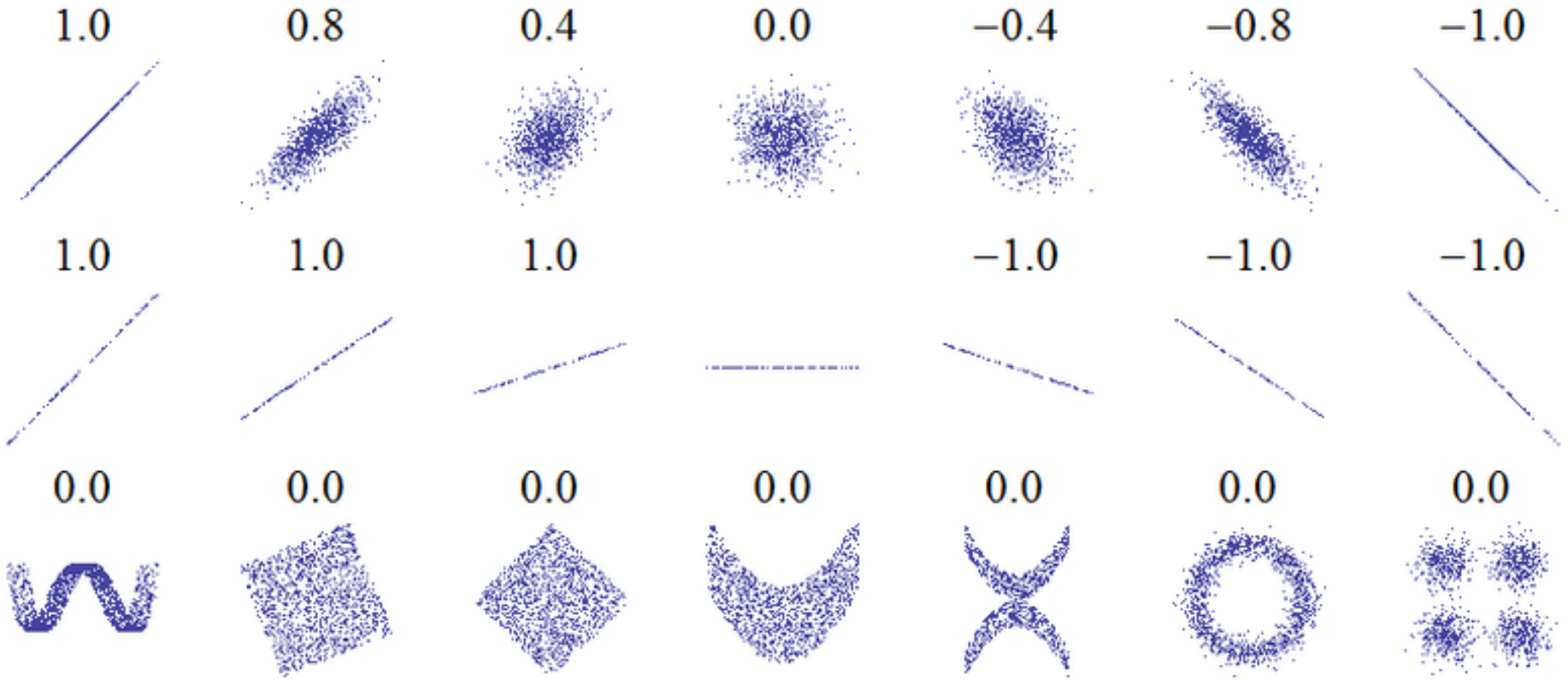
Při testování existence lineárního vztahu mezi proměnnými se používá odmocnina z koeficientu determinace r^2, která se nazývá **Pearsonův korelační koeficient *r***.
Korelační koeficient se počítá pomocí funkce =CORREL()
nebo ze vztahu:
(Meloun & Militký, 2013)

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

Meloun, M., Militký, J. (2013). *Kompendium statistického zpracování dat*. Praha: Karolinum.

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.

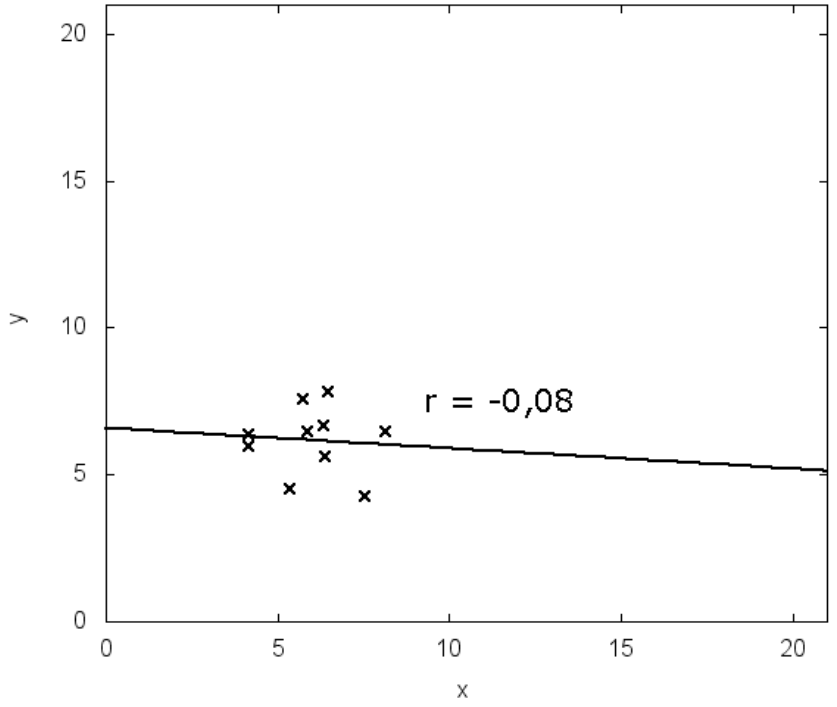


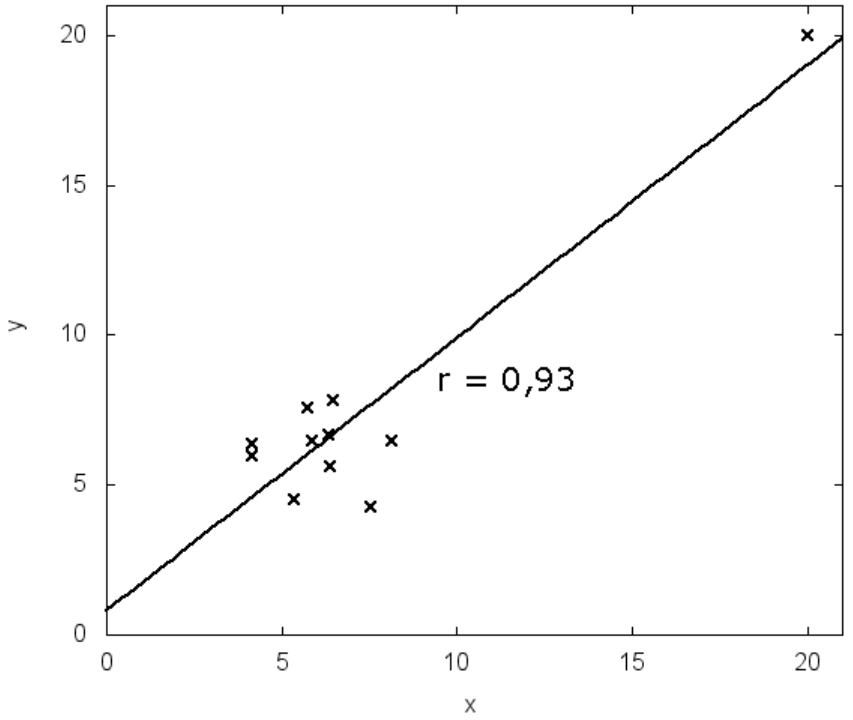
(Wikipedie, 2017)

Korelace (2017, December 23). Retrieved from https://cs.wikipedia.org/wiki/Korelace

Záludnosti korelačního koeficientu

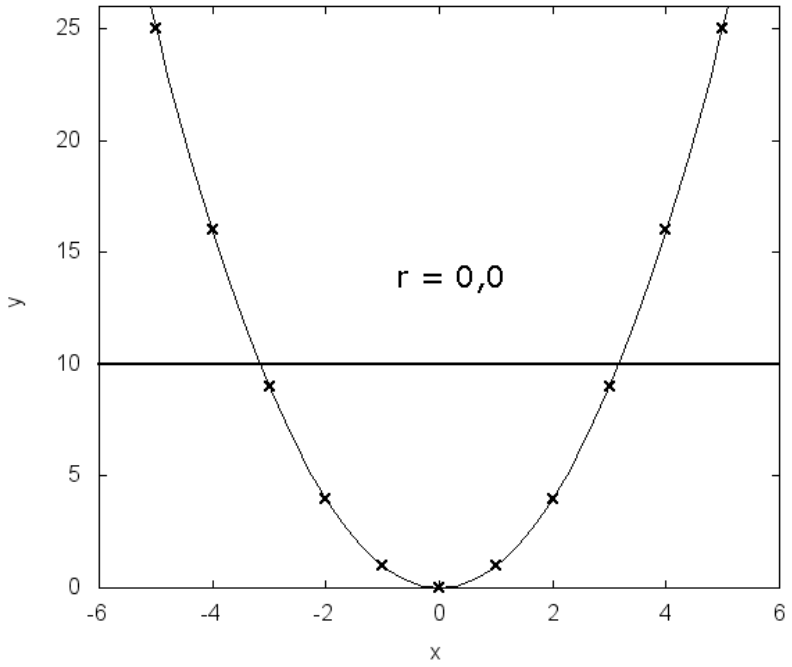
1) Je citlivý na odlehlé hodnoty.

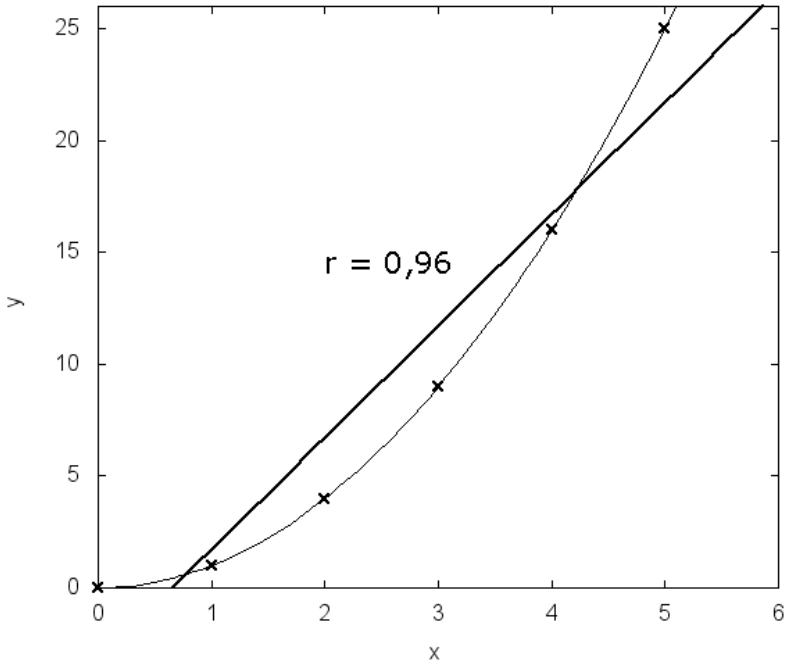




Záludnosti korelačního koeficientu

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). (Meloun & Militký, 2013)

Testovací kritérium:

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

Kritickou hodnotou $t_{1-\alpha}(N\text{-}2)$ 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)

Meloun, M., Militký, J. (2013). *Kompendium statistického zpracování dat*. Praha: Karolinum.