EFEKTIVNÍ VYUŽITÍ MATLABU PŘI ŘEŠENÍ ÚLOH REGRESE

Jiří Militký
Technická universita v Liberci

1.Úvod

V praxi se pomocí regresních modelů řeší řada přírodovědných a technických úloh. Speciálně pro případ, kdy není modelový vztah apriori určen se vychází z lineárního regresního modelu tj. lineární kombinace vysvětlujících proměnných. Tato úloha vede obecně z numerického hlediska na problém řešení přeurčené soustavy lineárních rovnic. Pro běžný případ kritéria minima součtu čtverců odchylek se dá také převést na problém řešení soustavy lineárních rovnic se čtvercovou, symetrickou a často pozitivně definitní maticí. Numerické potíže zde vznikají v případě špatné podmíněnosti této matice. To je běžný případ např. u polynomických modelů, kdy dochází ke špatné podmíněnosti vlivem mocnin vysvětlující proměnné.

V tomto příspěvku jsou porovnány možnosti jazyka MATLAB pro řešení úloh lineární regrese s ohledem na možnou špatnou podmíněnost. Je použit jednoduchý příklad z knihy [1], který umožňuje měnit podmíněnost podle zadaného parametru. Je zmíněn program OLIN pro porovnání jednotlivých numerických metod [2].

2. Základy regrese

Regresní analýza umožňuje nalezení závislosti výstupní veličiny (odezvy) y na nastavované kombinaci hodnot m-tice vstupních proměnných $\boldsymbol{x} = (x_1, x_2..x_m)$.
Vychází se z naměřených hodnot y při různých kombinacích nastavovaných proměnných $x_1, x_2, ...x_m$. Jde vlastně o n-tici bodů $\{y_i, x_{ij}\}$, $j = 1, ..., m$, $i = 1, ..., n$, vyjádřených ve zkráceném maticovém zápisu $\{\boldsymbol{y}, \boldsymbol{X}\}$. Vektor $\boldsymbol{y}$ má rozměr (n×1) a matice $\boldsymbol{X}$ (n×m). Cílem statistické analýzy je objasnění variability měřené, výstupní ***závisle proměnné*** (vysvětlované) veličiny y s využitím regresní funkce $y = f(\boldsymbol{x}, \beta)$ obsahující nastavované, vstupní, ***nezávisle proměnné*** (vysvětlující) veličiny $\boldsymbol{x}$. Běžně se předpokládá, že veličina y je náhodná a veličiny $\boldsymbol{x}$ jsou nenáhodné a libovolně nastavovatelné. Dalším předpokladem je aditivní model měření, kdy se náhodné veličiny ε_i adují na regresní model. Omezme se na lineární regresní modely, kde je regresní model lineární v parametrech a obyčejně je přímo lineární kombinací vysvětlujících proměnných. Podmíněná střední hodnota proměnné y pro dané $\boldsymbol{x}$ (regrese) je pak ve tvaru

$$E(y/x) = \sum_{i=1}^{m} \beta_j x_j \qquad (1)$$

Odhady $\boldsymbol{b}$ parametrů β je pak možné určit metodou nejmenších čtverců, která bývá v praxi nejpoužívanější. Ukažme si geometrický význam této metody.

V případě platnosti aditivního modelu měření pro lineární regresní model je možné zapsat výsledky experimentů jednoduše s pomocí lineární kombinace sloupcových vektorů.

$$\mathbf{Y} = \mathbf{X} * \beta + \varepsilon \qquad (2)$$

$$\begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & . & x_{1m} \\ x_{21} & x_{22} & . & x_{2m} \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ x_{n1} & x_{n2} & . & x_{nm} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ . \\ . \\ \beta_m \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ \varepsilon_n \end{bmatrix} \qquad (3)$$

(nx1) ***(nxm)*** ***(mx1)*** ***(nx1)***

Sloupce x_j matice ***X*** definují z geometrického hlediska m-rozměrný souřadnicový systém resp. nadrovinu L v *n*-rozměrném eukleidovském prostoru E^n. Vektor ***y*** obecně neleží v nadrovině L, (viz. obr. 1 pro případ dvou nezávisle proměnných m = 2). V nadrovině L však leží všechny lineární kombinace sloupců matice ***X*** tj. vektory **X** β. Parametry β lze tedy chápat jako koeficienty úměrnosti u jednotlivých složek x_j souřadnicového systému (vysvětlujících proměnných) jejichž lineární kombinace tvoří regresní model. Bez ohledu na užité kritérium regrese bude tedy u lineárních regresních modelů ležet modelová funkce **X b** stejně jako teoretický model **X** β v m-rozměrné nadrovině L. Symbol ε označuje vektor chyb.

Metoda nejmenších čtverců (MNČ) hledá odhady parametrů ***b*** tak, aby byla minimalizována vzdálenost mezi vektorem ***y*** a nadrovinou L. To je ekvivalentní požadavku minimální délky vektoru reziduí

$$\mathbf{e} = \mathbf{y} - \mathbf{y_P} \qquad (4)$$

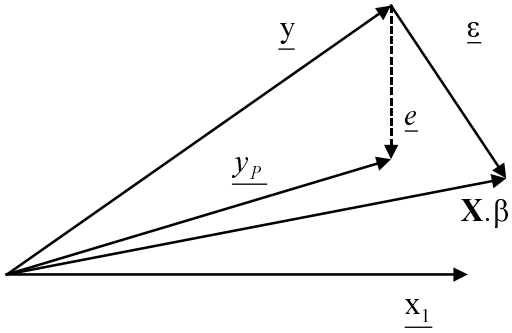
kde $\mathbf{y_p}$ = **X b** je vektor predikce. V eukleidovském prostoru lze délku vektoru e vyjádřit vztahem

$$d = \sqrt{\sum_{i=1}^{n} e_i^2} \qquad (5)$$

Čtverec délky vektoru ***e*** je tedy číselně shodný s hodnotou kriteriální podmínky *S(**b**)* metody nejmenších čtverců. Odhady modelových parametrů ***b*** pak minimalizují výraz

$$S(\mathbf{b}) = \sum_{i=1}^{n} \left[y_i - \sum_{j=1}^{m} x_{ij} b_j \right]^2 \qquad (6)$$

Vektory **e** a $\mathbf{y}_P$ jsou znázorněny na obr.1. Vektor $\mathbf{y}_P$ nazývaný ***vektor predikce*** představuje ***kolmou projekci*** vektoru ***y*** do nadroviny L. Vektor **e** nazývaný ***vektor reziduí*** leží v (n-m) rozměrné nadrovině L*, ***kolmé*** na nadrovinu L.



Obr. 1 Geometrie lineárního regresního modelu

Na základě tohoto geometrického znázornění lze hledat odhady parametrů ***b*** tak, aby byla minimalizována vzdálenost mezi vektorem ***y*** a nadrovinou L. Je patrné, že vektor reziduí **e** je kolmý na všechny sloupce matice **X**, a proto jsou odpovídající skalární součiny nulové.Tuto soustavu podmínek lze zapsat maticově jako

$$\mathbf{X}^T\mathbf{e} = 0 \qquad (7)$$

Po dosazení za **e** = **y - X b** a úpravě vyjde odhad **b**, minimalizující vzdálenost *d* ve tvaru

$$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \qquad (8)$$

kde symbol $\mathbf{A}^{-1}$ označuje inverzi matice **A**. Z rovnice (8) lze určit tvar projekční matice **H** pomocí které se promítá vektor ***y*** do nadroviny L**.** Tedy

$$\mathbf{y}_P = \mathbf{H}\,\mathbf{y} \qquad (9)$$

Pomocí vektoru **b** lze vyjádřit rovnici (9) ve tvaru

$$\mathbf{y}_P = \mathbf{X}\,\mathbf{b} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \qquad (10)$$

Projekční matice $\mathbf{H} = \mathbf{X}\,(\mathbf{X}^T\,\mathbf{X})^{-1}\,\mathbf{X}^T$ má tu vlastnost, že promítne libovolný vektor **V** do roviny L. Projekční matice **P** pro kolmou projekci do nadroviny L*, kolmé na nadrovinu L má tvar

$$\mathbf{P} = \mathbf{E} - \mathbf{H} \qquad (11)$$

kde **E** je jednotková matice. S využitím těchto projekčních matic lze provést rozklad vektoru **y** do dvou složek

$$\mathbf{y} = \mathbf{H}\mathbf{y} + \mathbf{P}\mathbf{y} = \mathbf{y}_P + \mathbf{e}$$

Geometricky to znamená, že vektor ***y*** byl rozložen na dva vzájemně kolmé vektory. Jeden souvisí s částí variability objasněné regresním modelem a druhý se zbytkovou (reziduální variabilitou). Ke stejným vztahům lze dospět analytickou minimalizací kritéria MNČ, tzn. derivováním rovnice (6) a dalšími úpravami.

3.Numerické aspekty MNČ

Určení odhadu **b** lineárního regresního modelu je podle rov. (3) úloha řešení **přeurčeného systému lineárních rovnic**, pro kterou má MATLAB operátor zpětné lomítko (\).

Lze také použít metodu SVD, která rozkládá libovolnou obdélníkovou matici X (nxm) na tři matice tj. $\mathbf{X} = \mathbf{U} * \mathbf{S} * \mathbf{V}^T$. Pomocí příkazu svd(x,0) se získá tzv. zkrácená SVD kterou uvažujeme v dalším (pro zkrácenou SVD se mění rozměry matic **U** a **S**) Pro zkrácenou SVD je matice **S** (mxm) diagonální a obsahuje na diagonále tzv. singulární čísla matice **X.** Pokud má matice **X** hodnost r (tj. obsahuje pouze r lineárně nezávislých sloupců) je právě r kladných nenulových singulárních čísel seřazených dle velikosti, tj. $S_{11} \geq S_{22} \geq S_{33} \geq ... \geq S_{rr}$. Matice **U** (nxm) a **V** (mxm) jsou ortogonální a normované, takže platí $\mathbf{U}^T\mathbf{U} = \mathbf{E}$ a $\mathbf{V}^T\mathbf{V} = \mathbf{E}$, kde **E** je jednotková matice.

Pro zkrácenou SVD platí, že kladná singulární čísla jsou odmocniny z vlastních čísel matice $\mathbf{X}^T\mathbf{X}$ (ale také matice $\mathbf{X}\mathbf{X}^T$), sloupce matice **U** jsou vlastní vektory matice $\mathbf{X}\mathbf{X}^T$ a řádky matice $\mathbf{V}^T$ jsou vlastní čísla matice $\mathbf{X}^T\mathbf{X}$. S využitím SVD lze rov. (3) vyjádřit ve tvaru

$$\mathbf{y} = \mathbf{U} * \mathbf{S} * \mathbf{V}^T\boldsymbol{\beta} + \boldsymbol{\varepsilon} \text{ resp. } \mathbf{y} = \mathbf{U} * \boldsymbol{\omega} + \boldsymbol{\varepsilon} \text{ kde } \boldsymbol{\omega} = \mathbf{S} * \mathbf{V}^T\boldsymbol{\beta}$$

Vektor $\boldsymbol{\omega}$ je stejného rozměru jako vektor $\boldsymbol{\beta}$ Vzhledem k ortogonalitě matice **U** lze získat odhad **o** parametrů $\boldsymbol{\omega}$ po dosazení do rov. (8), tedy

$$\mathbf{o} = (\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T\mathbf{y} \quad \text{tedy} \quad \mathbf{o} = \mathbf{U}^T\mathbf{y}$$

Protože však platí,že $\mathbf{o} = \mathbf{S} * \mathbf{V}^T\mathbf{b}$ vyjde pro odhad regresních parametrů, že

$$\mathbf{b} = (\mathbf{V}^T)^{-1}\mathbf{S}^{-1}\mathbf{o} \quad \text{resp.} \quad \mathbf{b} = \mathbf{V}\mathbf{S}^{-1}\mathbf{U}^T\mathbf{y}$$

Inverzní matice $\mathbf{S}^{-1}$ je pochopitelně také diagonální s prvky $S_{ii}^{-1} = 1/S_{ii}$ S na hlavní diagonále.

Označíme-li sloupce matice **U** jako $\mathbf{u}_j$ a řádky matice $\mathbf{V}^T$ jako $\mathbf{v}_j$ můžeme vyjádřit řešení úlohy lineární regrese v jednoduchém tvaru

$$b = \sum_{j=1}^{r} \frac{1}{S_{jj}} * \mathbf{u}_j * \mathbf{v}_j * \mathbf{y}$$

Pokud je r=m, jde o metodu klasických nejmenších čtverců. Pro r<m a r celé resultují tzv. odhady hlavních komponent a pro r necelé tzv. zobecněné vychýlené odhady.

Další možností je řešení soustavy tzv. normálních rovnic $\mathbf{X}^T\mathbf{X} * \mathbf{b} = \mathbf{X}^T\mathbf{y}$, vyjádřené rov. (8). Také pro tento případ lze s výhodou využít zabudovaných maticových operací jazyka MATLAB, kdy stačí použít příkaz b=inv(x'*x)*x'*y. Příkaz inv(A) realizuje inverzi čtvercové pozitivně definitní matice A Místo příkazu inv(x) lze použít příkaz X=pinv(A), který využívá Moore - Penrosovy pseudoinverse (platí,že A*X*A = A, X*A*X = X). Pro inverzi matice $\mathbf{X}^T\mathbf{X}$ je také možné využít rozkladu na vlastní čísla a vlastní vektory s využitím příkazu [P L]=eig(A), kde A= $\mathbf{X}^T\mathbf{X}$. Zde sloupce $\mathbf{p}_j$ matice **P** jsou vlastní vektory matice $\mathbf{X}^T\mathbf{X}$ a diagonální prvky matice **L** jsou odpovídající vlastní čísla L_j setříděná od nejmenšího k největšímu.

S využitím vlastních čísel a vlastních vektorů lze psát

$$\mathbf{X}^T\mathbf{X} = \sum_{j=r}^{m} L_j * \mathbf{P}_j * \mathbf{P}_j^T \quad \text{a} \quad (\mathbf{X}^T\mathbf{X})^{-1} = \sum_{j=r}^{m} (1/L_j) * \mathbf{P}_j^T * \mathbf{P}_j$$

Zde r =1 pro klasickou MNČ a r je větší než jedna pro případ, že některá vlastní čísla jsou blízká nule, nulová nebo záporná. Pro dobře podmíněné úlohy, kdy jsou vlastní čísla matice $\mathbf{X}^T\mathbf{X}$ všechna kladná a dostatečně vzdálená od nuly vedou všechny výše uvedené varianty ke

stejnému řešení. Rozdíly se projevují u špatně podmíněných úloh, kdy jsou vlastní čísla matice $\mathbf{X}^T\mathbf{X}$ sice kladná, ale některá z nich jsou blízká nule.

4. Porovnání jednotlivých metod

Pro ilustraci vhodnosti jednotlivých výše uvedených postupů byl sestaven program „olin“, kde lze měnit podmíněnost matice $\mathbf{X}^T\mathbf{X}$ pomocí parametru ep.byl vybrán jednoduchý případ, kdy je známo přesné řešení.Proměnná **bk** obsahuje odhady parametrů počítané klasickou inverzí, proměnná **bp** obsahuje odhady počítané s využitím pseudoinverse, proměnná **bz** obsahuje odhady počítané s využitím zpětného lomítka, proměnná **bsm** obsahuje odhady počítané s využitím SVD, proměnná **brs** obsahuje odhady počítané s využitím rozkladu na vlastní čísla.

Program olin

```
% test linearni MNC
% data y    x1    x2
%      3     1      1
%      ep    ep     0
%      2ep   0     ep
%kde e je presnost (male cislo, zde ep=1e-15).
%korektni reseni je b1=1 a b2=2
clc;clear all;
ep=1e-16; %mozno menit dle testu
y=[3 ep 2*ep]';x=[1 1;ep 0;0 ep];[n m]=size(x);
%klasicke reseni
bk=inv(x'*x)*x'*y;
% zpetne lomitko
bz=x\y;
%pseudoinverse
bp=pinv(x'*x)*x'*y;
%SVD maticove
[U S V]=svd(x,0);for i=1:m
if S(i,i)>0;S(i,i)=1/S(i,i);
else S(i,i)=0;
end;end
bsm=V*S*U'*y;
%racional slozkove
[V S]=eig(x'*x);for i=1:m
if S(i,i)>0;S(i,i)=1/(S(i,i));
else S(i,i)=0;
end;end
brs=zeros(m,1);for j=1:m
pom=V(:,j)*V(:,j)';
brs=brs+S(j,j)*pom*x'*y;
end
fprintf('vysledky \n');s='       bk       bz       bp       bsm       brs';s1=[bk bz bp bsm brs];
disp(s);disp(s1);
```

Pro volbu **ep=1e-7** jsou výsledky pro všechny metody prakticky ve shodě s přesným řešením

Vysledky¨

bk	**bz**	**bp**	**bsm**	**brs**
0.9996	1.0000	0.9790	1.0000	0.9996
2.0004	2.0000	2.0210	2.0000	2.0004

Pro rozmezí **ep=1e-8** až do **ep=1e-15** vycházejí již podstatné rozdíly mezi jednotlivými postupy

Vysledky¨

bk	**bz**	**bp**	**bsm**	**brs**
NaN	1.0000	1.5000	1.0000	1.5000
NaN	2.0000	1.5000	2.0000	1.5000

Pro **ep=1e-16** resp. menší (až do **realmin = 2.2251e-308**) vycházejí vždy stejně

Vysledky¨

bk	**bz**	**bp**	**bsm**	**brs**
NaN	3.0000	1.5000	1.0000	1.5000
NaN	0	1.5000	2.0000	1.5000

Je patrné, že použití příkazu inv(.) vede ke zkolabování výpočtu. V ostatních případech vede i špatná podmíněnost k výsledkům, které jsou však pro pseudoinverzi a rozklad na vlastní čísla a vlastní vektory vychýlené. Nelépe vychází postup využívající SVD, který je stabilní. Překvapivě dobré výsledky je možné získat použitím operátoru zpětného lomítka. Rozdíly jsou zde způsobeny numerickými metodami použitými u jednotlivých algoritmů.

Z uvedeného je patrné, že zdánlivě jednoduchá úloha lineární regrese může být v reálné situaci komplikovaná a její řešení může vyžadovat speciální postupy. Pro indikaci těchto potíží se používá metod pro indikaci tzv. multikolinearity (tj. přibližně lineární závislosti mezi sloupci matice X). Většina charakteristik multikolinearity vychází z prvků korelační matice vysvětlujících proměnných, kterou lze v MATLABu vyčíslit s využitím příkazu R=corrcoef(x). Mez základní míry multikolinearity patří:

1. prvky korelační matice CC (čím jsou větší, tím je multikolinearita vyšší)
2. VIF faktory, což jsou diagonální prvky inverzní matice R^{-1} (čím jsou větší, tím je multikolinearita vyšší, silná multikolinearita je pro VIF větší než 100)
3. korelační koeficienty regrese x_j na ostatních vysvětlujících proměnných **X,** pro které platí $R_j^2 = 1 - 1/VIF_j$ (čím jsou větší, tím je multikolinearita vyšší).

Pro výpočet charakteristik multikolinearity byl sestaven program „akoli“. Tento program byl použit pro testová data z programu olin.

Pro volbu ep=1e-7 vyšlo
**********.
VIF faktory .
VIF = 1. 3.336e+013.
VIF = 2. 3.336e+013.

a pro volbu ep=1e-4 vyšlo
**********.
VIF faktory .
VIF = 1. 3.333e+007.
VIF = 2. 3.333e+007.

Je patrné, že výrazná multikolinearita je indikovaná podstatně dříve,než dojde ke kolapsu i pro standardní metodu řešení MNČ.

5. Závěr

Je patrné, že i pro případy, kdy multikolinearita je poměrně vysoká vedou také jednoduché zabudované příkazy k akceptovatelným výsledkům [2]. Pro extrémní případy však může dojít ke kolapsu. Výhodné je použití SVD, kdy lze počítat také celou řadu dalších charakteristik a zejména řídit možné vychýlení odhadů (viz.[1]). Díky kvalitním maticovým operacím nejsou numerické problémy lineární regrese tak významné jak je běžné u jiných jazyků. Přesto lze doporučit aby při konstrukci programů pro lineární regresi bylo použito jak kritérií pro indikaci multikolinearity tak i metod umožňujících analýzu vlastních čísel.

Poděkování: *Tato práce vznikla s podporou výzkumného centra Textil LN00B090*

6. Literatura

[1] Meloun M., Militký J.*: Zpracování experimentálních dat*, East Publishing Praha 1998
[2] Militký J.,: *MATLAB a analýza dat*, elektronická příručka, TU Liberec 2002

Kontakt: Prof. Ing. Jiří Militký CSc,
Katedra textilních materiálů,
Technická universita v Liberci,
Hálkova 6
61 17 **Liberec**,
e- mail: jiri.miliky@vslib.cz