Měření závislostí

1. **Průběh závislosti**
   - **spojitá křivka s jednoduchou rovnicí (= jednoduchým průběhem), s malým počtem parametrů, která v rozmezí naměřených hodnot vystihuje průběh závislosti,**
   - **určení konkrétního typu křivky (přímka, parabola, hyperbola, …) není součástí úlohy,**
   - **ideální je, pokud uživatel je schopen odvodit typ křivky např. z fyzikálně mechanických zákonitostí sledovaného procesu nebo ji má ověřenou z jiných obdobných analýz,**
   - **parametry by měly být interpretovatelné,**
   - **neznámé parametry křivky se odhadují metodou nejmenších čtverců, která má řadu modifikací,**
   - **nejednodušší je verze pro funkce lineární v parametrech (první parciální derivace podle všech parametrů jsou lineární funkce), pro které vždy existuje analytické řešení,**
   - **pokud funkce není lineární v parametrech, může pro ni existovat linearizující transformace, což však souvisí s mnoha potížemi,**
   - **pokud funkce není lineární v parametrech a neexistuje pro ni linearizující transformace, jedná se o nelineární závislost (nelze řešit analyticky, pouze numericky, kdy se výchozí hodnoty parametrů zpřesňují iterativním algoritmem),**
   - **zvláštní problém představují vázané parametry — např. požadujeme, aby funkce procházela počátkem nebo jiným určeným bodem, přímka měla požadovanou směrnici apod. Toto je skutečně velký problém na hranici korektní statistiky.**

2. **Intenzita závislosti**
   - **bezrozměrná na určitém intervalu (např. $\langle 0;1\rangle$, $\langle -1;+1\rangle$) normovaná charakteristika,**
   - **ta je založena na porovnání součtu čtverců odchylek vyrovnaných a naměřených hodnot závisle proměnné od jejich průměru, kde platí rovnice rozkladu součtu čtverců**

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(y'_i-\bar{y})^2+\sum_{i=1}^{n}(y_i-y'_i)^2$$

kde y_i, y'_i **jsou naměřené a vypočtené hodnoty závisle proměnné a dále platí** $\sum_{i=1}^{n}(y_i-y'_i)=0$, $\bar{y}=\bar{y'}$,

- **index determinace je podíl (zpravidla násobený stem a udávaný v %)**

$$0\le\frac{\sum_{i=1}^{n}(y'_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}(\cdot 100)\le 100\,\%$$

- **index korelace je druhá odmocnina indexu determinace (vyjádřeného jako desetinné číslo)**

$$0\le\sqrt{\frac{\sum_{i=1}^{n}(y'_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}}\le 1$$

- **koeficient determinace a korelační koeficient jsou zvláštními případy indexu determinace a indexu korelace, pokud průběh závislosti měří <u>přímka</u>; korelační koeficient lze alternativně spočítat jako**

$$-1\le\frac{\operatorname{cov} xy}{\sqrt{\operatorname{var} x\cdot\operatorname{var} y}}\le 1$$

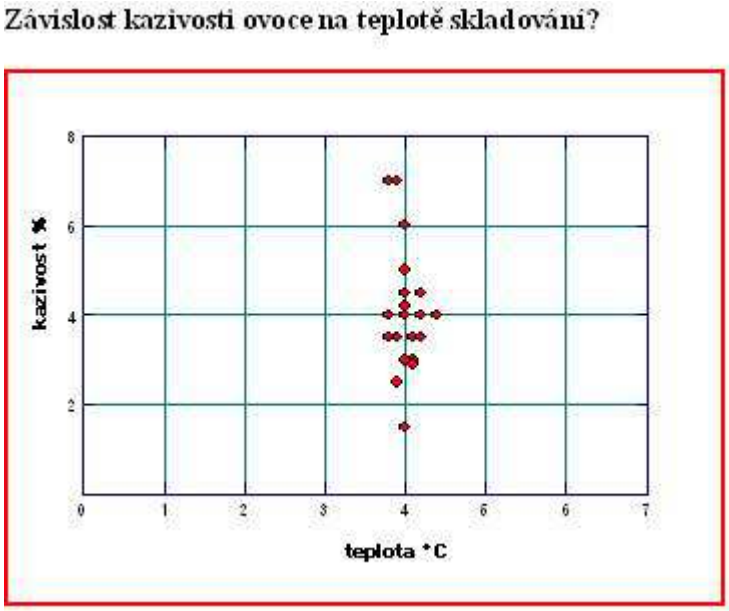
  kde cov *xy* **je kovariance a** var *x*, var *y* **jsou rozptyly nezávisle a závisle proměnné.**
- **rovnice rozkladu součtu čtverců <u>neplatí</u> (a tudíž nelze stanovit smysluplný index korelace)**
  - **u funkce jejíž rovnice byla vypočtena pomocí linearizující transformace a posléze inverzní**

transformací „vrácena“ do původního tvaru, např. $y' = b_0 b_1^x \rightarrow \log y' = \log b_0 + \log b_1 \cdot x$ **pro přímku vzniklou logaritmováním exponenciální funkce rovnice platí** $\overline{\log y} = \overline{\log y'}$ **, po inverzní transformaci už** $\bar{y} = \bar{y'}$ **neplatí,**

- **u funkce s vázanými parametry platí** $\bar{y} \neq \bar{y'}$ **, protože** $\sum_{i=1}^{n} (y_i - y'_i) \neq 0$ **,**
- **u nelineární regrese, kde rovnice platí jen po určité úpravě.**

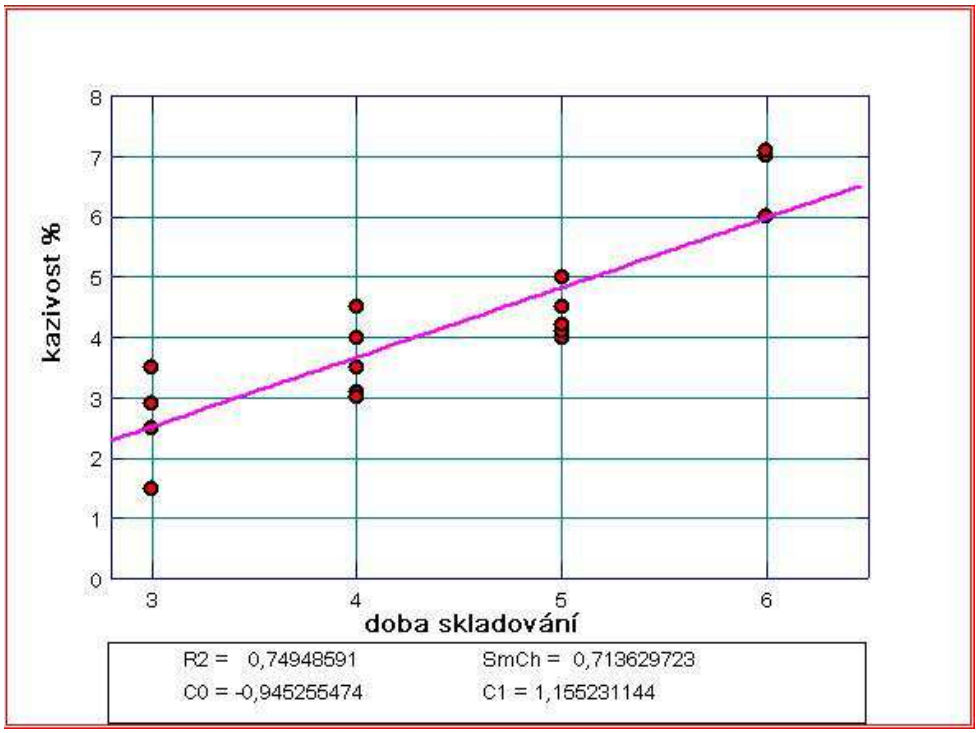
Dvě varianty úlohy o závislosti (modelové případy, život je složitější)

1. Je dána jednosměrná příčinná závislost (nezávisle proměnná je příčina a závisle proměnná je účinek), nezávisle proměnná je řízená, její hodnoty stanovuje (více nebo méně vhodně) experimentátor; závisle proměnná je pozorovaná (náhodná) veličina

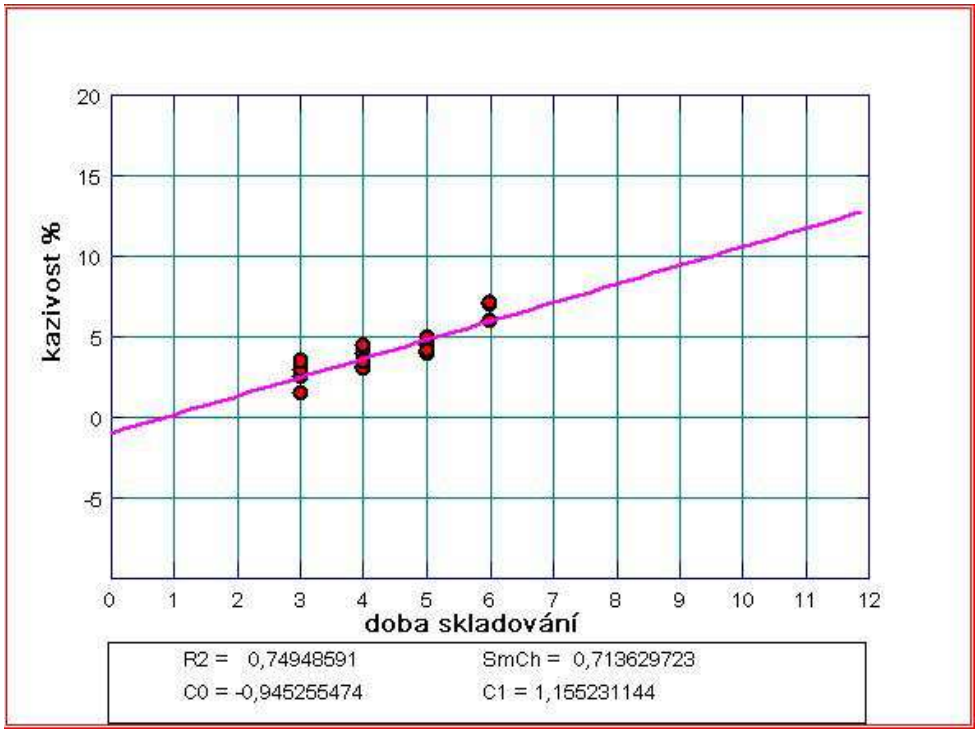


Příklad katastrofálně nevhodně zvolené nezávisle proměnné (s nepatrnou variabilitou).

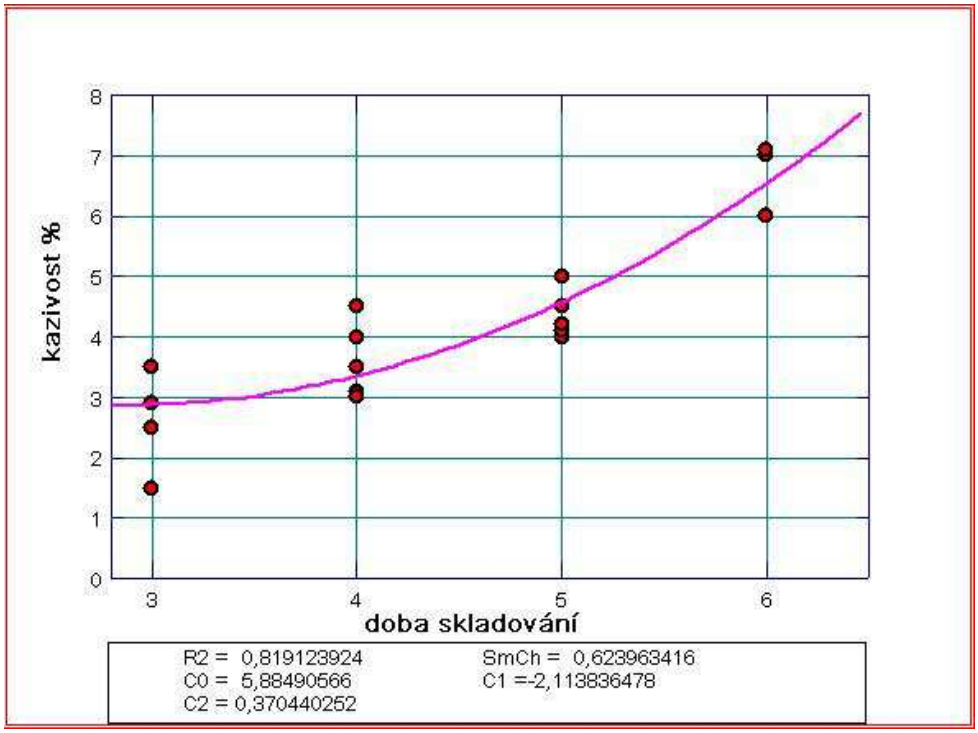
Závislost kazivosti ovoce na době skladování



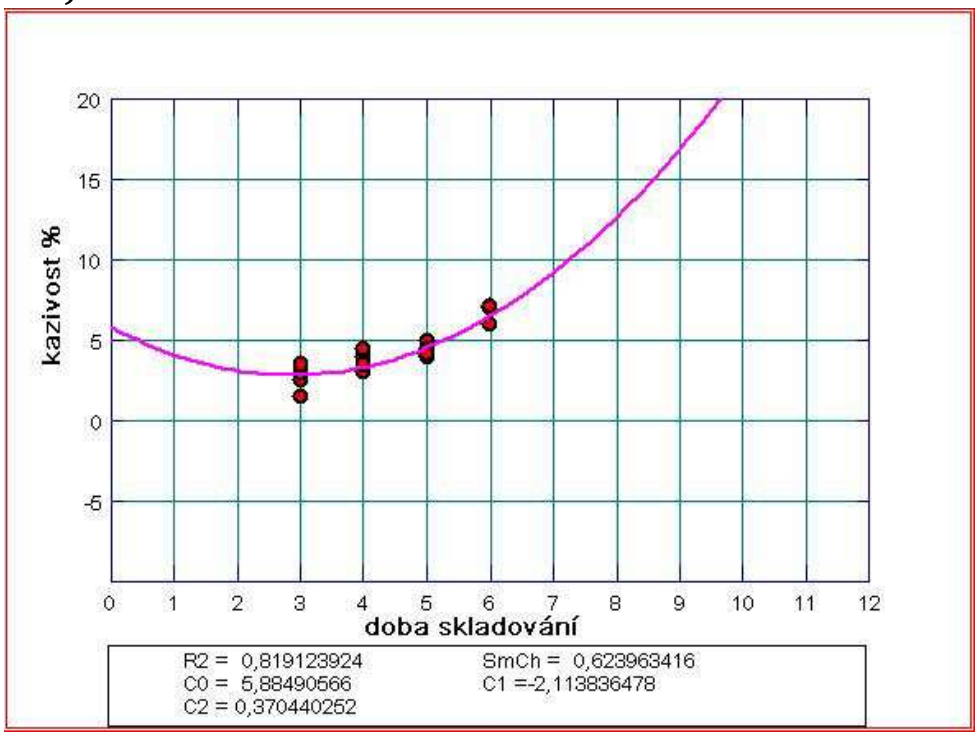
Vystižená přímkou o rovnici $y' = -0{,}9453 + 1{,}1552x$ **s indexem determinace 74,9 % (index korelace je 0,866). Současně jde o koeficient determinace a korelační koeficient (přímka).**



Rozšířením intervalů stupnic na obou osách zjišťujeme, že nemáme žádný důvod tvrdit, že přímka vystihuje lineární růst kazivosti po celou dobu skladování (0 až 12 měsíců).



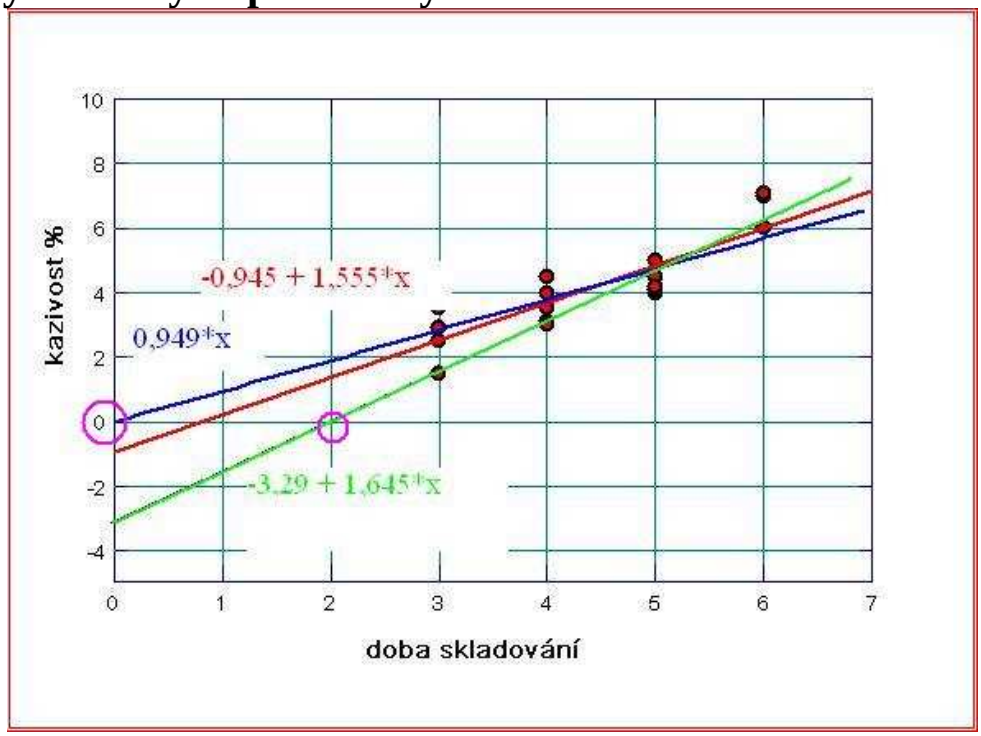
Kvadratická funkce vystihuje závislost o něco lépe, index korelace je roven 0,905.



Zopakujeme-li totéž co u přímky, vidíme že mimo interval měřených hodnot má funkce krajně nedůvěryhodný průběh — předpokládá nejprve pokles(?!) a pak strmý růst kazivosti. Daleko „šílenější“ věci vidíme, pokud používáme polynomy vyšších stupňů — interpolace se zlepšuje s rostoucím stupněm

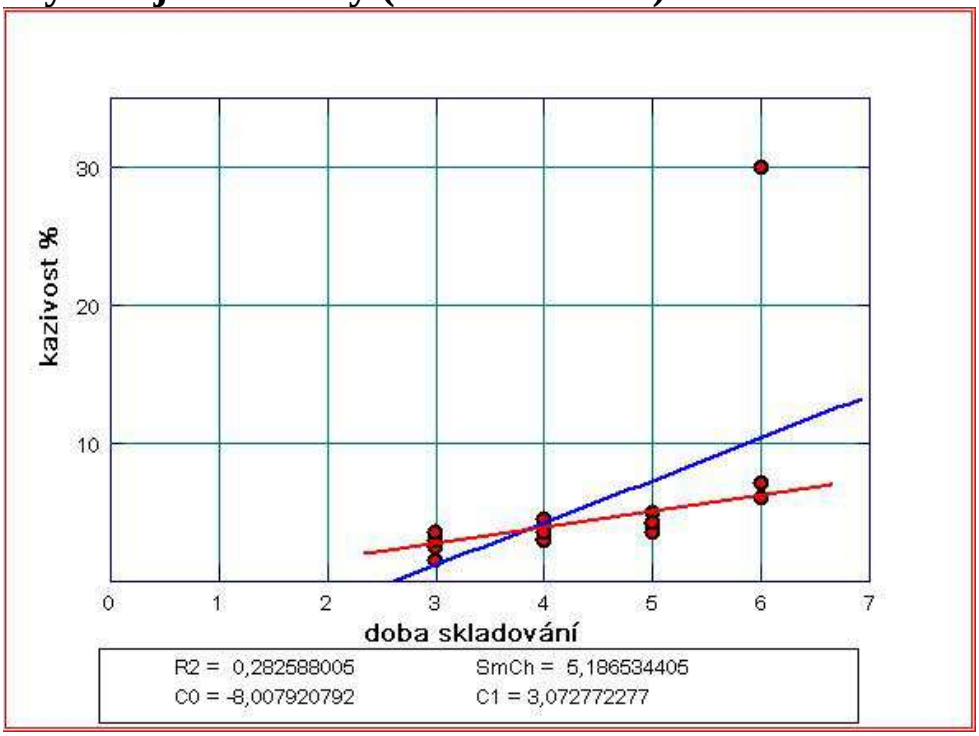
polynomu, ovšem extrémní růsty a pády mimo interval měřených hodnot, velký počet parametrů, žádná interpretace

Přímky s vázanými parametry



Červená je původní přímka o oběma odhadovanými parametry. Modrá přímka je vedena tak, aby procházela počátkem (žádné skladování, žádná kazivost).
Zelená přímka je vedena tak, aby procházela bodem o souřadnicích $[2;0]$ (předpokládáme, že první dva měsíce skladování nedochází k žádné změně v kazivosti).
U obou posledních přímek nelze smysluplně určit intenzitu závislosti (index korelace může vyjít i záporně nebo větší než jedna), součet odchylek vypočtených a naměřených hodnot závisle proměnné není nulový a rovněž oba průměry se liší.

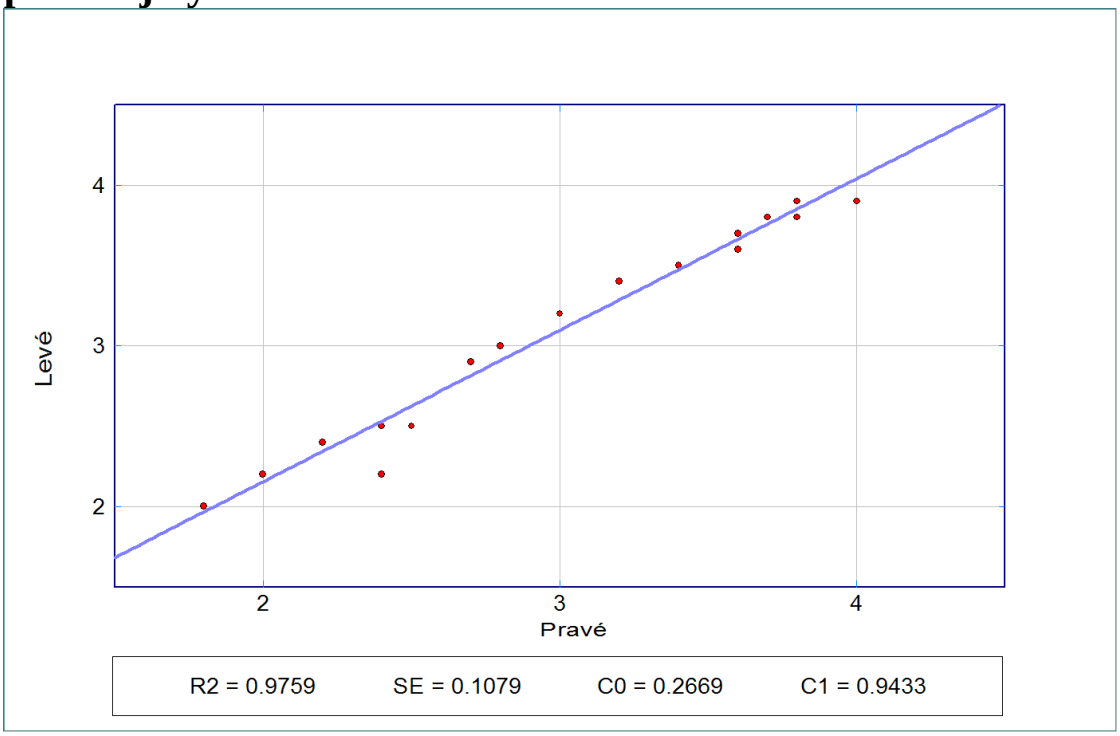
Výskyt vybočující hodnoty (vlivného bodu)

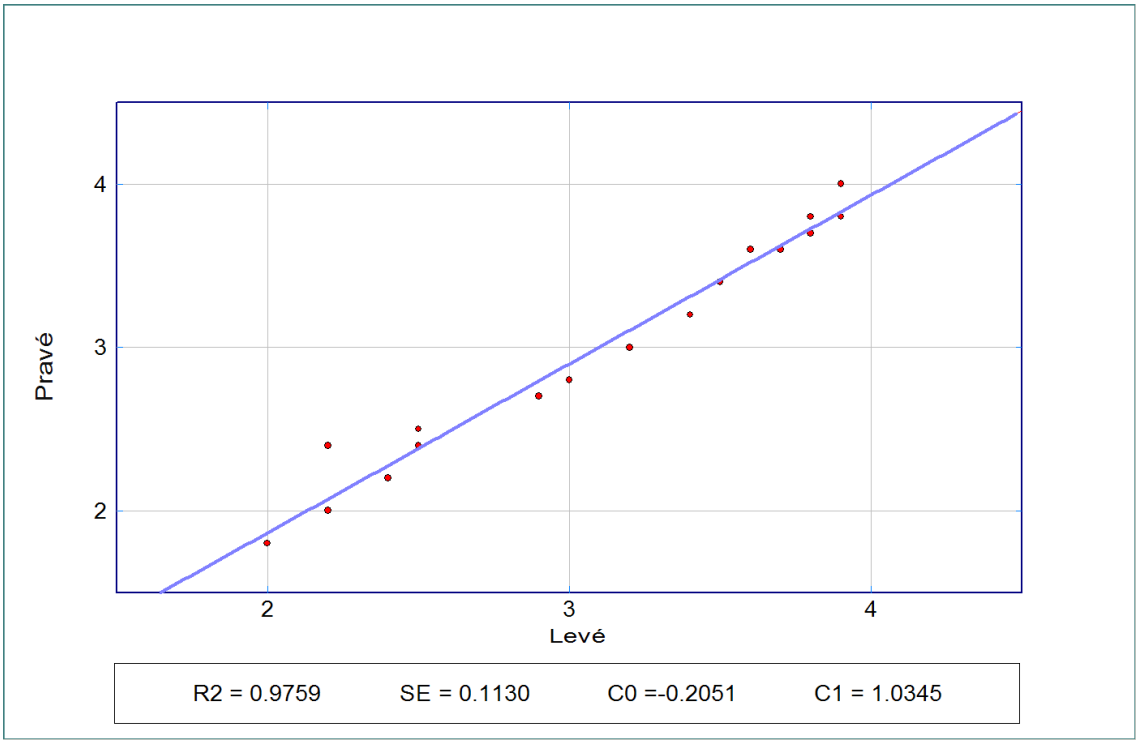


Ve statistice je nepřípustné, aby jediná hodnota výrazně změnila výsledek úlohy. Pokud takové měření v úloze existuje, označuje se jako vlivný bod.
Červená přímka je vypočtena z původních dat.
Modrá přímka vykazuje hodnoty parametrů a tudíž i průběh silně ovlivněný vybočující hodnotou. Dále (při splnění rovnice rozkladu součtu čtverců) vykazuje podstatně nižší intenzitu závislosti.
Závěr — dohledat příčiny (např. porušení ochranné atmosféry, porucha klimatizace) a vlivný bod z dat vyřadit.

2.Není zřejmé, která proměnná představuje příčinu a která účinek. Obě jsou pozorované proměnné (náhodné veličiny).
Data tvoří tzv. elipsu rozptylu. Průběh závislosti měří svazek dvou sdružených regresních přímek, intenzitu závislosti můžeme měřit buď indexem nebo koeficientem korelace (mají stejnou absolutní hodnotu, koeficient znaménkem informuje o směru závislosti).

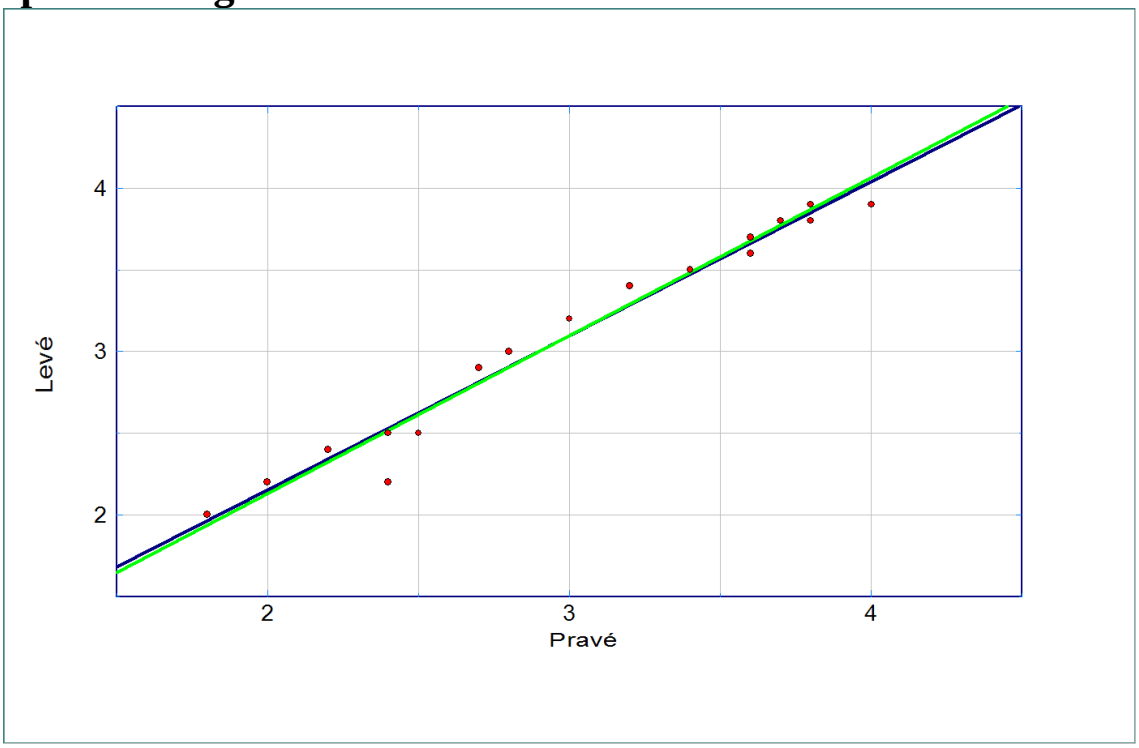
Výška dezénu na pravém a levém kole jedné nápravy při různém počtu ujetých kilometrů





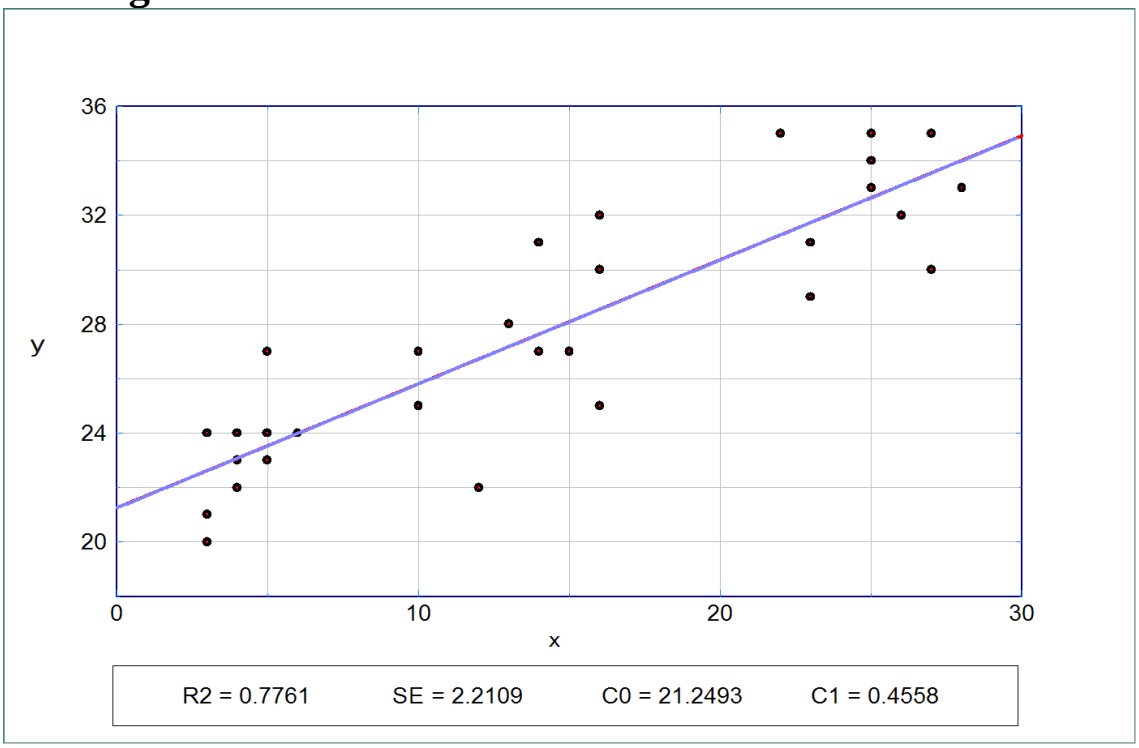
Obě přímky vykazují stejnou intenzitu závislosti.

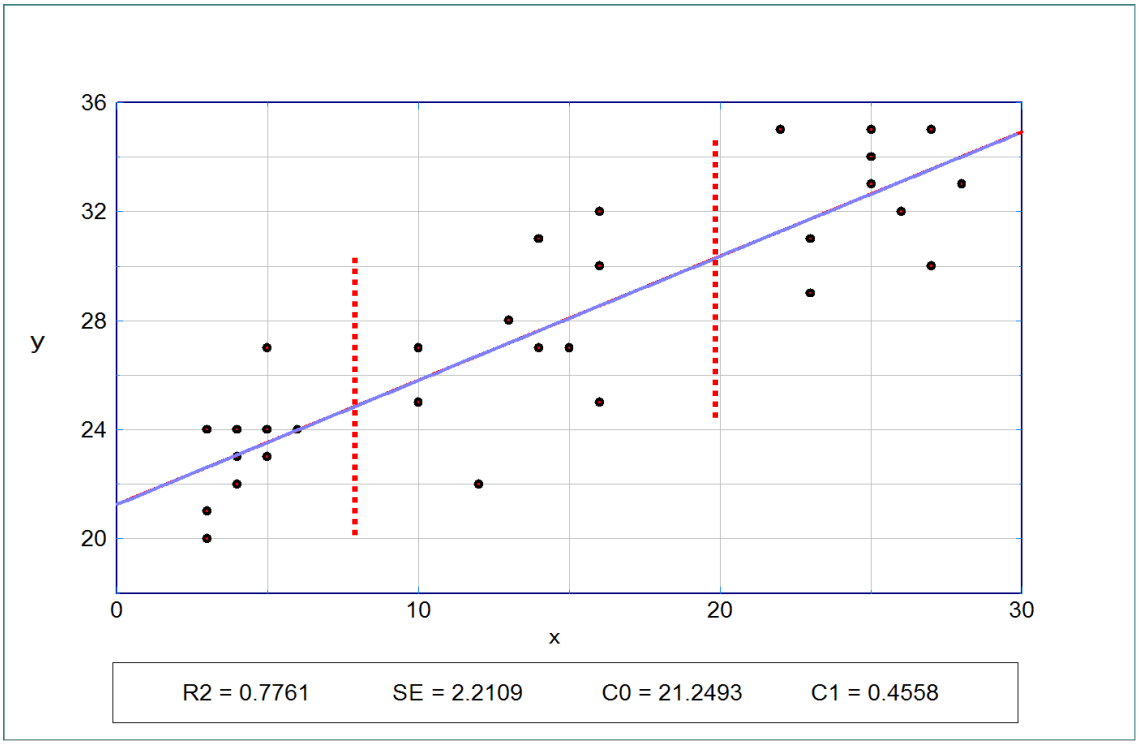
Obě přímky (pokud je potřebujeme) znázorňujeme zpravidla do společného grafu



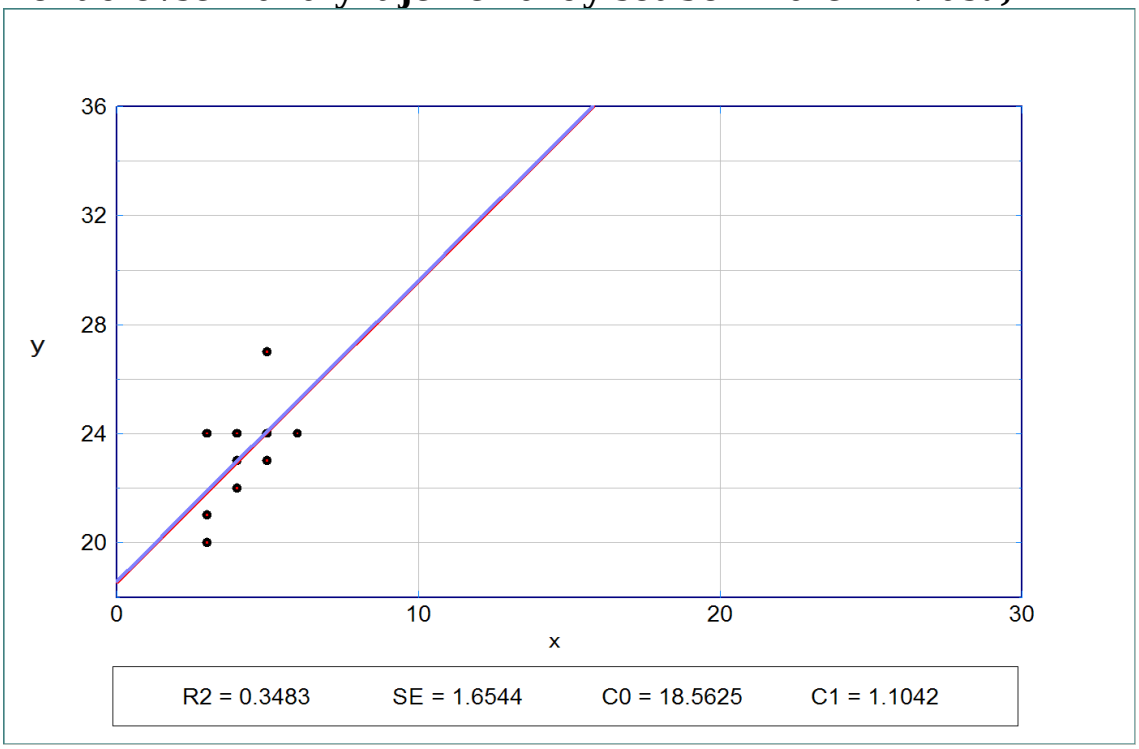
Přímky se protínají v bodě, který má souřadnice průměrů obou proměnných. Úhel, který svírají, je tím menší, čím je závislost intenzivnější.

Heterogenita v datech

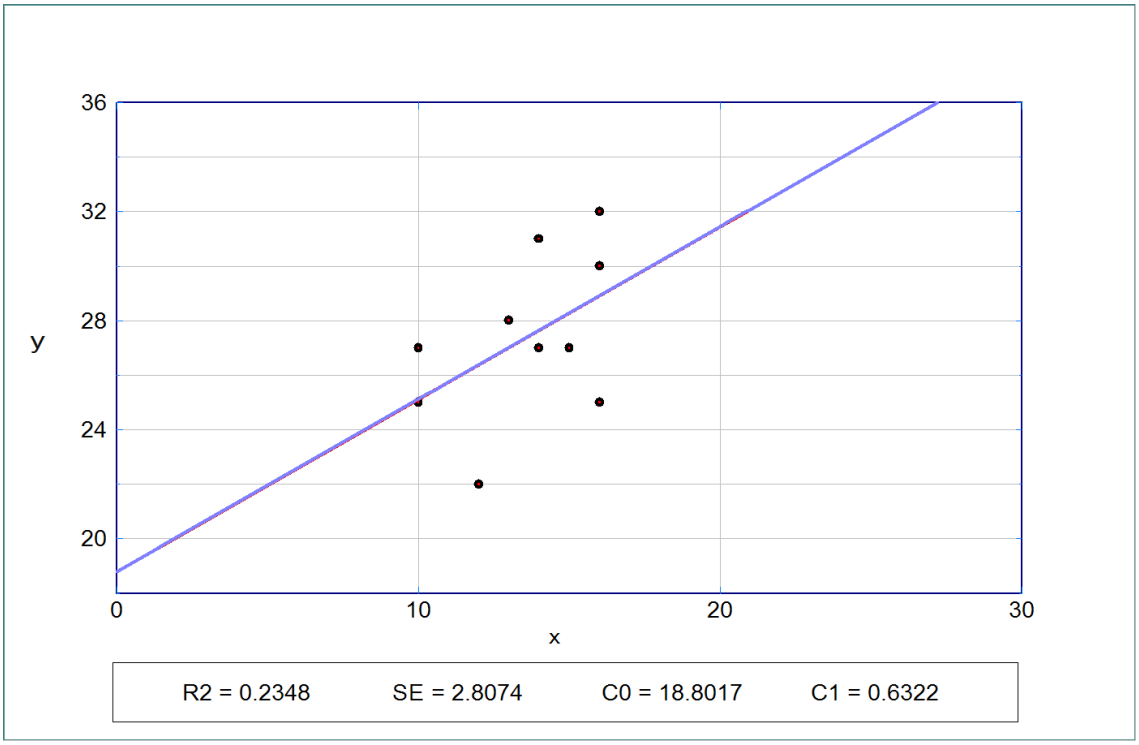


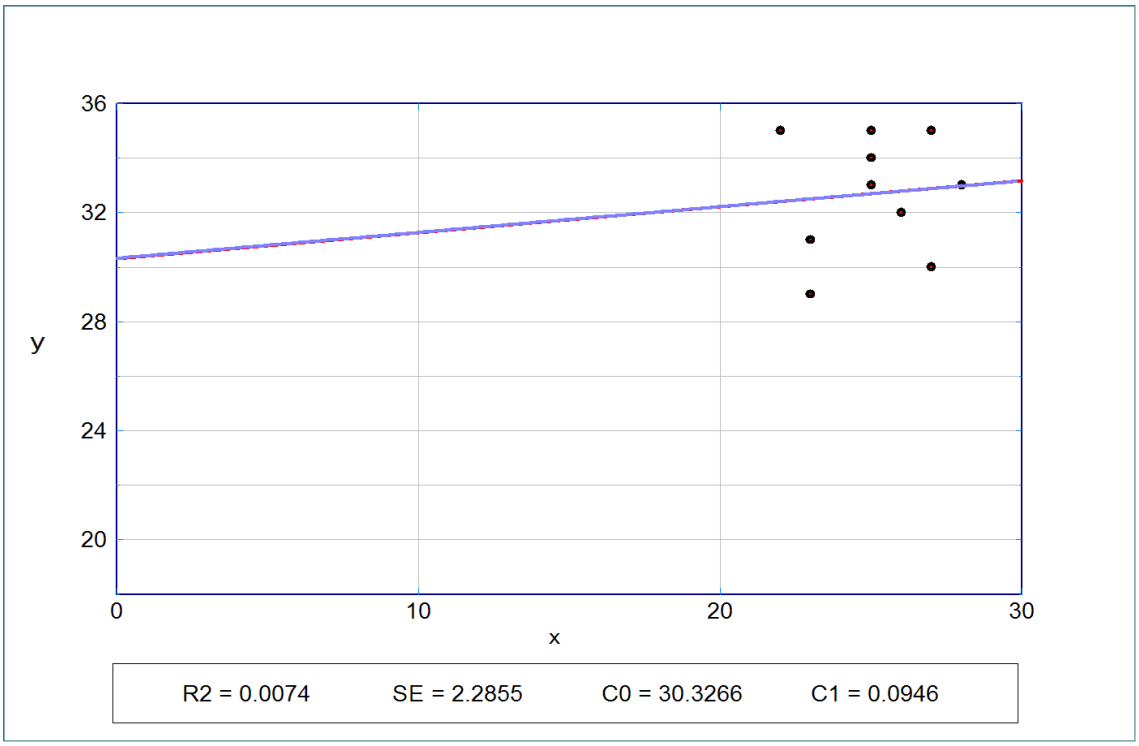


Běžně se vyskytující „závada“ v datech, kdy dojde ke spojení souborů, které by měly být analyzovány samostatně. To že data leží prakticky na společné přímce (viz obrázek se třemi dílčími soubory) vede k efektu „zesílení“ intenzity závislosti.
Pokud ovšem analyzujeme každý soubor měření zvlášť,

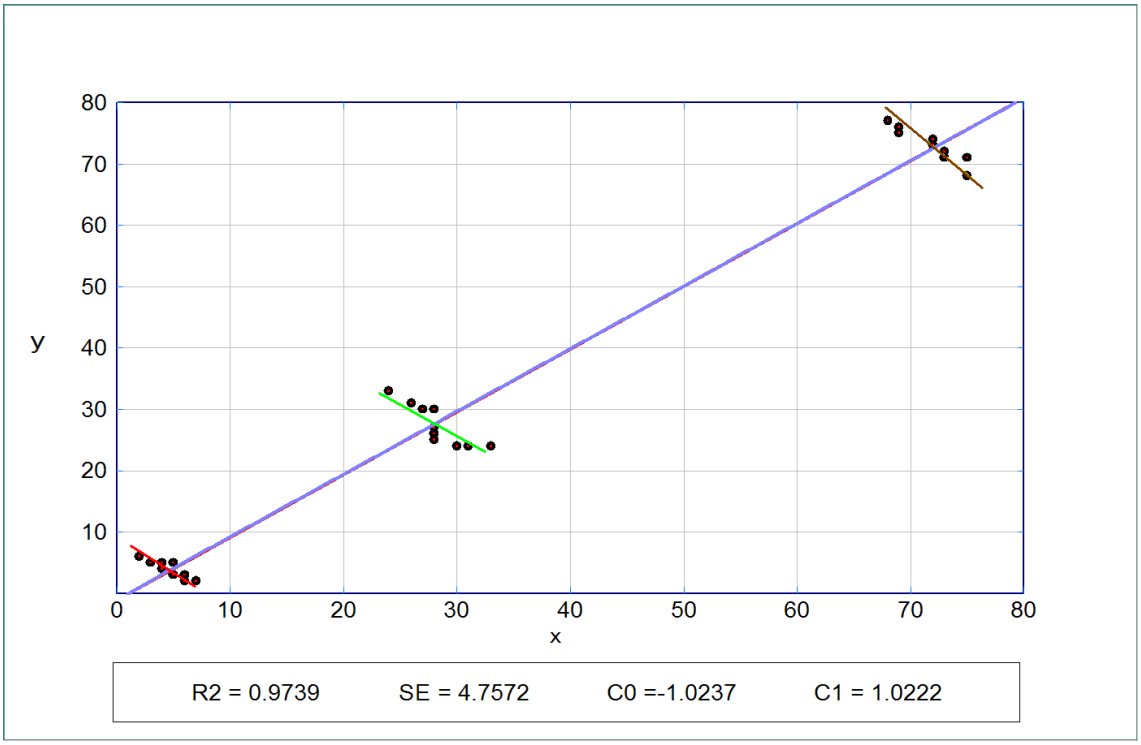


zjistíme (ne nutně!), že skutečnost je poněkud jiná. Svědčí o tom tento a následující dva obrázky.





Někdy to ovšem dopadne podstatně hůře.



Doposud jsme k měření závislostí přistupovali popisným způsobem, tj. analyzovali jsme jen konkrétní soubor měření bez širších souvislostí.

Induktivní úvahy o závislosti
Naměřené hodnoty (se svými chybami) jsou jedinečným, neopakovatelným „vzorkem“ reality a tak, jako je nedokážeme bezezbytku reprodukovat, nedokážeme přesně zopakovat ani z nich vypočtené charakteristiky závislosti.
Proto jsou z výběrového souboru vypočtené charakteristiky (u regresní přímky např. směrnice) náhodnými veličinami s nějakou střední hodnotou a směrodatnou odchylkou, která se u výběrových charakteristik nazývá směrodatná chyba. Vedle tohoto pojmu operujeme ještě s pojmem přípustná chyba, která udává kolika násobek směrodatné chyby považujeme ještě za přijatelný (svědčící o statistické shodě) a naopak kolika násobek již vypovídá o statistické neshodě vypočtených parametrů.
Velikost přípustné chyby u charakteristik závislosti závisí

- **na rozsahu souboru (počtu měření) — čím je větší, tím je směrodatná chyba menší,**

- **na intenzitě závislosti — čím je závislost intenzivnější, tím je přípustná chyba menší,**
- **na pravděpodobnosti (spolehlivosti, hladině významnosti) induktivní úvahy — čím je požadovaná pravděpodobnost bližší jedné, tím je přípustná chyba větší.**

Z úloh statistické indukce je třeba jmenovat

- **Bodový odhad, který slouží k určení výběrové charakteristiky v podobě jediného čísla. Jako příklady výběrových charakteristik můžeme jmenovat např. absolutní člen nebo směrnici přímky, regresní přímku jako celek (i to je výběrová charakteristika!), korelační koeficient, ale také např. rozdíl dvou směrnic, rozdíl dvou nebo více korelačních koeficientů apod. (tedy jednoduché funkce dvou nebo i více charakteristik)**

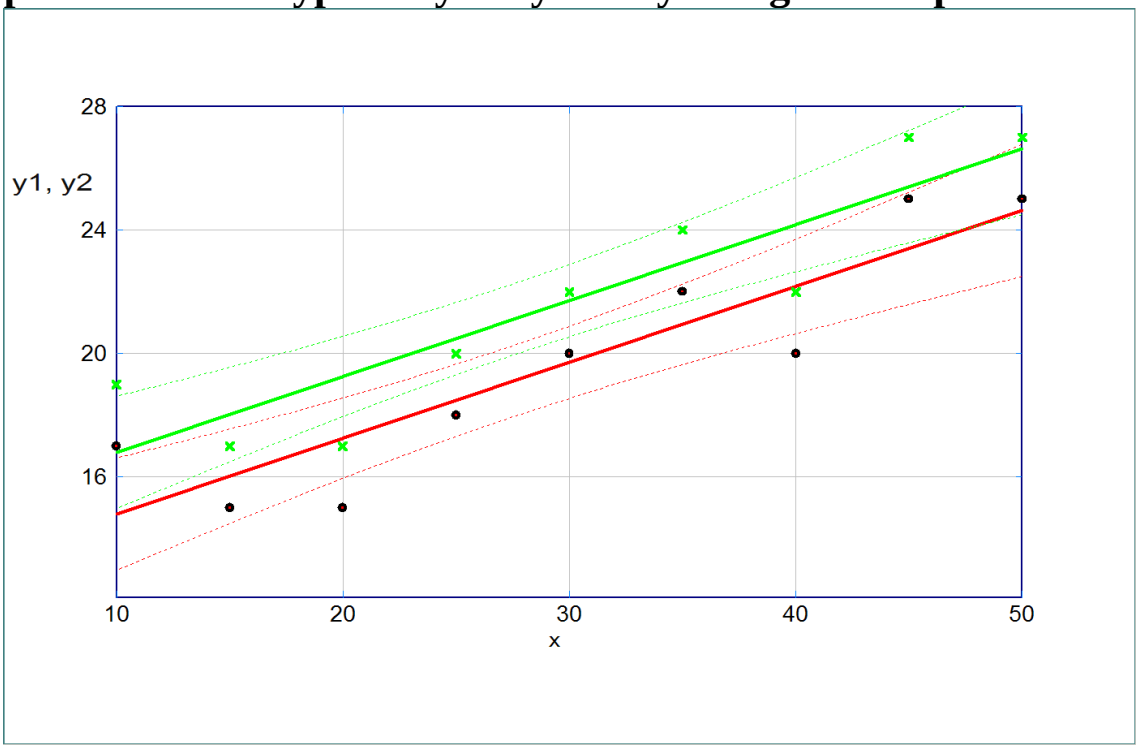
- **Intervalový odhad, kdy pro výběrové charakteristiky sestrojujeme tzv. konfidenční intervaly (též intervaly spolehlivosti), které si můžeme představit jako úsečky či polopřímky (podle potřeby volíme oboustranné či jednostranné intervaly), na kterých — s vysokou předem zvolenou pravděpodobností blízkou jedné — výběrová charakteristika leží. Riziko intervalového odhadu pak udává pravděpodobnost, s jakou výběrová charakteristika v konfidenčním intervalu neleží. Poloha hranic konfidenčních intervalů je odvozena od přípustné chyby.**

- **Testování hypotéz o charakteristikách závislosti. Předem vyslovený předpoklad o nějaké charakteristice (rozdílu dvou charakteristik apod.) závislosti se ověřuje na základě vypočtených hodnot výběrových charakteristik jejich srovnáním s předpokládanou hodnotou (nebo hodnotou pocházející z jiného analogického výběru). Výsledkem je vypočtená hodnota tzv. testového kritéria, která vypovídá o udržitelnosti/neudržitelnosti původního předpokladu. Pomocí testování hypotéz nelze dokázat, že předpoklad je**

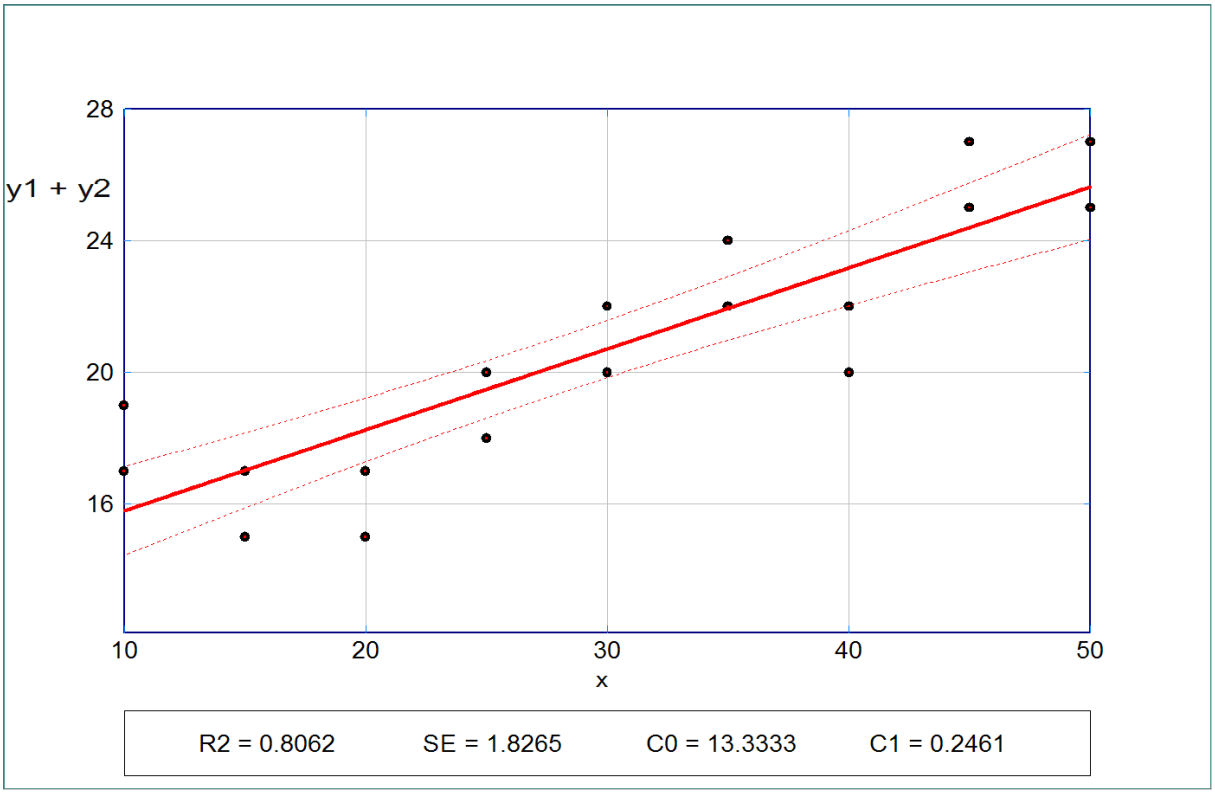
jednoznačně (s jistotou) pravdivý/nepravdivý. Naopak, existují dvě možnosti chybných rozhodnutí

- **pravdivý předpoklad se jeví jako neudržitelný (zamítnutí pravdivé hypotézy, chyba prvního druhu),**
- **nepravdivý předpoklad je jeví jako udržitelný (nezamítnutí nepravdivé hypotézy, chyba druhého druhu).**

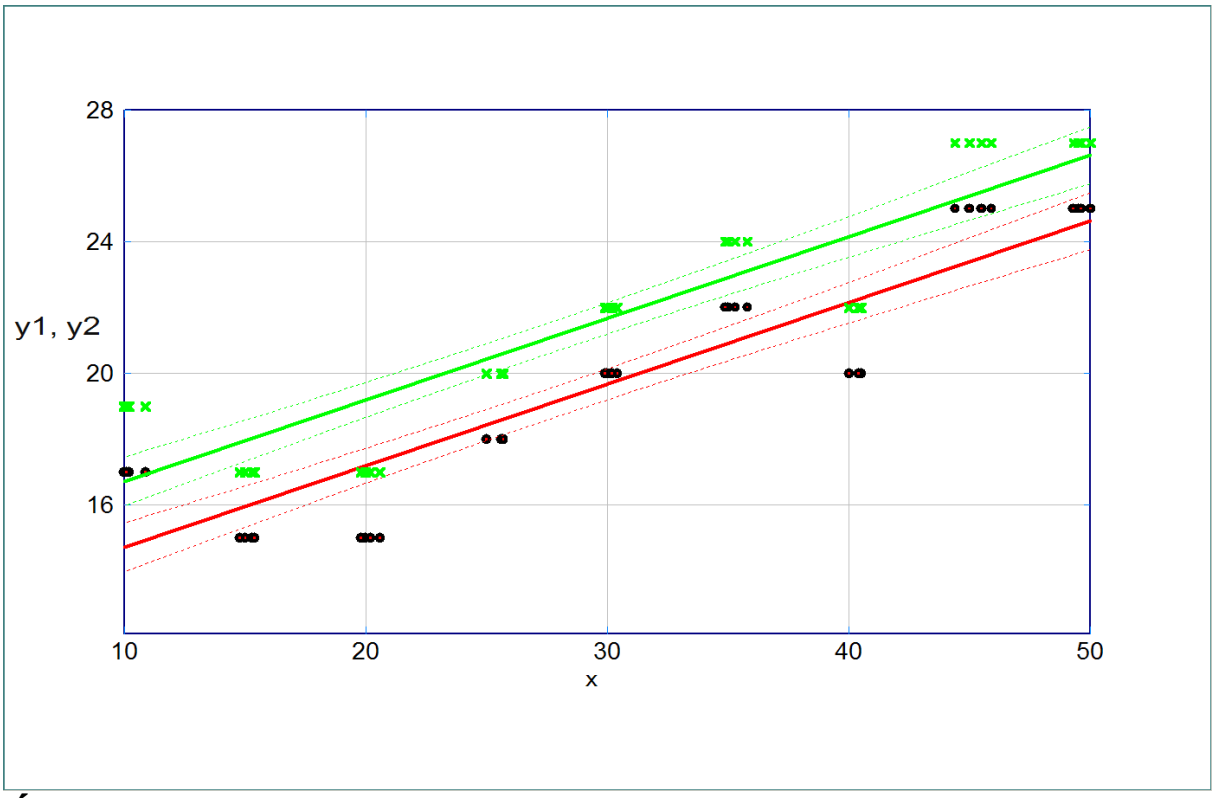
Příklad:
Jednou z možností je sestrojit konfidenční interval pro regresní přímku. Ten je symetrický kolem vypočtené přímky, nejužší v oblasti průměrů nezávisle a závisle proměnné a postupně se na obě strany rozšiřuje. Konferenční interval přímky je tím širší, z čím menšího počtu hodnot byla přímka vypočtena, čím slabší je závislost obou veličin a čím větší spolehlivost (menší riziko) odhadu je požadována.
Konfidenční intervaly lze využít pro ověření předpokladu o shodě průběhu dvou vypočtených výběrových regresních přímek.



Konfidenční intervaly obou přímek se překrývají, předpoklad o shodě je udržitelný.



Obě závislosti lze vyjádřit společnou regresní přímkou.



Úlohu jsme zopakovali se stejnými přímkami, vypočtenými ze čtyřnásobného počtu měření. Konfidenční intervaly se tentokrát nepřekrývají, průběh obou přímek se s vysokou pravděpodobností liší. Obě závislosti nelze vyjádřit společnou přímkou.

Příklad nelineární regrese

Tzv. *záběrová křivka traktoru*, která může být uvedena např. ve tvaru (*Grečenko*, 1994)

$$\mu = (\frac{b_0}{1+\delta} + b_1)(b_2\delta + 3 - \sqrt{(b_2\delta - 1)^2 + 8})$$

kde μ **je bezrozměrný *součinitel záběru*,** δ **je rovněž bezrozměrný *prokluz* a** b_0, b_1, b_2 **jsou *parametry*.**

Ze 182 **měření byl po vyrovnání získán polynom 6. stupně**

$$\mu = -0{,}114 + 0{,}506\delta - 0{,}0019\delta^2 + 4{,}5.10^{-5}\delta^3 - 6{,}4.10^{-7}\delta^4 + 4{,}8.10^{-9}\delta^5 - 1{,}4.10^{-11}\delta^6$$

s $RSS = 0{,}0231$, **který z mnoha důvodů nevyhovuje.**

***Grečenkovou* metodou byl získán výchozí odhad parametrů a záběrová křivka je pak ve tvaru**

$$\mu = (\frac{0{,}07505}{1+\delta} + 0{,}11832)(13{,}9245\delta + 3 - \sqrt{(13{,}9245\delta - 1)^2 + 8})$$

s reziduálním součtem čtverců odchylek $RSS = 0{,}0366$.

***Marquartovým* algoritmem nelineární regrese byly výchozí hodnoty parametrů zlepšeny a záběrová křivka je ve tvaru**

$$\mu = (\frac{0{,}00028}{1+\delta} + 0{,}15810)(17{,}46185\delta + 3 - \sqrt{(17{,}46185\delta - 1)^2 + 8})$$

s reziduálním součtem čtverců odchylek $RSS = 0{,}0248$.

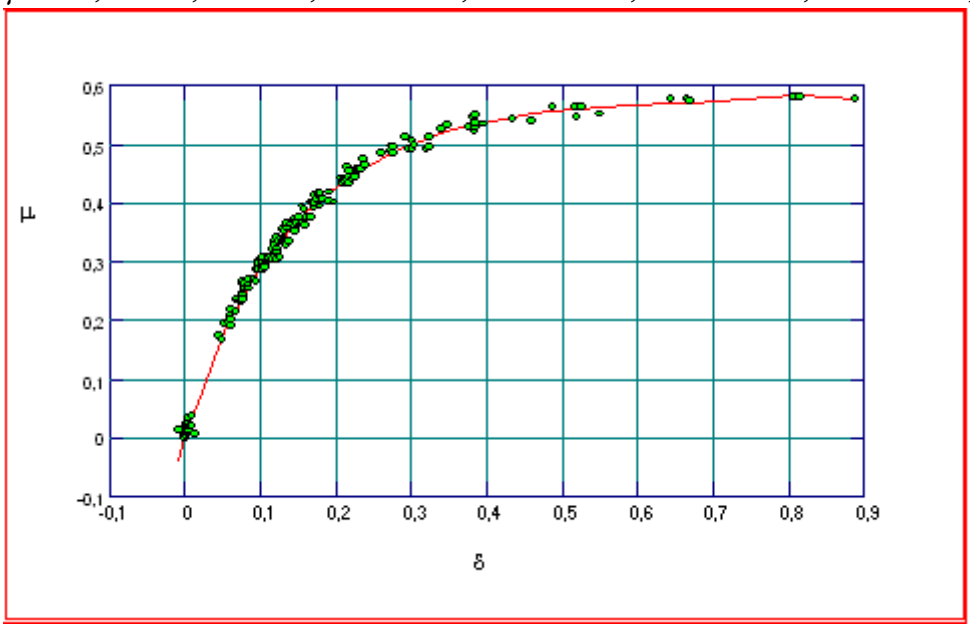
Testováním parametrů byla zjištěna statistická nevýznamnost parametru b_0**. Výsledný tvar křivky je pak**

$$\mu = 0{,}15827(17{,}4770\delta + 3 - \sqrt{(17{,}4770\delta - 1)^2 + 8})$$

s reziduálním součtem čtverců odchylek $RSS = 0{,}0248$.

Polynom 6. stupně

$$\mu = -0{,}114 + 0{,}506\delta - 0{,}0019\delta^2 + 4{,}5.10^{-5}\delta^3 - 6{,}4.10^{-7}\delta^4 + 4{,}8.10^{-9}\delta^5 - 1{,}4.10^{-11}\delta^6$$



Záběrová křivka

$$\mu = 0{,}15827(17{,}4770\delta + 3 - \sqrt{(17{,}4770\delta - 1)^2 + 8})$$