Základy zpracování geologických dat

R. Čopjaková

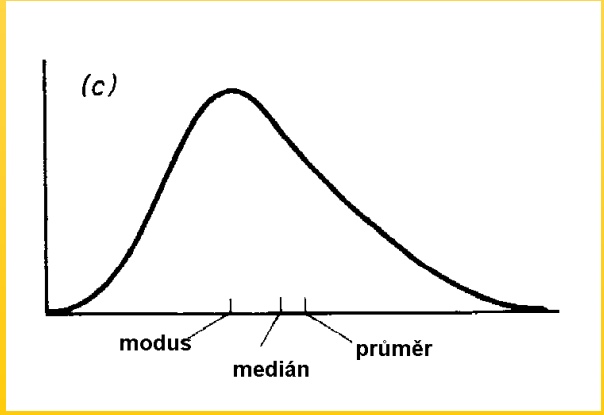
Základní charakteristiky náhodné veličiny

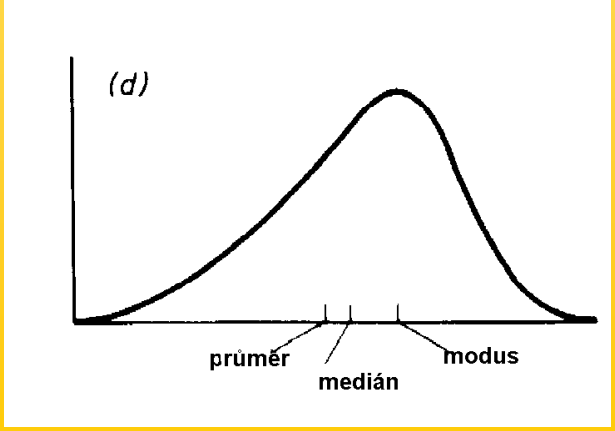
- **Koeficient šikmosti** je charakteristika rozdělení náhodné veličiny, která popisuje jeho nesymetrii
- *Šikmost* označuje stupeň asymetričnosti rozdělení veličiny kolem střední hodnoty
- Nulová šikmost - hodnoty náhodné veličiny jsou rovnoměrně rozděleny vlevo a vpravo od střední hodnoty – symetrické rozdělení
- Výběrový koeficient šikmosti je definován vzorcem

$$\frac{n}{(n-1)(n-2)}\sum\left(\frac{x_i-\bar{x}}{s}\right)^3$$

- **>0 pozitivně šikmé** **<0 negativně šikmé**

  rozdělení má tzv. *pravý ocas*





Regrese a korelace – základní termíny

Regrese versus korelace

- Regrese popisuje vztah = závislost dvou a více kvantitativních proměnných formou funkční závislosti
- Korelace měří těsnost (sílu) vztahu = závislosti mezi dvěma proměnnými

- Liší se chápání proměnných u obou metod?

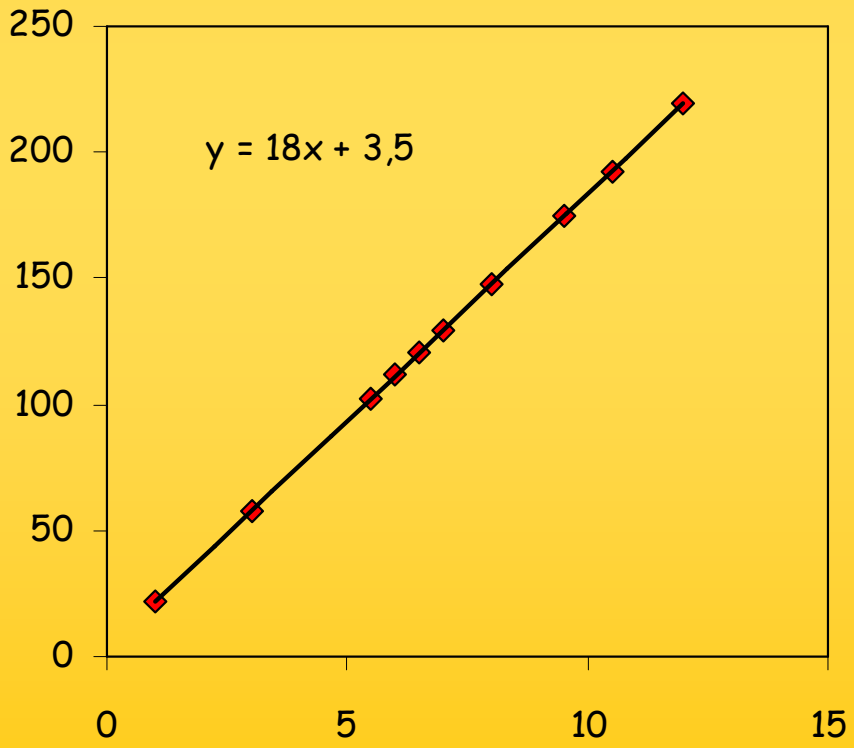
  U regrese lze rozlišit, která proměnná závisí na které, čili rozlišuje se tzv. nezávislá (x) a závislá proměnná (y); nezávislá proměnná *x* je na horizontální ose *x*, závislá proměnná *y* je na vertikální ose *y*.

  U korelace se nerozlišují proměnné na závislou a nezávislou

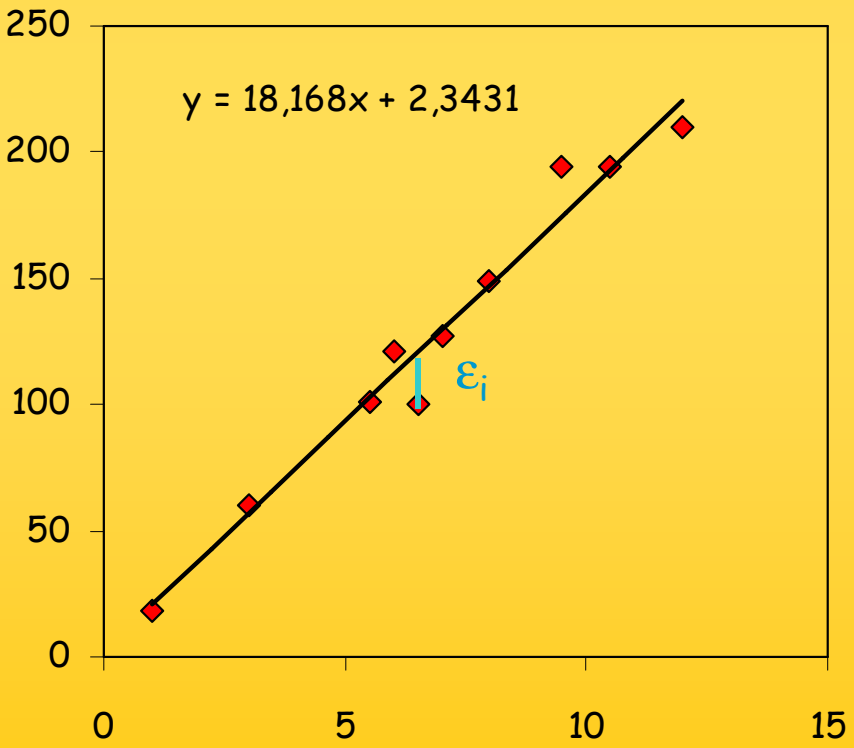
- Regresní analýza – sestavení modelu, kterým lze formálně popsat vztahy (pokud existují)
- Regresní model – vztah jedné proměnné označované jako závisle proměnná (vysvětlovaná) k dalším proměnným, které se označují jako nezávislé (vysvětlující)

Závislost dvou souborů dat

- Funkční /deterministická závislost/: vzájemný vztah mezi proměnnými daný jednoznačně y=f(x)
- Statistická závislost: vyjadřuje, že mezi proměnnými neexistuje jednoznačný vztah, tedy Y=f(X) + ε, kde ε jsou pozorované náhodné odchylky od modelu

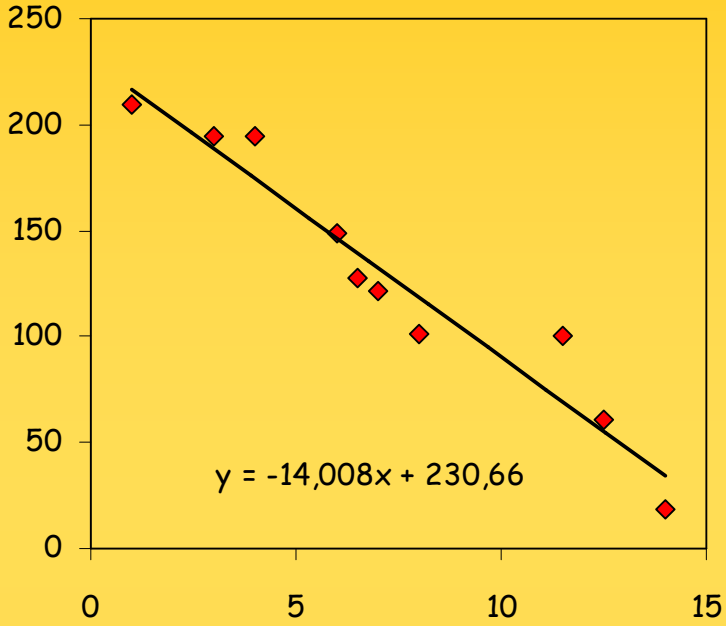


funkční závislost

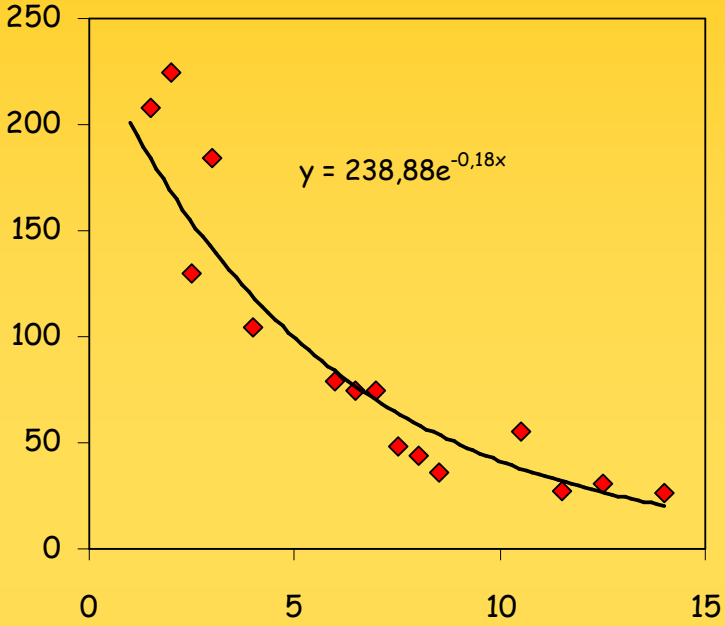


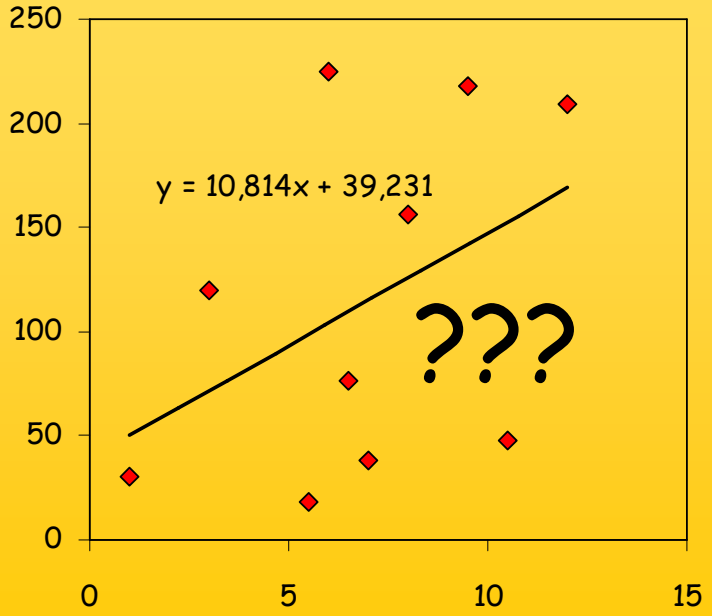
stochastická závislost

závislost lineární



závislost exponenciální





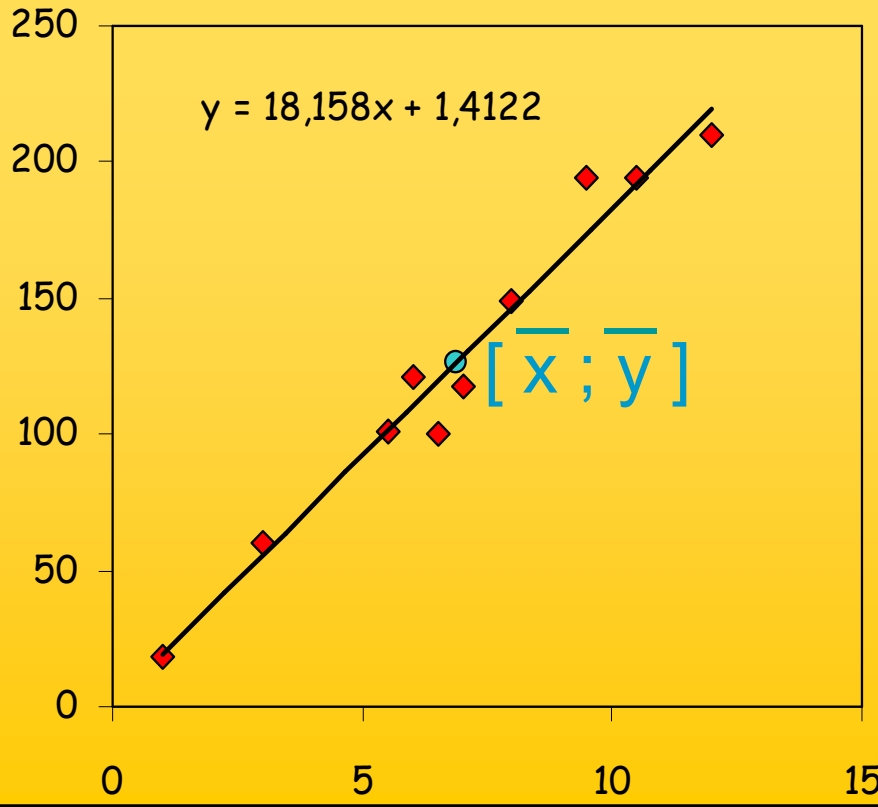
závislost neexistuje, nemá smysl prokládat regresní funkci

Jednoduchý lineární regresní model:

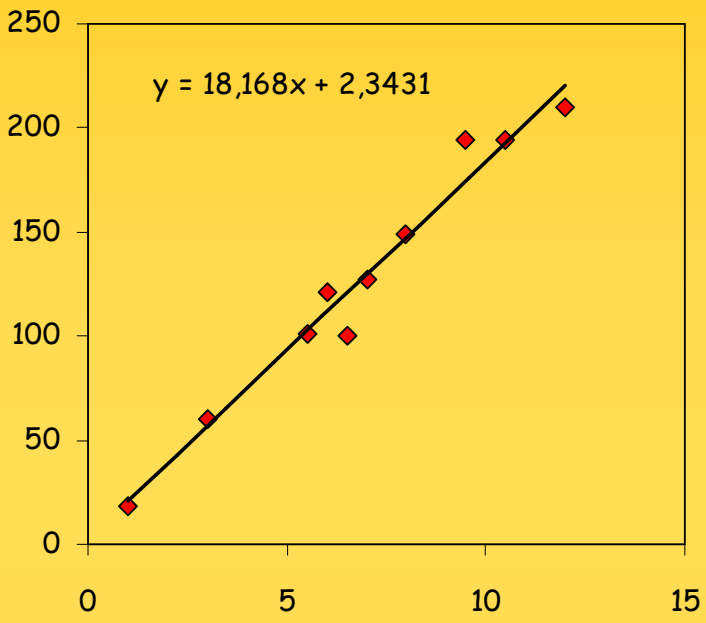
- nejjednodušší případ regrese:
  - „jednoduchá“ = pouze 1 nezávislá a 1 závislá proměnná
  - „lineární“ = závislost *y* na *x* vyjadřujeme přímkou
- Některé předpoklady lineární regrese:

  1. homogenní rozptyl: všechna *Y* mají stejnou rozptýlenost

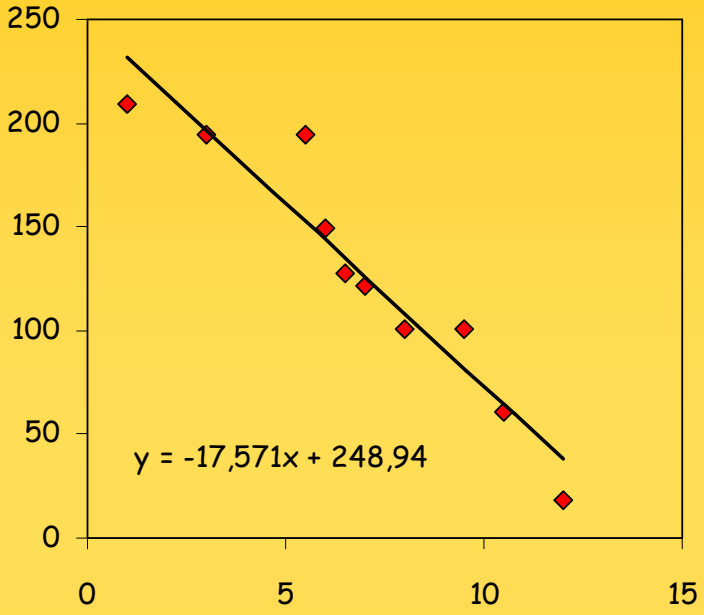
  2. linearita: střední hodnoty obou proměnných X a Y leží na regresní přímce



lineární závislost přímá



lineární závislost nepřímá

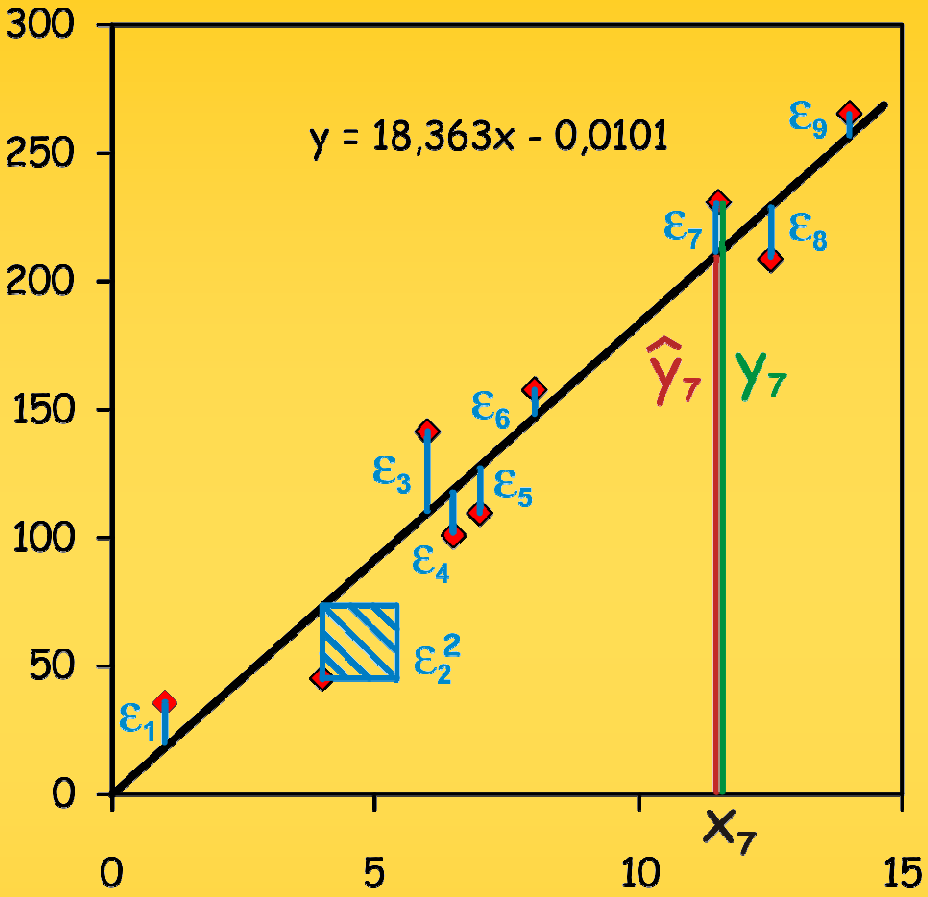


Regresní analýza

- **napozorovaná (empirická) hodnota** - hodnota proměnné, kterou jsme získali jako výsledek pozorování (měření, vážení atd.).

  značíme ji y
- **odhadnutá (teoretická) hodnota** - hodnota proměnné, kterou jsme získali jako výsledek modelování této proměnné.

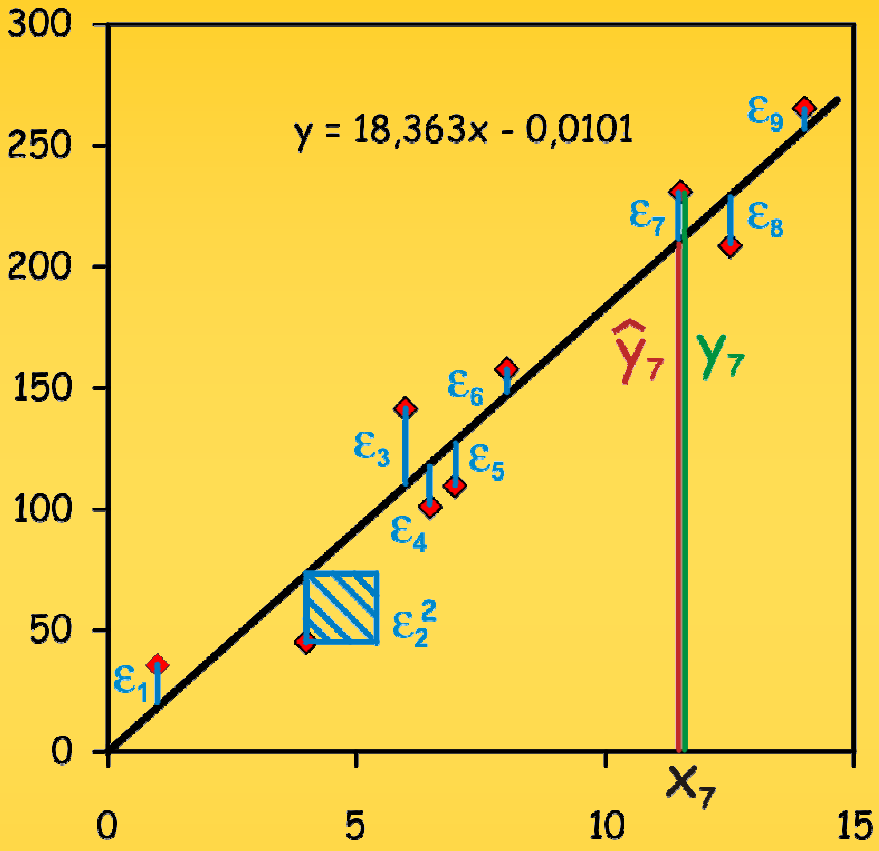
  značíme ji $\hat{y}$
- **reziduum** - rozdíl mezi napozorovanou a odhadnutou hodnotou. Reziduum značíme symbolem ε a v příslušném bodě počítáme jako rozdíl empirické hodnoty a teoretické. Reziduum tedy můžeme chápat jako velikost chyby, které se v příslušném bodě při odhadu dopouštíme.



- **Jak nalézt funkci, která „nejlépe“proloží naše data?**

Jak nalézt funkci, která „nejlépe"proloží naše data?

- postup odhadu parametrů regresní funkce, který dává nejmenší hodnoty reziduí (tedy „nejmenší chybu") a to najednou ve všech odhadovaných bodech.
- Nestačí pouze rezidua sečíst - vlivem kladných a záporných znamének u jednotlivých hodnot by mohlo dojít k tomu, že součet reziduí bude nulový, přestože jednotlivá rezidua (tedy jednotlivé chyby) jsou veliké.
- Z celé škály vyrovnávacích kritérií se jako nejpoužívanější (ne však vždy nejvhodnější) jeví tzv. **metoda nejmenších čtverců** = **musí platit, aby (reziduální) součet čtverců odchylek skutečných od očekávaných hodnot byl minimální**



$$S_e^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \min$$

Metoda nejmenších čtverců pro přímku

- Hledáme minimum výrazu $S_e^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$
- Kde $Y_i = b_o + b_1 X_i + \varepsilon_i$ a $\hat{Y}_i = b_0 + b_1 X_i$
- Po dosazení obdržíme $S_e^2 = \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2$
- Hodnota veličiny S závisí na volitelných hodnotách b_0 a b_1 a je to tedy funkce dvou proměnných. Její extrém (minimum) se najde nulováním parciálních derivací podle těchto proměnných. Zderivujeme výraz parciálně podle b_0 a b_1 a dostaneme soustavu normálních rovnic

$$S_e^2 = \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2 = \min$$

$$2\sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)(-1) = 0$$

$$2\sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)(-X_i) = 0$$

- Z těchto rovnic můžeme po příslušných úpravách vyjádřit parametr b_1 - tedy směrnici regresní přímky

$$b_1 = \frac{n\sum x_i y_i - \left(\sum x_i\right)\left(\sum y_i\right)}{n\sum x_i^2 - \left(\sum x_i\right)^2} = \frac{Q_{(x,y)}}{Q_{(x)}}$$

- Z rovnice lineární funkce potom dopočteme parametr b_0, za předpokladu že x a y leží na regresní přímce

$$b_1 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma (x_i - \bar{x})^2} = \frac{cov_{xy}}{s_x^2}$$

$$b_0 = \bar{y} - b_1 \bar{x} = \frac{\sum y_i - b_1 \sum x_i}{n}$$

Reziduální rozptyl

Reziduální rozptyl - velikost chyb ε je popsána rozptylem $\sigma^2_{(y-\hat{y})}$ nebo σ^2_e odchylek od regresní přímky

$$s^2_{(y-\hat{y})} = s^2_e = \frac{1}{n-2}\sum (y_i - \hat{y}_i)^2 = \frac{Q(e)}{n-2}$$

Kovariance

- Nástroj kovariance můžete použít k testování závislosti dvou sad dat (u lineární závislosti dvou proměnných s přibližně normálním rozdělením).

- Závislost znamená, že velké hodnoty v jedné sadě odpovídají velkým hodnotám ve druhé sadě (kladná kovariance), nebo že velké hodnoty v jedné sadě odpovídají malým hodnotám ve druhé sadě (záporná kovariance). Teoreticky se pohybuje od $-\infty$ do $+\infty$

- Pokud jsou hodnoty v obou množinách nezávislé => blízká nule.

- nelze usuzovat na sílu vztahu, pouze na směr působení + přímé – nepřímé
- Kovariance je ≤ součinu směrodatných odchylek proměnné X a Y

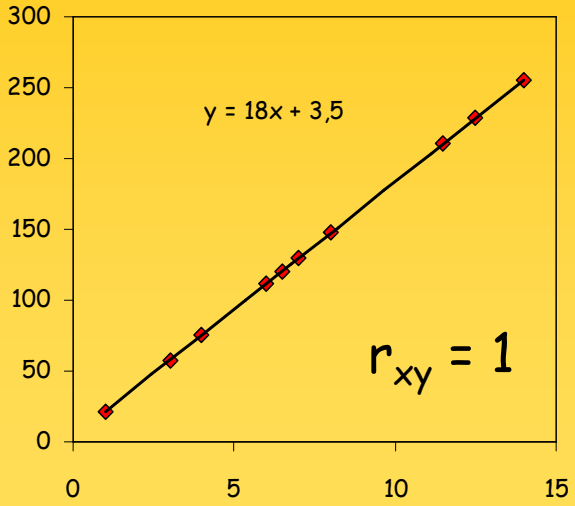
$$S_{xy} = \mathrm{cov}(X,Y) = \mathrm{cov}(Y,X) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n}$$

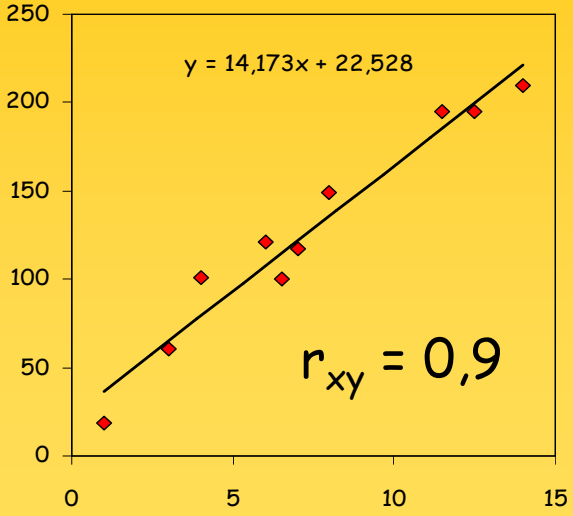
Pearsonův korelační koeficient

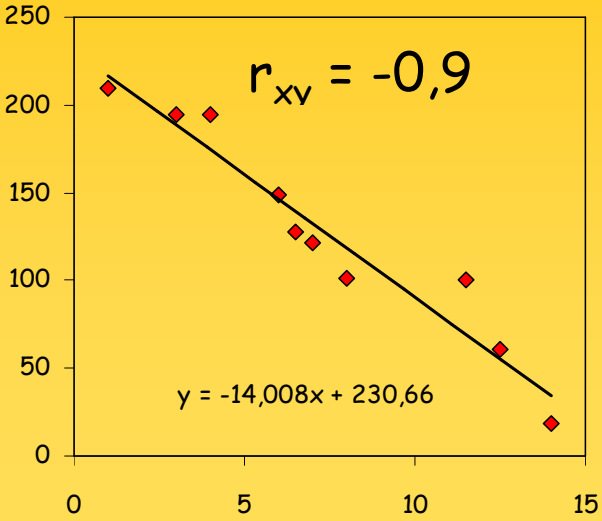
- Tzv. standardizovaná kovariance
- určení síly vztahu mezi proměnnou X a Y (s přibližně normálním rozdělením) bez nutnosti definovat závislou a nezávislou veličinu (pouze pro lineární závislost)
- Hodnota korelačního koeficientu –1 značí zcela nepřímou (funkční) závislost, tedy čím více se zvětší hodnoty v první skupině znaků, tím více se zmenší hodnoty v druhé skupině znaků.
- Hodnota korelačního koeficientu +1 značí zcela přímou (funkční) závislost.
- Pokud je korelační koeficient roven 0, pak mezi znaky není žádná statisticky zjistitelná závislost,
- Korelační koeficient může nabývat hodnot <-1;+1>

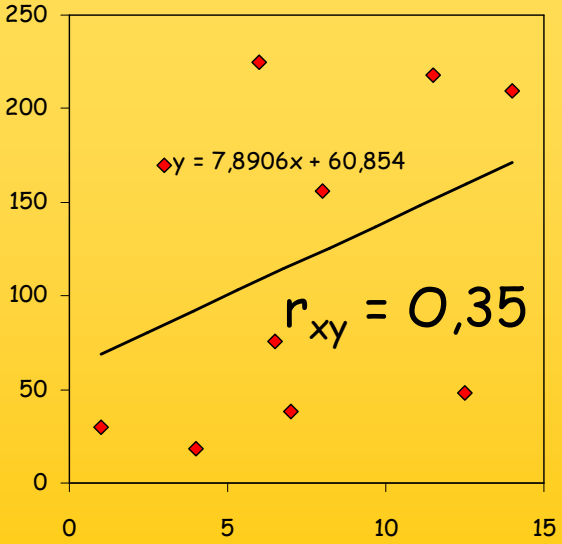
$$r_{yx} = r_{xy} = \frac{s_{xy}}{\sqrt{s_x^2 s_y^2}} = \frac{S_{xy}}{S_x S_y}$$

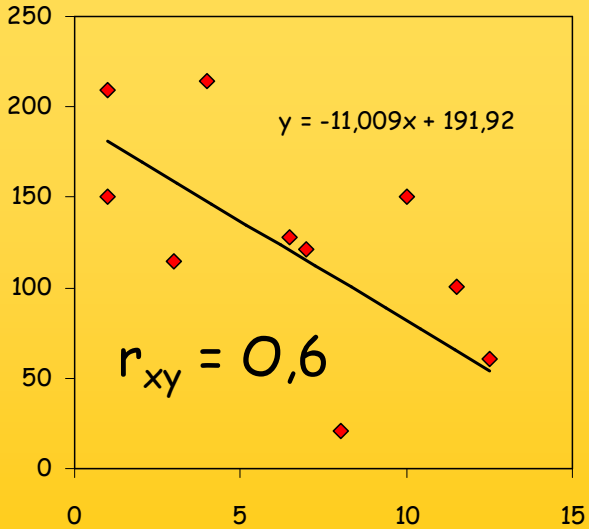
Pearsonův korelační koeficient

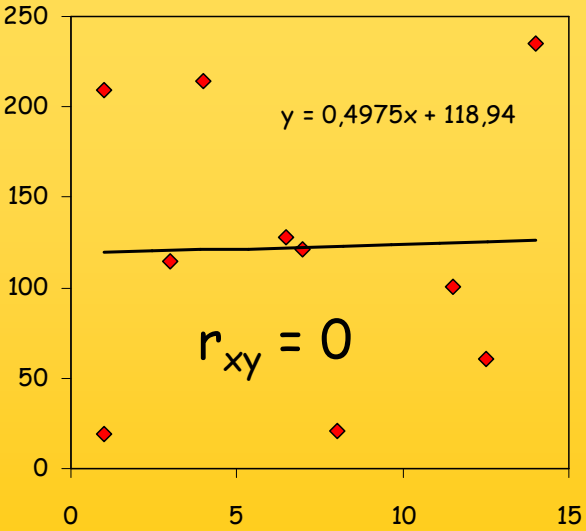












Spearmanův koeficient pořadové korelace

- Univerzální – nejen pro lineární závislost
- Chci-li spočítat hodnotu Spearmanova koeficintu, převedu naměřená data pro soubor X_i a Y_i na pořadové hodnoty X_{ip} a Y_{ip}.
- Spočtu rozdíly v pořadí jednotlivých párů $d_i = X_{ip} - Y_{ip}$, které použiji při výpočtu tohoto koeficientu

$$SR = 1 - \frac{6\sum d_i^2}{n\left(n^2 - 1\right)}$$

Spearmanův koeficient pořadové korelace

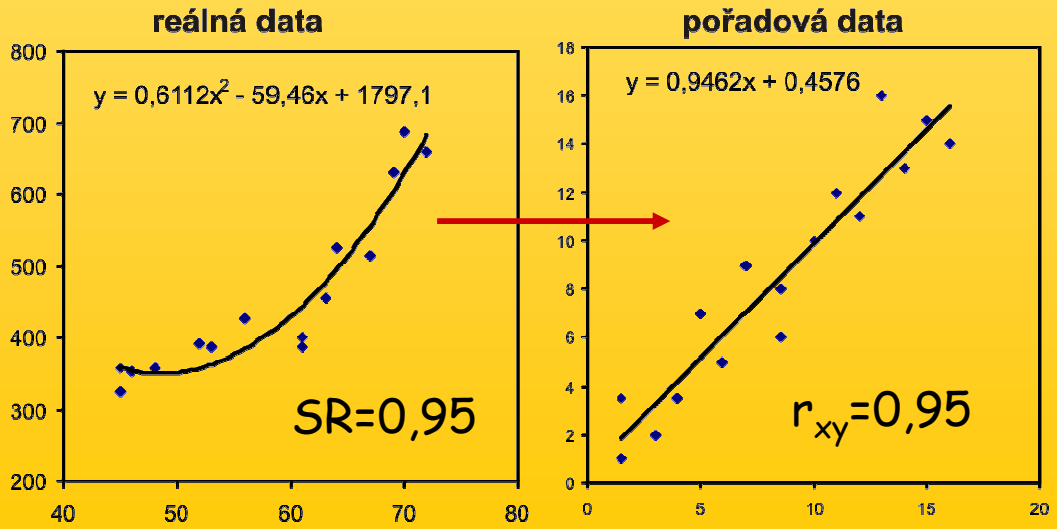
Reálná naměřená data s nelineární závislostí převedu na pořadové hodnoty a spočtu Spearmanův koeficient pořadové korelace

n	X	Y	rank X	rank Y	úprava X	úprava Y	di	di^2
1	45	359	1	3	1,5	3,5	-2	4
2	45	326	1	1	1,5	1	0,5	0,25
3	46	354	3	2	3	2	1	1
4	48	359	4	3	4	3,5	0,5	0,25
5	52	392	5	7	5	7	-2	4
6	53	386	6	5	6	5	1	1
7	56	426	7	9	7	9	-2	4
8	61	401	8	8	8,5	8	0,5	0,25
9	61	387	8	6	8,5	6	2,5	6,25
10	63	455	10	10	10	10	0	0
11	64	526	11	12	11	12	-1	1
12	67	515	12	11	12	11	1	1
13	69	630	14	13	14	13	1	1
14	72	659	16	14	16	14	2	4
15	70	689	15	15	15	15	0	0
16	68	708	13	16	13	16	-3	9
							suma	37

$$SR = 1 - \frac{6\sum d_i^2}{n\left(n^2 - 1\right)}$$

$$= 1 - 6*37/16(16^2-1) = 0,95$$

$$r_{xy} = \frac{Sxy}{SxSy} = \frac{20,05}{4,60*4,61} = 0,95$$



Spočtu-li pearsonův koeficient korelace pro pořadové hodnoty (lineární závislost), bude velice blízký hodnotě Spearmanova koeficientu pořadové korelace pro naměřené hodnoty proměnné X a Y