Milí studenti,

projděte si nejprve soubor geost_4_odpovedi.pdf. Výsledky úkolů U1, U2, ... vložte do elaborátu nazvaného Vaším jménem a číslem lekce.

S pozdravem,

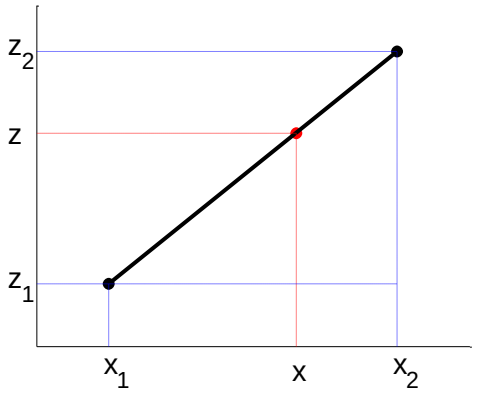
JJ

V předchozí lekci jsme dokončili metodu IDW, o které lze říci, že je historickým a ideovým předchůdcem geostatistiky (i když její současný název vznikl později). Pro geostatistiku je důležitá také **metoda nejmenších čtverců**, kterou se budeme zabývat v této lekci. Vlastně jsme na ni narazili na konci minulé lekce, když jsme prokládali přímku znázorňující výsledky procedury CV, kde jsme použili funkci `primka_LSQ`. Máte jistě také nějaké znalosti, které jste získali v základních kurzech matematiky a statistiky.

Půjdeme opět cestou, kterou lze nazvat „škola hrou“.

Proložení přímky 2 body

Na začátku 2. lekce jsme vyšli z následujícího obrázku a nalezli vzorec pro interpolaci v bodě *x*.



Nyní nás zajímá rovnice přímky procházející body [x_1, z_1] a [x_2, z_2] . Rovnici hledáme ve tvaru

$$z = a + bx$$

Víte, že *a* je úsek na ose *y* , *b* je směrnice přímky. Přímka prochází oběma datovými body, což nám dává dvě rovnice

$$a + bx_1 = z_1 \qquad (1)$$
$$a + bx_2 = z_2$$

Tuto soustavu rovnic zapíšeme maticově

$$\boldsymbol{LB} = \boldsymbol{P} \qquad (2)$$

kde ***L*** je matice levé strany soustavy

$$\boldsymbol{L} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \end{bmatrix}$$

B je sloupcový vektor parametrů přímky

$$\boldsymbol{B} = \begin{bmatrix} a \\ b \end{bmatrix}$$

a ***P*** je sloupcový vektor pravé strany soustavy

$$\boldsymbol{P} = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}$$

(pro jistotu zkuste dosazením ***L***, ***B*** a ***P*** do (2) a roznásobením ukázat, že dostanete soustavu (1)). Hledáme vektor **B**. Levou i pravou stranu soustavy (2) vynásobíme inverzní maticí k matici ***L*** a dostáváme řešení ve tvaru

$$\boldsymbol{B} = \boldsymbol{L}^{-1}\boldsymbol{P} \qquad (3)$$

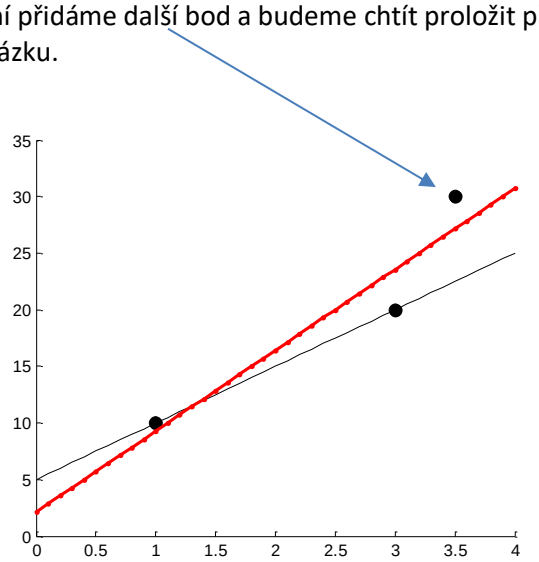
kde $\boldsymbol{L}^{-1}$ je inverzní matice.
K výpočtu inverzní matice je k dispozici řada metod (a některou z nich jste se patrně učili v základním kurzu matematiky). Nepřekvapí vás, že programy jako Matlab inverzní matice umějí počítat. Takže k praktickému proložení přímky stačí umět formálně sestavit matice ***L*** a ***P***. V Matlabu by tomu odpovídaly žlutě vyznačené řádky:

```
function primka(xd,zd,x)

% prolozeni primky dvema body

figure
plot(xd,zd,'o')
hold on

L=[ 1 xd(1);
    1 xd(2)] % matice leve strany soustavy rovnic

P=[zd(1) zd(2)]' % vektor prave strany

B=inv(L)*P % reseni soustavy

a=B(1)
b=B(2)

z=a+b*x

plot(x,z,'r-')
```

Funkce `inv` umí spočíst inverzní matici.

U1: Funkci `primka` vytvořte v Matlabu a spusťte na data xd=[1,3], zd=[10,20]. Vstupní vektor udává místa, kde se vypočtou hodnoty bodů na přímce. Zadáte-li např. x=[0,4], zobrazí se přímka v tomto rozsahu.

Nyní přidáme další bod a budeme chtít proložit přímku mezi 3 body, jak je naznačeno červeně na obrázku.



Obr. 1

Víte, že vhodná metoda k tomu se jmenuje **metoda nejmenších čtverců** (anglicky *Least Squares*, používaná zkratka LS nebo LSQ).

Také víte, že ve statistice by se proložení přímky měřenými body nazvalo **lineární regrese**. Statistický model vychází z představy, že přímka zobrazuje zákonitost, ke které se přidávají nahodilé odchylky (mohou to být chyby měření nebo nějaká dodatečná variabilita). Odpovídající vzorec je

$$z_i = a + bx_i + e_i \qquad (4)$$

Takže z_i jsou naměřené hodnoty a e_i náhodné odchylky, které ovšem neznáme. Předpokládáme, že jsou vzájemně nezávislé. Tu vlastnost jsem zdůraznil, neboť to v souvislosti s geostatistikou hraje velmi důležitou roli, jak uvidíme později.

Snažíme se proložit přímku tak, aby v nějakém smyslu byla co nejlépe mezi těmi body. Praktické je požadovat, aby odchylky bodů od ní byly „v součtu“ co nejmenší. Metoda nejmenších čtverců požaduje, aby byla minimální suma čtverců (kvadrátů) reziduí, tj. odchylek od přímky

$$S = \sum[z_i - (a + bx_i)]^2 = min \qquad (5)$$

Kdybychom měli přebytek času, mohli bychom zkoušet opakovaně volit parametry *a* a *b* přímky, jakoby hýbat přímkou v obrázku výše, a počítat tuto sumu. Za nejlepší přímku bychom vybrali tu, která dává sumu nejmenší.
Z kurzu matematiky ale víte, že to nemusíme takto pracně dělat. Suma je totiž kvadratickou funkcí parametrů *a* a *b*, které jsou naše *neznámé*. Stačí proto hledat minimum tak, že výraz pro sumu podle nich derivujeme a derivace položíme rovny nule

$$\frac{\partial S}{\partial a} = 0, \qquad \frac{\partial S}{\partial b} = 0$$

U2: Tužka papír. Pokuste se derivovat sumu (5) podle *a* a *b*. Výsledky vhodným způsobem upravte tak, abyste dostali soustavu dvou rovnic pro hledané parametry *a* a *b*. Výsledek mi vložte do elaborátu.

Pokud se to povedlo odvodili jste tzv. *normální rovnice*. Jejich zápis, resp. zápis levé a pravé strany soustavy v Matlabu by mohl být

```
nd=length(xd);

L=[ nd       sum(xd);
    sum(xd) sum(xd.*xd)] % matice leve strany soustavy

P=[sum(zd) sum(xd.*zd)]' % vektor prave strany
```

(apostrof je znaménku transpozice)

U3: Funci `primka` uložte jako `primka_LSQ_pokus`. Výše uvedenými 3 příkazy nahraďte tu část, kterou jsme dříve vyznačili žlutě. Doplňte data o 3. bod označený v obrázku výše šipkou a funkci spusťte. Výsledný obrázek mi vložte do elaborátu.

U4: Vyzkoušejte spuštění funkce na libovolná data obsahující více bodů a také na data sestávající pouze ze dvou bodů z U1.

Pokud se to povedlo, umíme proložit přímku metodou nejmenších čtverců 3 i více body a (možná k malému překvapení) i pouhými těmi dvěma body, se kterými jsme začínali (výsledek splývá s U1).

Základní vzorec lineární regrese

Pro lineární regresi existuje vzorec, který pokrývá nejen případ přímky, ale i mnoho dalších možností a je dobrý jak pro numerické výpočty, tak i pro teoretické úvahy. Lze ho odvodit analogickou cestou, jako jste se snažili v U2 a vypadá následovně

$$\boldsymbol{B} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{z} \qquad (7)$$

kde matice $\boldsymbol{X}$ se nazývá matice designu nebo regresní matice a její sloupečky regresory . V případě prokládání přímky má matice jen 2 sloupce, v prvním jsou jedničky a ve druhém souřadnice datových bodů

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ & \cdots \\ 1 & x_n \end{bmatrix} \qquad (8)$$

U5: Podívejte se do funkce, `primka_LSQ`, kterou jsem vám zaslal minule, a najděte tam řádek, kde je regresní matice vytvořena. Spusťte funkci `primka_LSQ` na stejná data jako v U3 a ověřte, že dává stejné výsledky.

Charakteristiky proložení

Připomeneme si některé věci, které znáte ze základního kurzu statistiky.

Když jsme proložili přímku, můžeme vyčíslit i sumu čtverců reziduí *S*. Ve statistických textech i programech se často označuje zkratkou *RSS* (*Residual Sum of Squares*)

$$RSS = \sum [z_i - (a + bx_i)]^2 = \sum {e_i}^2$$

Tato suma svojí velikostí říká jak moc datové body "varýrují" kolem přímky, neboli nakolik přímka shluk bodů vystihuje, nakolik je vhodným modelem studované veličiny. Suma je ovšem závislá na počtu dat a na jednotkách.

Rozumné je tedy ji vztáhnout k počtu dat, nejjednodušejí tím počtem dat vydělit. Z kurzu statistiky víte, že z teoretického hlediska je vhodnější vydělit sumu tzv. *počtem stupňu volnosti*. Tím dostáváme nevychýlený (anglicky *unbiased*; říká se též nestranný) odhad reziduálního rozptylu

$$s^2 = RSS/(n-k)$$

(*n* je počet dat, *k* je počet regresorů).

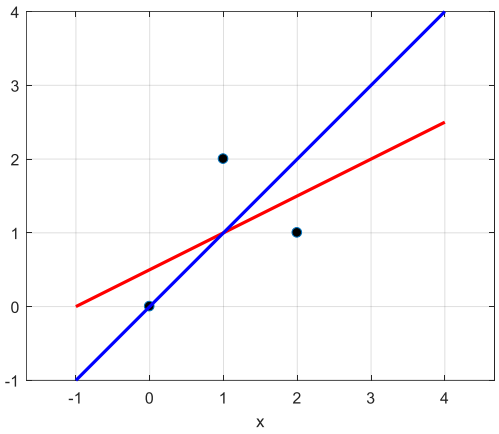
Reziduální rozptyl (σ^2), resp. jeho odmocnina, *směrodatná odchylka* (σ), určuje míru velikosti náhodných odchylek e_i ve vzorci (4). Tj. míru kolísání hodnot kolem přímky. V klasickém regresním modelu se předpokládá, že tyto odchylky mají normální rozdělení, $e_i \sim N(0,\sigma^2)$. To jest s 95% pravděpodobností jsou v intervalu $\pm 1.96\sigma$.

Abychom se zbavili závislosti na jednotkách, zavádí se *koeficient determinace*, který se označuje R^2

$$R^2 = 1 - \frac{RSS}{\sum (z_i - \bar{z})^2}$$

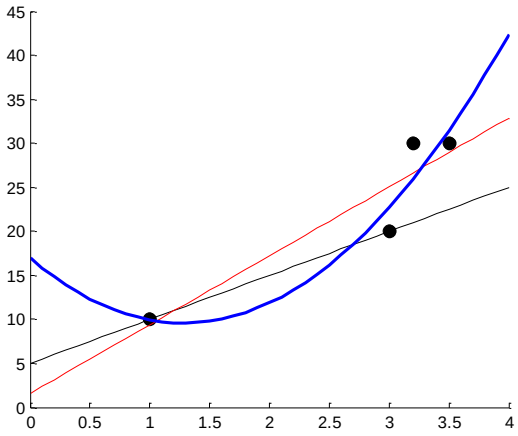
Koeficient determinace se zavádí i pro jiné situace nežli prokládání přímky (a je několik různých definic). Z našeho hlediska, s malým zjednodušením, můžeme říci, že hodnoty koeficientu determinace jsou v intervalu $\langle 0,1 \rangle$ a čím blíže jsou hodnoty R^2 k jedničce, tím je závislost lépe popsána přímkou (případně jinou křivkou či plochou) a datové body se k lépe přimykají.

U6: Na následujícím obrázku je trojice bodů proložených dvěma přímkami (jen 3 body, aby se to snadno spočítalo).
Napište rovnice obou přímek a zjistěte, která z nich vystihuje data lépe ve smyslu metody nejmenších čtverců.



Proložení polynomu – další regresor

Poté, co jsme zvládli proložit přímku, nebude problém proložit metodou nejmenších čtverců také polynom. Do obrázku obr. 1 přidáme další bod a chceme proložit polynom 2. stupně.



Rovnice polynomu je

$$z = a + bx + cx^2$$

Analogicky k postupu u přímky sestavíme sumu čtverců reziduí

$$S = \sum [z_i - (a + bx_i + cx_i^2)]^2 = min$$

Odvození nás dovede k regresní matici obsahující pouze jeden sloupeček navíc

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ & \cdots & \\ 1 & x_n & x_n^2 \end{bmatrix}$$

Pro výpočet koeficientů polynomu zůstává v platnosti vzorec (7)

$$\boldsymbol{B} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{z}$$

Povšimněme si, že formálně je to stále lineární regrese, jenom je více regresorů. Přidaný regresor je x^2. Sestává z kvadrátů souřadnic, ale závisle proměnná *z* na něm závisí lineárně prostřednictvím koeficientu *c*.

Otevřete zaslanou funkci `polynom_LSQ` a porovnejte její kód s funkcí `primka_LSQ` (zjistěte, kde se liší). Pokuste se funkci `polynom_LSQ` spustit tak, abyste vykreslili obrázek se 4 datovými body výše.Do elaborátu toto vkládat nemusíte, jen kdyby byl nějaký problém.

Závěrečná poznámka

Asi vás napadlo, že pro polynomy vyššího stupně bude stačit přidat další regresor

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_1 & x_1^2 & \dots \\ 1 & x_2 & x_2^2 & \dots \\ & & \dots & \\ 1 & x_n & x_n^2 & \dots \end{bmatrix}$$

A také, že v případě plošných dat, kdy máme navíc souřadnici *y*, bude regresní matice obsahovat sloupec s *y*, popř. jeho mocninami

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 & y_1 & x_1^2 & \dots \\ 1 & x_2 & y_2 & x_2^2 & \dots \\ & & \dots & & \\ 1 & x_n & y_n & x_n^2 & \dots \end{bmatrix}$$

Přečtěte si text GPI, str. 49-52. Prostorovou regresí se budeme zabývat příště.