

Aplikace matematiky

Pavel Bureš

Algoritmy. 15. ANAREG. Mnohonásobná lineární regrese

Aplikace matematiky, Vol. 13 (1968), No. 4, 361–365

Persistent URL: <http://dml.cz/dmlcz/103180>

Terms of use:

© Institute of Mathematics AS CR, 1968

Institute of Mathematics of the Czech Academy of Sciences provides access to digitized documents strictly for personal use. Each copy of any part of this document must contain these *Terms of use*.



This document has been digitized, optimized for electronic delivery and stamped with digital signature within the project *DML-CZ: The Czech Digital Mathematics Library* <http://dml.cz>

ALGORITMY

15. ANAREG

MNOHONÁSOBNÁ LINEÁRNÍ REGRESE

PAVEL BUREŠ, prom. mat., Ústav výpočtové techniky ČSAV a ČVUT Horská 3, Praha 2

Používá se metody nejmenších čtverců, která vede na soustavu lineárních rovnic a ta se řeší upravenou Gauss-Jordanovou metodou.

procedure ANAREG ($n, m, f, data$); **value** n, m, f ;

integer n, m ; **real** f ; **array** $data$;

comment n je počet náhodných veličin, m je počet pozorování v každé veličině, f je číslo z intervalu (0,3) a udává, zda veličina má být přijata do regresní závislosti, $data$ je identifikátor pole s rozměry $n + 2$ a m , které obsahuje $n + 2$ m -rozměrných vektorů jednotlivých pozorování. Na prvních n místech jsou vektory nezávisle proměnných, na $n + 1$ místě je vektor závisle proměnné veličiny a na $n + 2$ místě je vektor koeficientů vah (obvykle jsou to jednotky);

begin

real array mat [$1 : n + 1, 1 : n + 1$], $prum, odch, b$ [$1 : n + 1$];

real $a, b0, sy, vmin, vmax, v, fi, c1, c2, c3, c4, c5, yn, roz, pom$;

integer $i, j, k, nmin, nmax$;

boolean p ;

$a := 0$;

for $i := 1$ **step** 1 **until** m **do** $a := a + data$ [$n + 2, i$];

for $i := 1$ **step** 1 **until** $n + 1$ **do**

begin $prum$ [i] := 0;

for $j := 1$ **step** 1 **until** m **do** $prum$ [i] := $prum$ [i] +
 $data$ [$n + 2, j$] \times $data$ [i, j];

$prum$ [i] := $prum$ [i] / a

end i ;

for $i := 1$ **step** 1 **until** $n + 1$ **do**

for $j := 1$ **step** 1 **until** i **do**

begin mat [i, j] := 0;

for $k := 1$ **step** 1 **until** m **do** mat [i, j] := mat [i, j] + $data$ [i, k]
 \times $data$ [j, k] \times $data$ [$n + 2, k$];

mat [i, j] := mat [i, j] - $prum$ [i] \times $prum$ [j] \times a ;

end j ;

```

for  $i := 1$  step 1 until  $n + 1$  do  $odch [i] := sqrt (mat [i, i]);$ 
comment lze tisknout součet vah  $a$ , průměry  $prum [i]$  a odchylky  $odch [i]$  jednotlivých proměnných;
for  $i := 1$  step 1 until  $n + 1$  do
    for  $j := 1$  step 1 until  $i$  do  $mat [j, i] := mat [i, j] := mat [i, j]/odch [i]$ 
     $/odch [j];$ 
comment lze tisknout korelační matici  $mat [i, j];$ 
 $fi := a - 1;$ 
11 : for  $i := 1$  step 1 until  $n$  do  $b [i] := 0;$ 
     $sy := odch [n + 1] \times sqrt (mat [n + 1, n + 1]/fi); vmin = 10^4;$ 
     $vmax := nmin := nmax := 0;$ 
    for  $i := 1$  step 1 until  $n$  do
        begin
            if  $mat [i, i] < 10^{-6}$  then go to l2;
             $v := mat [i, n + 1] \times mat [n + 1, i]/mat [i, i];$ 
            if  $v = 0$  then go to l2; if  $v > 0$  then go to l3;
             $b [i] := mat [i, n + 1] \times odch [n + 1]/odch [i];$ 
            if  $abs (v) \geq abs (vmin)$  then go to l2;  $vmin := v; nmin := i;$ 
            go to l2;
13: if  $v \leq vmax$  then go to l2;  $vmax := v; nmax := i; b [i] := 0;$ 
            l2 : end  $i;$ 
             $b0 := 0; \text{for } i := 1 \text{ step } 1 \text{ until } n \text{ do } b0 := b0 + b [i] \times prum [i];$ 
             $b0 := prum [n + 1] - b0;$ 
            if  $abs (vmin) \times fi/mat [n + 1, n + 1] < f$  then
                begin  $k := nmin; fi := fi + 1; \text{go to } l4$  end;
if  $vmax \times fi/(mat [n + 1, n + 1] - vmax) < f$  then go to l5;
             $k := nmax; fi := fi - 1;$ 
14 : for  $i := 1$  step 1 until  $n + 1$  do begin
            for  $j := 1$  step 1 until  $n + 1$  do begin
                 $p := \text{true}; \text{if } j = k \text{ then begin } c1 := 0; c2 := 1; p := \text{false} \text{ end}$ 
                else begin  $c1 := mat [i, j]; c2 := mat [k, j]$  end;
            if  $i = k$  then begin if  $p$  then  $b [j] := c2/mat [k, k]$  else
            begin  $c4 := c2/mat [k, k]; c3 := mat [k, k]$  end end
            else if  $p$  then  $mat [i, j] := c1 - c2 \times mat [i, k]/mat [k, k]$ 
            else  $c3 := c1 - c2 \times mat [i, k]/mat [k, k];$ 
            end  $j; mat [i, k] := c3;$ 
            end  $i;$ 
            for  $i := 1$  step 1 until  $n + 1$  do  $mat [k, i] := b [i];$ 
             $mat [k, k] := c4; \text{go to } l1;$ 
15 : comment zde tiskneme rozptyl regresní proměnné  $sy$  o absolutní koeficient regrese  $b0;$ 
            for  $i := 1$  step 1 until  $n$  do

```

```

begin if  $b[i] = 0$  then go to l6;
     $c1 = sy \times \text{sqrt}(\text{mat}[i, i]) / \text{odch}[i]$ ;
    comment zde tiskneme podle pořadí  $i$  regresní koeficienty  $b[i]$  a jejich roz-
    ptyly  $sb_i$  jež jsou postupně počítány v  $c1$ ;
l6 : end  $i$ ;
 $c1 := c2 := c3 := 0$ ;
for  $i := 1$  step 1 until  $m$  do begin
 $yn := 0$ ;
for  $j := 1$  step 1 until  $n$  do  $yn := yn + b[j] \times \text{data}[j, i]$ ;
 $yn := b0 + yn$ ;  $c1 := c1 + yn \times \text{data}[n + 1, i] \times \text{data}[n + 2, i]$ ;
 $\text{roz} := yn - \text{data}[n + 1, i]$ ;
comment zde lze tisknout spočtenou regresní proměnnou, jejíž hodnoty nabývá
postupně  $yn$  a rozdíly  $e_i$  mezi  $yn_i$  a veličinou  $y_i$   $\text{roz}$ ;
 $c2 := c2 + \text{roz} \uparrow 2$ ; if  $i = 1$  then  $\text{pom} := \text{roz}$ ;
 $c3 := c3 + (\text{roz} - \text{pom}) \uparrow 2$ ;  $\text{pom} := \text{roz}$ 
end  $i$ ;
 $c4 := \text{odch}[n + 1] \uparrow 2 + \text{prum}[n + 1] \uparrow 2 \times a$ ;
 $c5 := \text{odch}[n + 1] \uparrow 2$ ;
 $c2 := c3 / c2$ ;  $c3 := (c4 - c1) / (m - n)$ ;  $c5 := 1 + (c1 - c4) / c5$ ;
 $c4 := c1 / c4$ ;
comment lze tisknout Durban-Watsonovu statistiku  $d$  v  $c2$ , míru variance  $S_Y^2$  v  $c3$ ,
míru korelace  $R1$  a  $R2$  tj.  $c4$  a  $c5$ ;
end proc anereg;

```

Algoritmus je určen hlavně pro nalezení lineárních závislostí mezi náhodnými veličinami. Při použití transformací ho lze též využít pro aproximaci hledané funkce, známe-li předem její tvar.

Přesnost výsledků závisí na počtu pozorování m a na přesnosti vnitřního zobrazení čísla v počítači.

Vektory označujeme velkými písmeny, jejich složky odpovídajícími malými písmeny s indexy.

Mějme m pozorování od každé náhodné veličiny X_1, X_2, \dots, X_n, Y a dále m rozměrný vektor koeficientů vah ω (obvykle $\omega_j = 1$ pro $j = 1, 2, \dots, m$). Tyto údaje jsou zadány v dvojrozměrném poli *data*. Hledá se lineární závislost mezi veličinami Y a X_1, \dots, X_n tak, aby výraz

$$\sum_{j=1}^m (y_j \omega_j - \sum_{i=1}^n b_i x_{ij} \omega_j)^2$$

byl minimální. Tato metoda nejmenších čtverců vede na řešení soustavy n lineárních rovnic, které se řeší upravenou Gauss-Jordanovou metodou.

Algoritmus počítá průměry μ_{X_i} resp. μ_Y a směrodatné odchylky σ_{X_i} resp. σ_Y náhodných veličin X_1, \dots, X_n resp. Y ;

dále matici korelačních koeficientů mezi všemi náhodnými veličinami

$$\rho_{X_i X_k} = \frac{\sqrt{\left(\frac{1}{m} \sum_{j=1}^m (x_{ij} - \mu_{X_i})(x_{kj} - \mu_{X_k})\right)}}{\sigma_{X_i} \sigma_{X_k}}$$

a

$$\rho_{X_i Y} = \frac{\sqrt{\left(\frac{1}{m} \sum_{j=1}^m (x_{ij} - \mu_{X_i})(y_j - \mu_Y)\right)}}{\sigma_{X_i} \sigma_Y}.$$

Regresní proměnná je pak dána vztahem

$$Y_n = b_0 + \sum_{i=1}^n b_i X_i.$$

Algoritmus dále počítá její rozptyl s_y ,

koeficienty regrese b_i , jejich rozptyly sb_i a koeficient b_0 , dále rozdíl mezi veličinou Y a vypočtenou regresní proměnnou Y_n , $e_j = y_n j - y_j$, kde $j = 1, 2, \dots, m$.

Durban-Watsonovu statistiku

$$d = \frac{\sum_{j=2}^m (e_j - e_{j-1})^2}{\sum_{j=1}^m e_j^2}$$

je možno použít k testování hypotézy o normálním rozložení veličiny Y . V případě, že $d \sim 2$ nelze tuto hypotézu zamítnout. Naposledy se počítá míra variance

$$S_Y^2 = \frac{\sum_{j=1}^m (\omega_j y_j^2 - \omega_j y_n j y_j)}{m - n}$$

a míry korelace

$$R_1 = \frac{\sum_{j=1}^m y_n j y_j \omega_j}{\sum_{j=1}^m \omega_j y_j^2},$$

$$R_2 = 1 + \frac{\sum_{j=1}^m y_n j y_j \omega_j - \sum_{j=1}^m \omega_j y_j^2}{\sum_{j=1}^m \omega_j y_j^2 - \left(\sum_{j=1}^m \omega_j y_j\right)^2 / \sum_{j=1}^m \omega_j}.$$

Z technických důvodů byla v algoritmu použita jiná označení, než se ve statistice běžně užívají. Přehled těchto odchylek je následující:

$$\begin{array}{lll} \text{prum } [i] & \dots \mu_{X_i} & \text{mat } [i, k] \dots \varrho_{X_i X_k} \\ \text{prum } [n + 1] & \dots \mu_Y & \text{mat } [i, n + 1] \dots \varrho_{X_i Y} \\ \text{odch } [i] & \dots \sigma_{X_i} & b[i] \dots b_i \\ \text{odch } [i] & \dots \sigma_Y & \text{pro } i = 1, 2, \dots n; \quad k = 1, 2, \dots n \end{array}$$

Kontrolní příklad

Vstupní údaje:

$$\begin{aligned} n &= 11; \\ m &= 11; \\ f &= 0.5; \end{aligned}$$

prvky pole *data*:

$$\begin{aligned} X &= 2.063, 1.721, 1.403, 1.125, 0.898, 1.586, 1.376, 1.142, 1.268, 1.103, 1.019; \\ Y &= 3.601, 3.192, 2.892, 2.581, 2.326, 3.117, 2.886, 2.616, 2.805, 2.627, 2.585; \\ \omega &= 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1; \end{aligned}$$

Vypočtené hodnoty:

$$\begin{aligned} a &= 11; \text{prum } [1] = 1.3367; \text{prum } [2] = 2.8389; \text{odch } [1] = 1.0857; \\ \text{odch } [2] &= 1.1303; \\ \text{mat } [1,2] &= 0.9937; \\ s_y &= 0.0423; \end{aligned}$$

Regresní rovnice:

$$Y_n = 1.456 + 1.0345X; \text{ tj. } b_0 = 1.456; b = 1.0345;$$

rozptyl regresního koef. $sb_1 = 0.038965$;

$$Y_n = 3.5903, 3.2364, 2.9075, 2.6199, 2.385, 3.0968, 2.8795, 2.6375, 2.7678, 2.5971, 2.5102;$$

Vektor rozdílů $(Y_n - Y) = E$

$$E = -0.0107, +0.0444, -0.0155, +0.0389, +0.059, -0.0202, -0.0065, +0.0215, -0.0372, -0.0299, -0.0748;$$

$$d = 1,0926; S_Y^2 = 0.0018;$$

$$R_1 = 0.9998;$$

$$R_2 = 0.9874;$$

Algoritmus byl naprogramován ve strojním kódu počítače URAL 2 jako standardní program a ověřen v jazyku Algol.

[1] *Ralston - Wilf*: Numerical Methods of Digital Computers.

[2] *A. Hald*: Statistical Theory with Engineering Applications.

[3] *Lucy J. Slater*: Regression Analysis (The Computer Journal Vol. IV., No. 4).