Základy zpracování geologických dat

regresní a korelační analýza

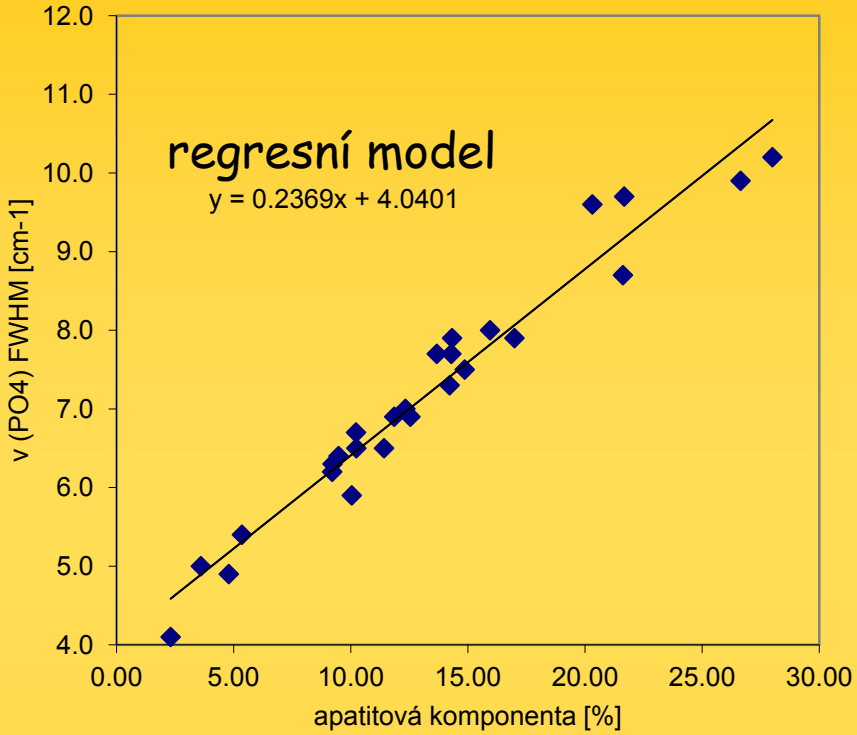
R. Čopjaková

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 – kvantifikuje jak hodně jsou hodnoty blízké ideálnímu regresnímu modelu

- Regresní analýza – sestavení modelu (regresní funkce), 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)



korelační koeficient
$r_{xy} = 0,98$

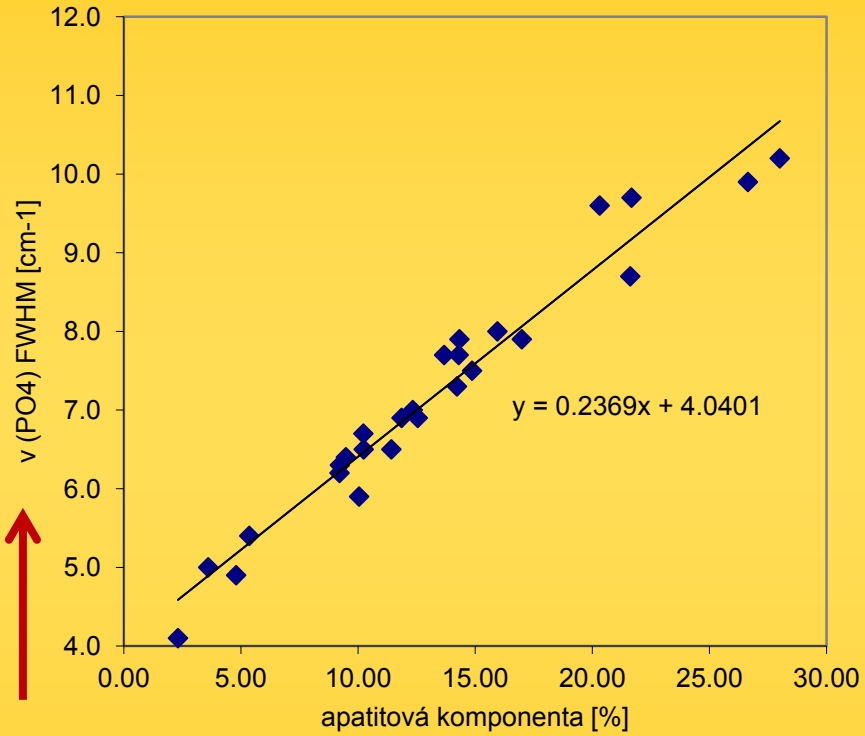
Regrese a korelace – základní termíny

- 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*.

pokud bychom přehodili x a y, získáme jinou rovnici regresní funkce

U korelace se nerozlišují proměnné na závislou a nezávislou $r_{xy}=r_{yx}$



Závisle proměnná na ose y

Nezávisle proměnná na ose x

Závislost dvou souborů dat

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

funkční závislost

stochastická závislost

závislost lineární

závislost exponenciální

???

závislost neexistuje, nemá smysl prokládat regresní funkci

pozor – regresní funkci lze vždy spočítat, (i když nemá smysl, protože žádná závislost mezi soubory dat není)

lineární regresní model

Lineární funkce: $Y = b_1X + b_0$

b_1 směrnice přímky, udává sklon

b_0 průsečík s osou y

lineární závislost přímá

směrnice přímky je kladná

lineární závislost nepřímá

směrnice přímky je záporná

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

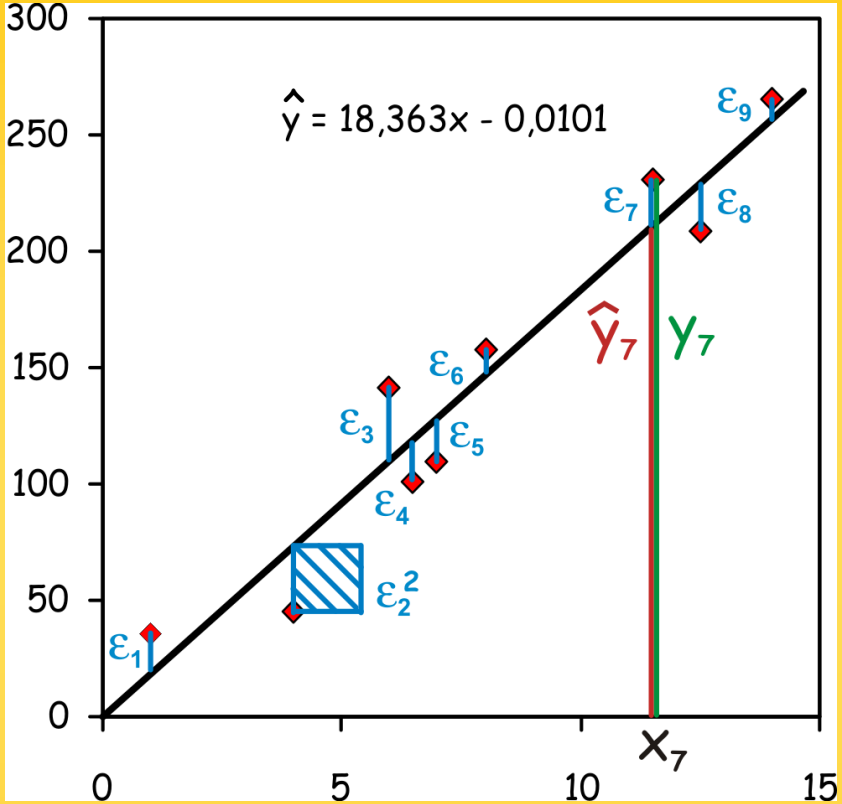
$[\overline{x}; \overline{y}]$

lineární regresní model

- **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.

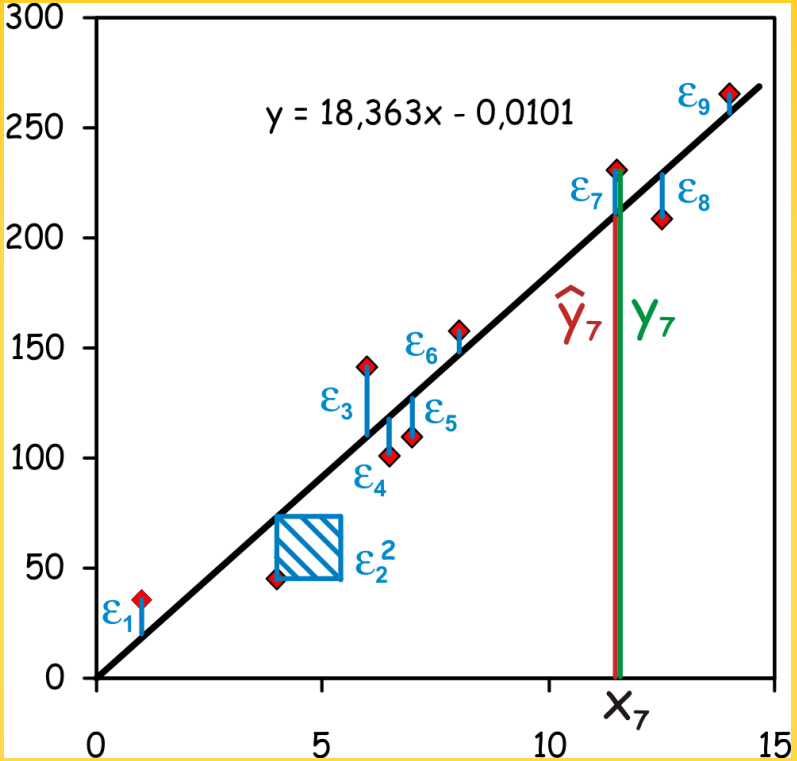
  značíme ji $\hat{y}$
- **reziduum** - rozdíl mezi napozorovanou a odhadnutou hodnotou. Reziduum značíme symbolem ε a v příslušném bodě počítáme jako rozdíl empirické hodnoty a teoretické. Reziduum tedy můžeme chápat jako velikost chyby, které se v příslušném bodě při odhadu dopouštíme.



- **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**



$$S_e^2 = \sum_{i=} \qquad = \min$$

Metoda nejmenších čtverců pro přímku

- Hledáme minimum výrazu $S_e^2 = \sum_{i=} = min$
- Kde $Y_i = b_o + b_1 X_i + \varepsilon_i$ a $\hat{Y}_i = \quad +$
- Po dosazení obdržíme $S_e^2 = \sum_{i=}$
- 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 $S_e^2 = \sum_{i=} = \min$

$$2\sum_{i=} \quad - \quad =$$

$$2\sum_{i=} \quad - \quad =$$

- Z těchto rovnic můžeme po příslušných úpravách vyjádřit parametr b_1 - tedy směrnici regresní přímky
- Z rovnice lineární funkce potom dopočteme parametr b_0, za předpokladu že $\bar{x}$ a $\bar{y}$ leží na regresní přímce

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

Kovariance, cov(x,y), S_{xy}

- 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 -∞ do + ∞

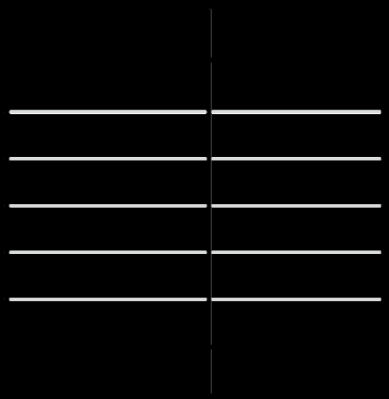
- Pokud jsou hodnoty v obou množinách nezávislé => blízká nule.

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

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

Lineární regresní model

Bylo provedeno 6 měření indexu lomu roztoku NaCl ve vodě pro koncentrace NaCl 2, 4, 6, 8 a 10 % a pro destilovanou vodu. Teplota byla konstantní. Vyšetři závislost indexu lomu na koncentraci NaCl v roztoku.

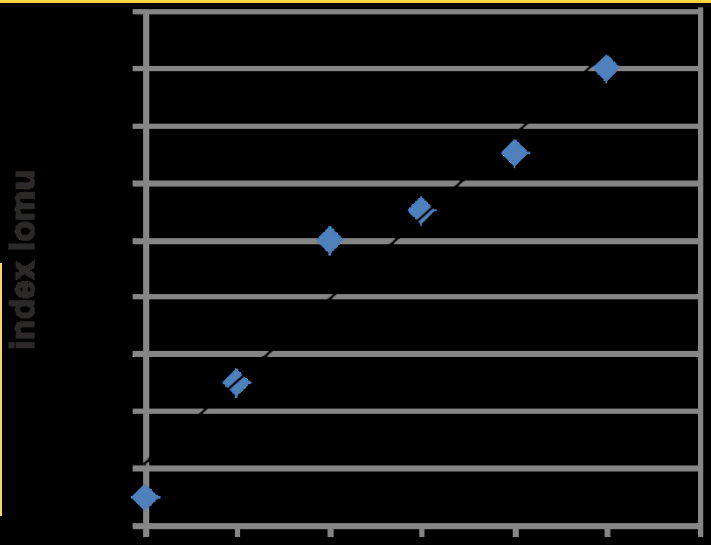


11,7

Funkce v excelu

kovariance (COVARIANCE.P/ COVARIANCE.S)

rozptyl (VAR.P/VAR.S)

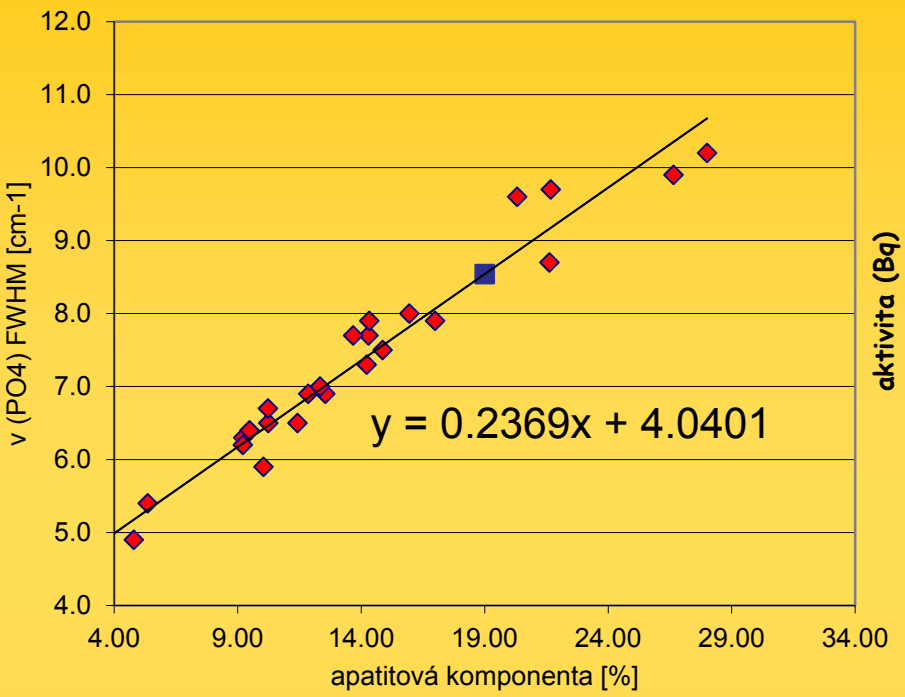


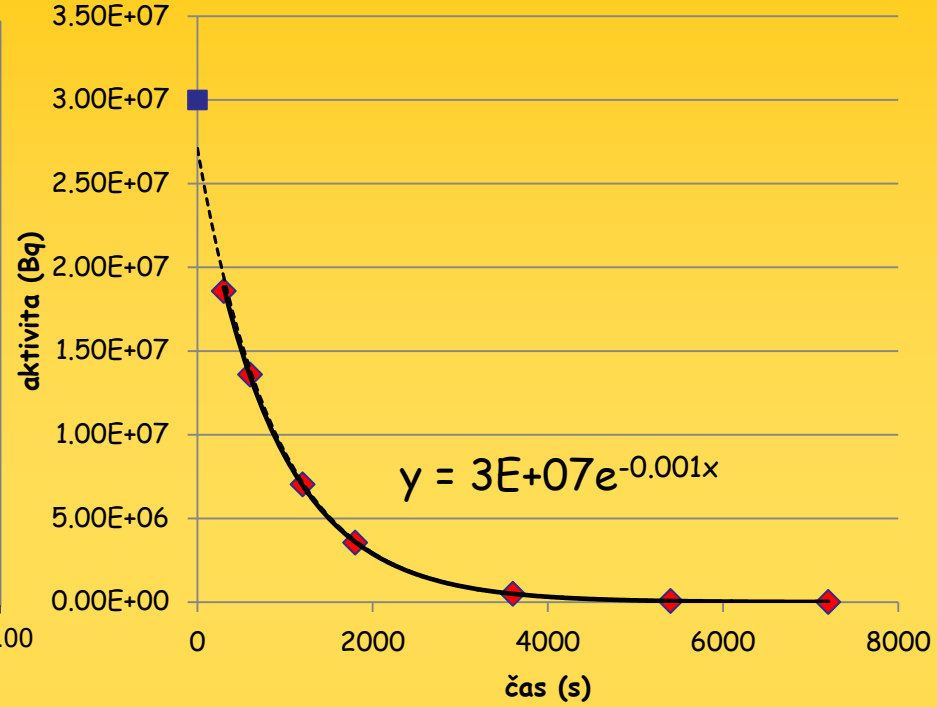
1) dopočtením ze vzorců

2) nebo přidáním spojnice trendu do grafu (typ trendu - lineární ; a přídat rovnici do grafu)

V případě nelineární závislosti nutno vybrat typ regresního modelu - funkce

Regresní analýza





◆ experimentální data

■ intrapolace - dopočtená šířka v pro 19% apt komponenty

◆ experimentální data

■ exptrapolace - stanovení počáteční aktivity v čase 0 s

$y = 0{,}2369*19+4{,}0401 = 8{,}54$

$y = 30000000*e^{-0{,}001*0} = 30000000$

pomocí stanovené rovnice regresní funkce můžu extrapolovat či interpolovat hodnoty y_i pro různá x_i a obráceně hodnoty x_i pro různá y_i

- **Interpolace** - výpočet hodnot mezi naměřenými body
- **Extrapolace** - výpočet mimo proměřenou oblast - pozor, zda je reálné pokračování dat podle této funkce i mimo empiricky vyšetřenou oblast

Pearsonův korelační koeficient

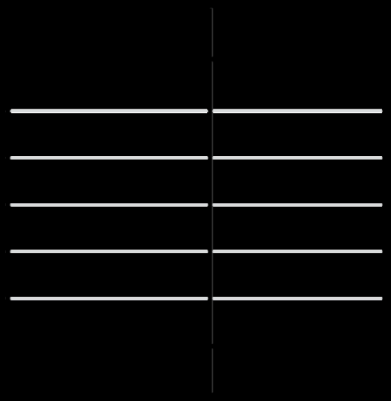
- Tzv. standardizovaná kovariance
- určení síly vztahu mezi proměnnou X a Y (s přibližně normálním rozdělením) bez nutnosti definovat závislou a nezávislou veličinu (pouze pro lineární závislost)
- Korelační koeficient může nabývat hodnot **<-1;+1>**
- Hodnota korelačního koeficientu –1 značí zcela nepřímou (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,
- V Excelu funkce CORREL

$$= \frac{S_{xy}}{S_x \, S_y}$$

- R^2 – koeficient determinace = čtverec korelačního koeficientu; **<0;+1>**

Lineární regresní model – síla závislosti

Bylo provedeno 6 měření indexu lomu roztoku NaCl ve vodě pro koncentrace NaCl 2, 4, 6, 8 a 10 % a pro destilovanou vodu. Teplota byla konstantní. Vyšetři závislost indexu lomu na koncentraci NaCl v roztoku.

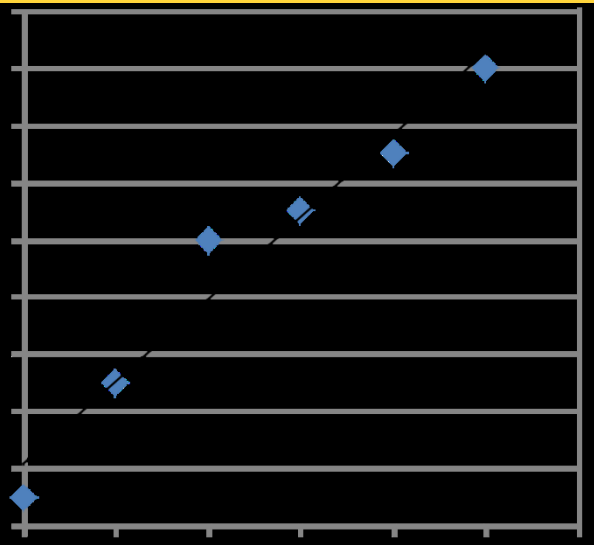


11,7

Funkce v excelu

kovariance (COVARIANCE.P/ COVARIANCE.S)

rozptyl (VAR.P/VAR.S)



1) dopočtením ze vzorců

3) Síla závislosti

r = 0.978 pearsonův korelační koeficient fce CORREL

R^2 = 0.957 koeficient determinace (čtverec korelačního koeficientu) v excelu: spojnice trendu v grafu; zobrazit hodnotu spolehlivosti R na druhou)

Pearsonův korelační koeficient

$r_{xy} = -0,9$
$R^2 = 0,81$

$r_{xy} = 1$
$R^2 = 1$

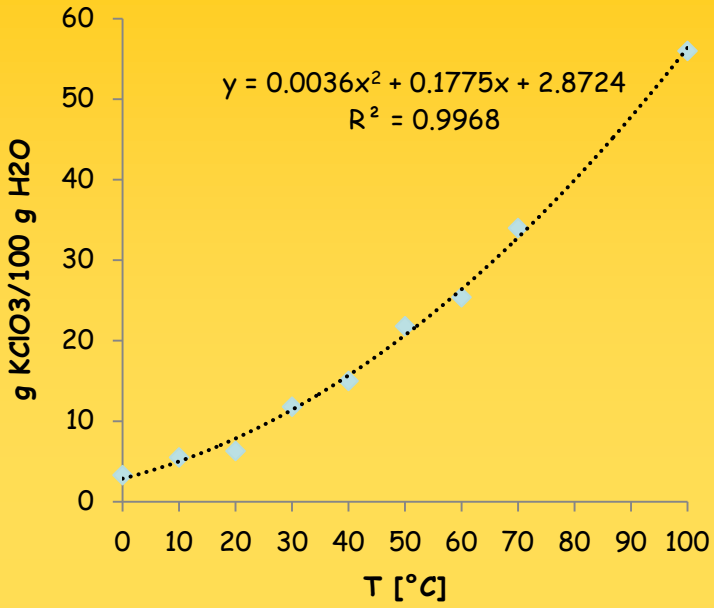
$r_{xy} = 0,9$
$R^2 = 0,81$

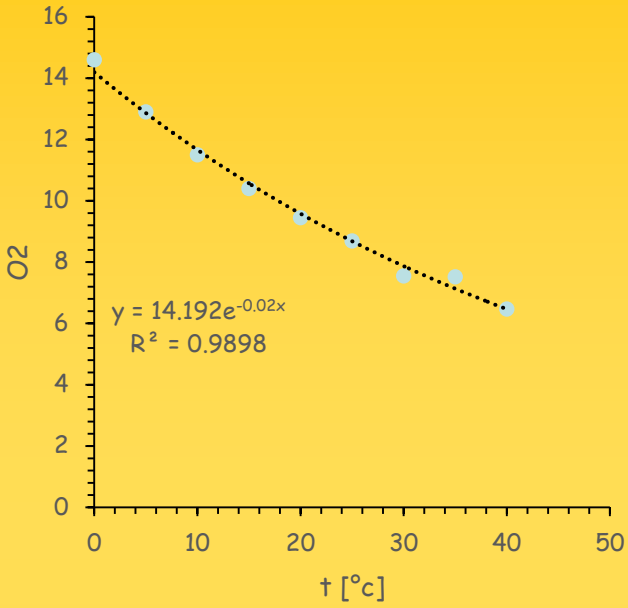
$r_{xy} = 0,35$
$R^2 = 0,12$

$r_{xy} = -0,6$
$R^2 = 0,36$

$r_{xy} = 0$
$R^2 = 0$

Nelineární závislost





Nutná pečlivá volba regresního modelu – kritéria: co nejvyšší r

reálnost pokračování regresního modelu i mimo proměřenou oblast

Nepočítat Pearsonův korelační keficient

Pro stanovení síly závislosti lze využít koeficient determinace v Excelu

- **Interpolace** - výpočet hodnot mezi naměřenými body – bez problémů
- **Extrapolace** - výpočet mimo proměřenou oblast - často problematická (zejména u kvadratické funkce), zvážit, zda je reálné pokračování dat podle této funkce i mimo empiricky vyšetřenou oblast

Děkuji za pozornost