KGG/STG Statistika pro geografy

10. Regresní analýza

Mgr. David Fiedor
27. dubna 2015

Nelineární závislost - korelační poměr

- užití v případě, kdy regresní čára není přímka, ale je vyjádřena složitější matematickou funkcí
- prvky výběru závisle proměnné y_i rozdělíme podle hodnot nezávisle proměnné x_i do skupin označených y_j a pro každou skupinu se vypočítá průměr $\bar{y}_j$
- korelační poměr se vypočítá podle vztahu

$$\eta_{yx} = \sqrt{\frac{\sum(\bar{y}_j - \bar{y}) \cdot n_j}{\sum(y_i - \bar{y})^2}} = \sqrt{\frac{\sum(\bar{y}_j n_j - n\bar{y})^2}{\sum y_i^2 - n\bar{y}^2}}$$

- n_j je četnost v y_j
- porovnání hodnot korelačního koeficientu a korelačního poměru lze použít jako kritéria linearity vztahu

Koeficient mnohonásobné korelace

- vztah dvou proměnných je často ovlivněn dalšími proměnnými - chceme zjistit celkovou sílu vztahu mezi zvolenou proměnnou na jedné straně a několika dalšími proměnnými
- pro hodnocení korelační závislosti tří nebo více výběrů náhodných veličin - koeficient mnohonásobné korelace
- vyjadřuje číselně míru predikce cílové proměnné X pomocí proměnných Y a Z

$$r_{x.yz} = \sqrt{\frac{r_{xz}^2 + r_{yz}^2 - 2r_{xy} \cdot r_{xz} \cdot r_{yz}}{1 - r_{xy}^2}}$$

Parciální korelační koeficient

- zabývá se otázkou vlivu jedné nebo více nezávisle proměnných na závisle proměnnou při vyloučení vlivu zbývajících nezávisle proměnných, u nichž předpokládáme konstantní hodnotu
- tento koeficient lze považovat za zvláštní případ koeficientu mnohonásobné korelace, kdy další proměnné považujeme za „rušivé“

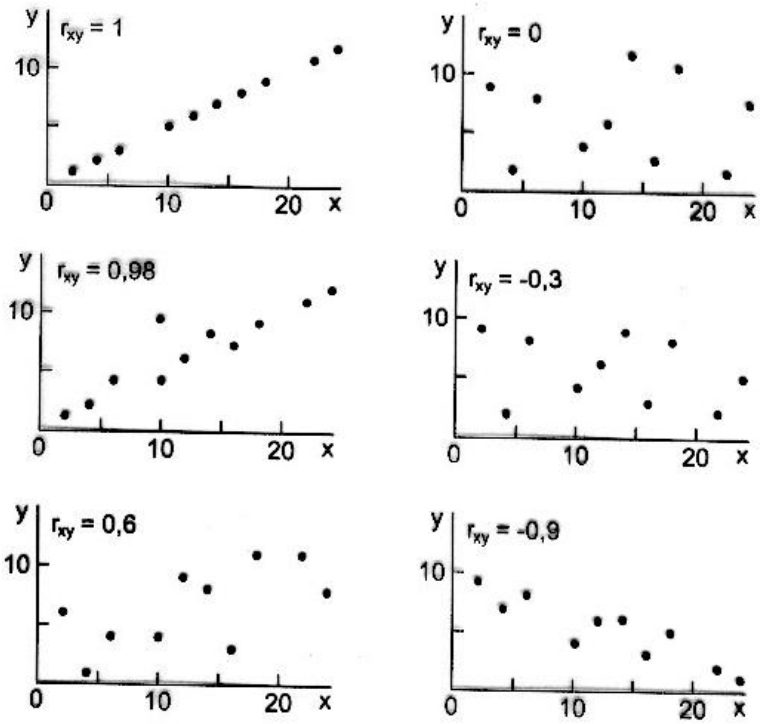
$$r_{xy.z} = \frac{r_{xy} - r_{xz} \cdot r_{yz}}{\sqrt{(1 - r_{xz}^2) \cdot (1 - r_{yz}^2)}}$$

- příkladem z lékařské praxe mohou být proměnné věk, krevní tlak a koncentrace cholesterolu v krvi u žen (parametr věku bereme jako rušivý element)

Korelační pole

- bodový graf znázorňující obě náhodné veličiny (závisle proměnnou na ose x a nezávisle proměnnou na ose y
- pomocí tohoto grafu lze posoudit dvourozměrnou normalitu dat (při dostatečně velkém počtu pozorování by měly body tvořit elipsu)

Korelační pole



Regresní analýza - úvod

- již se nejedná o určení síly závislosti statistických znaků (korelační analýza), ale o určení druhu závislosti
- úkolem je tedy sestavit vztah (model) závislosti mezi závisle a nezávisle proměnnou
- regresní analýza se zabývá odhadem neznámých parametrů regresní funkce, testováním hypotéz o těchto parametrech a také ověřováním předpokladů regresního modelu
- především se budeme věnovat lineární regresní závislosti (regresní přímkou)

Lineární regresní závislost

- nejjednodušší případ regresní závislosti - přímka

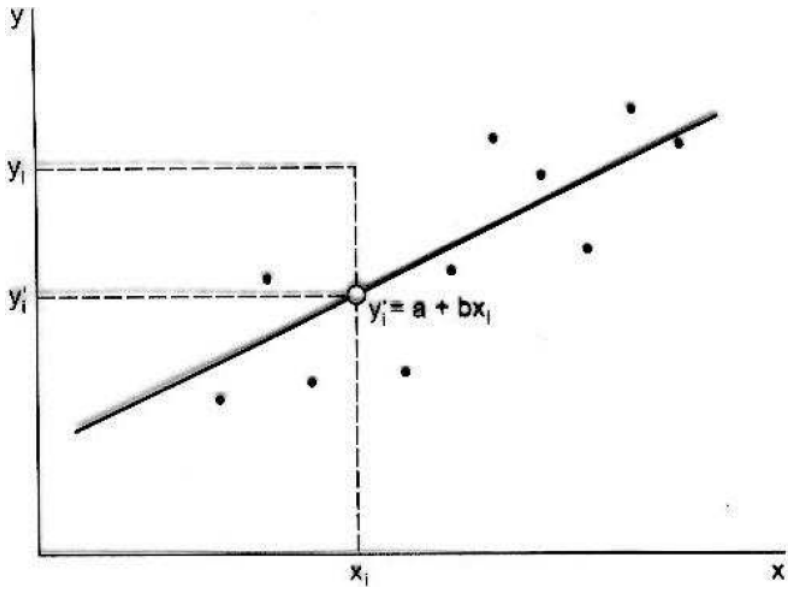
$$y' = a + bx$$

- symbol y' se používá pro označení nejpravděpodobnější teoretické hodnoty y odpovídající danému x - hodnoty, která „leží“ na regresní přímce (většinou se tato hodnota odlišuje od konkrétní hodnoty y_i nacházející se mimo přímku)
- vyvstává otázka: jak určit rovnici regresní přímky?

Metoda nejmenších čtverců

- určuje průběh regresní přímky - její parametry
- určující podmínka: součet čtverců vzdáleností všech bodů pole od přímky musí být minimální, tj.:

$$\sum (y_i - y'_i)^2 = \min$$



Metoda nejmenších čtverců

- výpočet vertikální vzdálenosti bodů korelačního pole od regresní přímky se provádí podle předchozího obrázku
- vzdálenost konkrétní hodnoty závisle proměnné y_i od bodu regresní přímky y'_i musí platit vztah:

$$y_i - y'_i = y_i - a - bx_i$$

- součet čtverců „svislých“ vzdáleností y_i od regresní přímky je potom

$$\sum (y_i - y'_i)^2 = \sum (y_i - a - bx_i)^2 = A$$

- pro metodu nejmenších čtverců musí platit $A = \sum (y_i - a - bx_i)^2 = \min$

Výpočet koeficientů regresní přímky

Úpravami vztahů dostaneme výrazy pro výpočet koeficientů regresní přímky:

$$b = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2}$$

Tento koeficient je směrnicí přímky (tangenta úhlu, který svírá přímka s osou x.

$$a = \bar{y} - b\bar{x}$$

Udává průsečík přímky a osy y.

Výpočet koeficientů regresní přímky

Výpočet koeficientu b lze zjednodušit pomocí vztahů pro kovarianci a směrodatnou odchylku, tedy:

$$b = \frac{S_{xy}}{S_x^2}$$

Intervaly a pásy spolehlivosti lineární regresní závislosti

- regresní přímku konstruujeme z dat výběrového souboru
- z tohoto důvodu se mohou rovnice této přímky lišit pro různé náhodné výběry ze stejného základního souboru
- obdobně jako bodový a intervalový odhad parametrů základního souboru funguje i regresní přímka
- doplnění průběhu regresní přímky intervalem spolehlivosti - při vykreslení vzniknou „pásy spolehlivosti“
- určujeme interval v němž se pro dané x s danou pravděpodobností bude nacházet i hodnota y příslušná hodnotě x

Intervaly a pásy spolehlivosti lineární regresní závislosti

- zvolíme míru spolehlivosti pro pásy spolehlivosti
- poloviční šířka tohoto intervalu je dána vztahem

$$l = t_{1-\alpha}(n-2) \cdot \frac{h\sqrt{A}}{\sqrt{n-2}},$$

kde $h = \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{(n-1)S_x^2}}$

- hodnota $t_{1-\alpha}(n-2)$ je hodnota kvantilu Studentova t rozdělení $1-\alpha$ pro $n-2$ stupňů volnosti

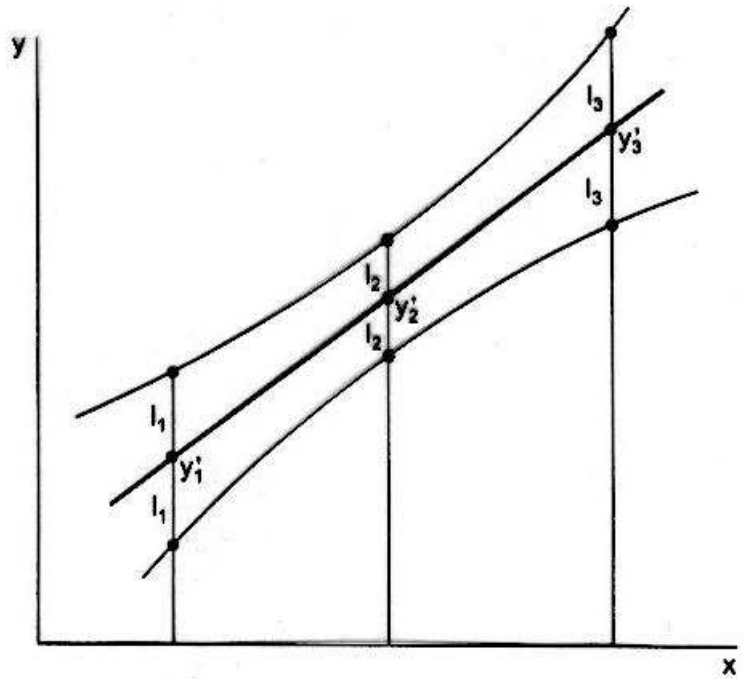
Pásy spolehlivosti

Dolní mez:

$$y' - l$$

Horní mez:

$$y' + l$$

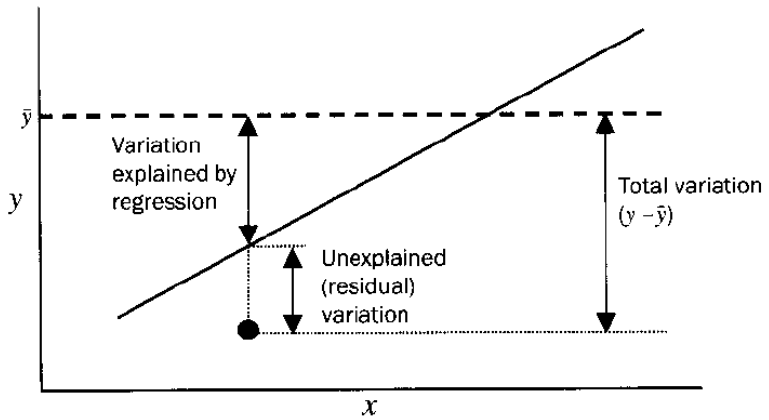


Testování významnosti regresní závislosti

- k testování lze využít jednovýběrového t-testu (nulová hypotéza bude tvrdit, že se směrnice přímky *b* neliší významně od nuly)
- častěji používáme ale analýzu rozptylu
  - zjistíme celkovou variabilitu hodnot *y*
  - vypočítáme, z kolika procent je tato celková variabilita vysvětlena variabilitou hodnot *x*
  - celková variabilita: celková suma čtverců odchylek hodnot od průměru
  - rozdělíme ji na variabilitu regresní (tj. vysvětlena regresní přímkou) a variabilitu reziduální (tj. zbytkovou - nevysvětlenou regresním modelem)

Testování významnosti regresní závislosti

- konkrétním způsobem se však již nebudeme zabývat (postačí postup v systému STATISTICA)



$$SS_{reziduální} = SS_{total} - SS_{regrese}$$

Příklad

Příklad z přednášky Korelační analýza

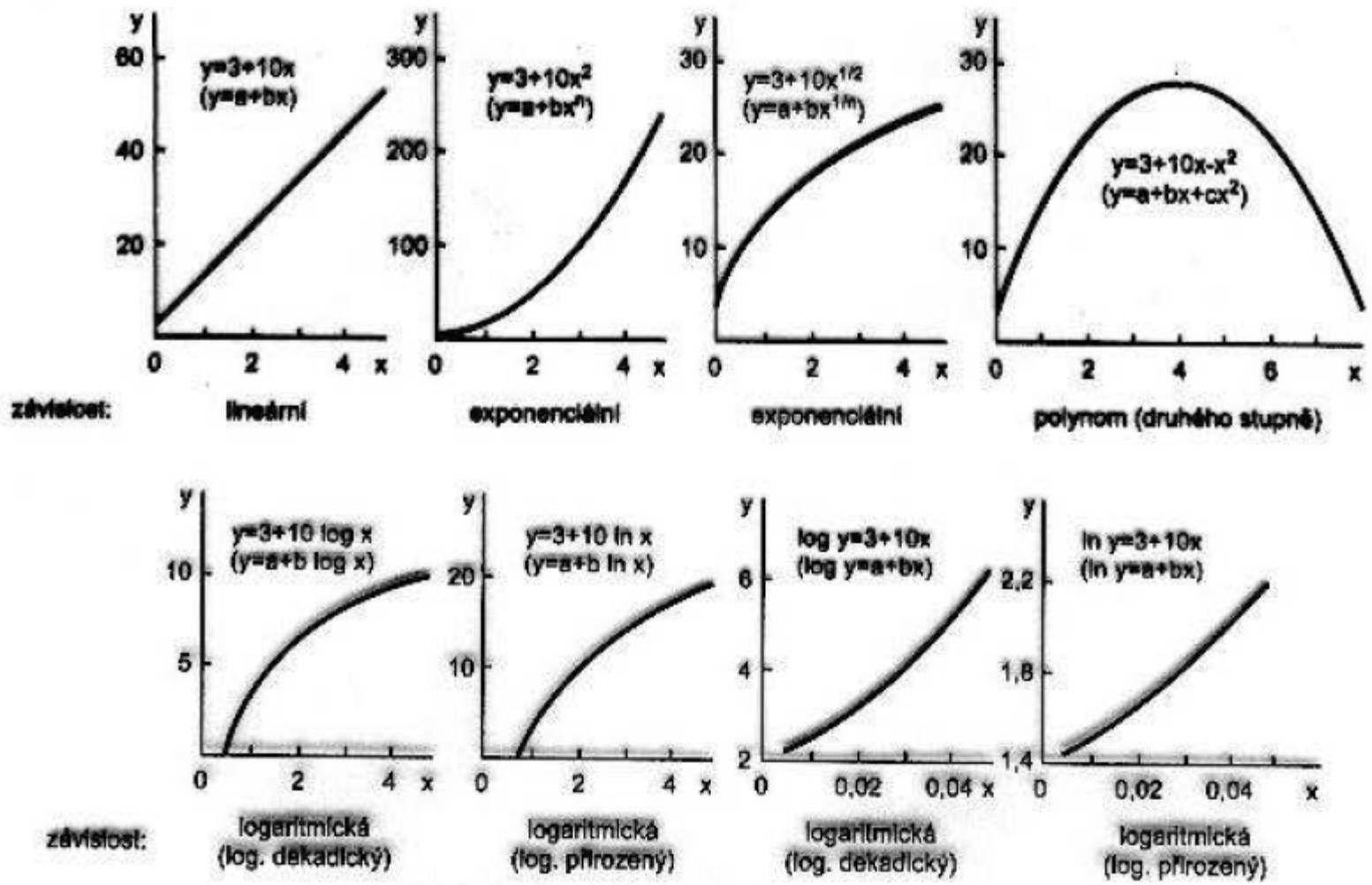
Jaká je závislost mezi pH půdy na výsypkách a počtem rostlinných druhů?

x	y	x^2	y^2	xy
2.8	17	7.8	289	47.6
2.9	7	8.4	49	20.3
3.8	10	14.4	100	38.0
4.5	22	20.3	484	99.0
7.1	40	50.4	1600	284.0
6.5	25	42.3	625	162.5
3.0	5	9.0	25	15.0
4.7	5	22.1	25	23.5
5.2	22	27.0	484	114.4
4.0	7	16.0	49	28.0
4.8	6	23.0	36	28.8
6.3	43	39.7	1849	270.9
7.2	19	51.8	361	136.8

Řešení

- *Statistiky–Vícenásboná regrese*
- volba proměnných - závisle a nezávisle proměnná
- *Výpočet: Výsledky regrese*
- *OK*, na záložce *Bodové grafy–Korelace 2 proměnných* - vykreslení grafu korelačního pole s pásy spolehlivosti a regresní přímkou

Další typy regresních funkcí - nelineární



Hledání vhodného regresního modelu

Postupovat lze dvěma způsoby:

- volba vhodného modelu na základě praktické zkušenosti či teoretických předpokladů
- posouzením bodového grafu a interpretací nástrojů regresní analýzy

Způsoby hodnocení vhodnosti regresního modelu:

- analýza reziduálních hodnot
- výpočet směrodatné chyby odhadu
- výpočet koeficientu determinace

Analýza reziduálních hodnot

- rezidua jsou vzdálenosti skutečných hodnot y_i od modelem odhadnutých hodnot $y´_i$
- model je vhodný, pokud reziduální hodnoty splňují následující podmínky
  1. rezidua jsou náhodná a nezávislá
  2. mají normální rozdělení s nulovým průměrem a konstantním rozptylem

Směrodatná chyba odhadu

- vyjadřuje směrodatnou odchylku, resp. rozptyl reziduálních hodnot a je vhodnou mírou pro posouzení vhodnosti použité regresní závislosti
- čím je hodnota reziduálního rozptylu nižší, tím je model vhodnější

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

Koeficient determinace

- zavedli jsme jej již dříve (korelační počet)
- čím je hodnota koeficientu determinace větší, tím je model vhodnější

$$r^2 = \frac{SS_{regres}}{SS_{total}}$$

Vícerozměrná regrese

- popisuje závislost více proměnných vysvětlujících jednu proměnnou
- v případě, že máme dvě vysvětlující proměnné, tak je regresní model rovinou
- odhad parametrů se opět provádí pomocí metody nejmenších čtverců
- $y' = a + b_1 x_1 + b_2 x_2 + \ldots$

Děkuji za pozornost...