ZASTOSOWANIA MATEMATYKI
APPLICATIONES MATHEMATICAE
XV, 2 (1976)

ALGORITHM 44

ANNA BARTKOWIAK (Wrocław)

AN ABBREVIATED METHOD OF CALCULATING THE MAHALANOBIS DISTANCE OR RESIDUAL SUM OF SQUARES IN A LINEAR REGRESSION MODEL

1. Procedure declaration. Procedure *linres* calculates, for a given vector $\boldsymbol{b}$ and a symmetric Grammian matrix $\boldsymbol{C}$, the value $z = y - \boldsymbol{b}'\boldsymbol{C}^{-1}\boldsymbol{b}$ involved in residual sum of squares for a linear regression model, multiple or partial correlation coefficients, Hotelling's T^2 statistics or Mahalanobis distance D^2.

The procedure operates only on the one-dimensional simulation of the lower triangle of the matrix $\boldsymbol{C}$ which is virtually the (pooled) covariance matrix. Procedure *linres* does not calculate the whole inverse $\boldsymbol{C}^{-1}$, but following an abbreviated form of the modified Gauss-Jordan algorithm performs only the operations necessary to obtain just the value of z for the declared variables. Dependent variables cause no perturbations — they are simply omitted.

Procedure *linres* is stepwise, it can be called several times, but only for strictly increasing values of *last* and *q1*, and with $q2 \geqslant q1$. The value *last* and the matrix $\boldsymbol{C}$ must not be destroyed between successive calls.

Data:

n — order of matrix $\boldsymbol{C}$ (the number of predictor variables $x_1, \ldots, x_{n-1}$ plus one criterion variable y in the case of regression; the number of variables under consideration plus one row allowing for inplantation of the vector $\boldsymbol{b}$ and the value of y otherwise);

$q1 \leqslant q2$ — integers indicating row numbers of variables for which the desired statistics are to be calculated; the values *q1* and *q2* must be non-decreasing, and $q1 > last$;

c[*1*:(*n*+*1*)×*n*÷*2*] — array containing the covariance matrix C and the vector $\boldsymbol{b}$ in the following row order:

$$\begin{array}{llll} C_{11} & & & \\ C_{21} & C_{22} & & \\ \cdots & \cdots & \cdots & \\ C_{n-1,1} & C_{n-1,2} & \ldots\ C_{n-1,n-1} & \\ b_1 & b_2 & b_{n-1} & y \end{array}$$

if the residual sum of squares is required, put the value of y equal to the total sum of squares; in calculations involved with T^2 or D^2 put y equal to 0;

eps — small floating-point constant, depending on the accuracy of the computer used;

last — when entering the procedure the first time, the value *last* must be set equal to zero, else it indicates the last computed variable; subsequent calls of *linres* are feasible only in the case where the entering value of *q1* is greater than the last calculated value *last*.

Results:

ind[*1*:*n*−*1*] — array indicating whether the operations needed for introduction of the declared variables into the calculated statistics were executable; *ind*[*i*] ($q1 \leqslant i \leqslant q2$) takes the value 1 if the variable number i could be dealt with by the algorithm, and 0 otherwise;

last — number of the variable dealt with *last* in the procedure;

linres — calculated value $z = y - \boldsymbol{b}'\boldsymbol{C}^{-1}\boldsymbol{b}$.

2. Method used. Procedure *linres* uses the modified Gauss-Jordan algorithm performing transformations T_q on the covariance matrix C_{ij} as follows (C'_{ij} denote the elements of C_{ij} after transformation):

$$C'_{ij} = C_{ij} - C_{iq}C_{jq}/C_{qq}, \quad i = q+1, \ldots, n; \ j = q+1, \ldots, i.$$

The transformations T_q are performed for $q = q1, \ldots, q2$ under the condition that $C_{qq} > eps$. If this is true, the value of *ind*[*q*] is set equal to 1. If $C_{qq} \leqslant eps$, the value of *ind*[*q*] is set equal to 0 and the next admissible value of q is considered.

3. Certification. Linear regression. We apply *linres* when only the residual sum of squares and not the coefficients of regression are required. The results quoted in the sequel have been obtained on the Odra 1204 computer in floating-point arithmetic with 11 decimal places of accuracy.

```
real procedure linres(n,q1,q2,c,eps,ind,last);
 value n,q1,q2,eps;
 integer n,q1,q2,last;
 real eps;
 integer array ind;
 array c;
 if q1>last
  then
  begin
   array d[1:n];
   integer i,j,k,l,q;
   real x,y;
   k:=q1×(q1-1)+2;
   for q:=q1 step 1 until q2 do
    begin
     k:=k+q;
     x:=c[k];
     if abs(x)≤eps
      then ind[q]:=0
      else
      begin
       ind[q]:=1;
       x:=-1.0/x;
       l:=k+q;
       for i:=q+1 step 1 until n do
        begin
         y:=d[i]:=c[l];
         y:=y×x;
         for j:=q+1 step 1 until i do
          begin
```

```
              l:=l+1;
              c[l]:=c[l]+y×d[j]
             end j;
            l:=l+q
           end i
         end abs(x) gt eps
       end q;
       last:=q2;
       linres:=c[n×(n+1)+2]
     end linres
```

Example 1 (Rao's example of prediction of cranial capacity ([3], p. 226-228)). Let us take the covariance matrix of three important measurements x_1, x_2, x_3 from which the cranial capacity y may be predicted. The matrix C is the following:

	x_1	x_2	x_3
x_1	0.01 875	0.00 848	0.00 684
x_2	0.00 848	0.02 904	0.00 878
x_3	0.00 684	0.00 878	0.02 886

The covariances of the dependent variable y with the measurements x_1, x_2, x_3 are represented by the vector

$$\sigma' = [\mathrm{cov}(y, x_1), \mathrm{cov}(y, x_2), \mathrm{cov}(y, x_3)] = [0.03\,030, 0.04\,410, 0.03\,629].$$

The total variance $\sigma_{yy} = 0.12\,692$.

Putting *last* $= 0$ and calling *linres* with the values

$$n = 4, \quad q1 = 1, \quad q2 = 3,$$

$$c = [0.01\,875, 0.00\,848, 0.02\,904, 0.00\,684, 0.00\,878, 0.02\,886, 0.03\,030, 0.04\,410, 0.03\,629, 0.12\,692],$$

$$eps = 10^{-10},$$

we get the following results:

$$linres = 0.02\,783, \quad ind = [1, 1, 1], \quad last = 3.$$

The calculated value *linres* can be interpreted as the residual sum of squares for the observed sample of size n, i.e.

$$linres = \sum_{i=1}^{n}\left(y_i - \sum_{l=1}^{3}\hat{b}_l x_{li}\right)^3 = \sigma_{yy} - \sigma' C^{-1}\sigma$$

(see Rao [3], formula 4g.1.10), where $(\hat{b}_1, \hat{b}_2, \hat{b}_3)$ denotes the estimators of coefficients of regression of y upon x_1, x_2, x_3.

4. Further examples of application.

Example 2. Multiple and partial correlation coefficients.

Using Rao's data from Example 1 we may easily calculate the multiple correlation coefficient between the variables x_1, x_2 and y. Putting ***last*** $= 0$ and calling ***linres*** with the values $n = 4$, ***q1*** $= 1$, ***q2*** $= 2$, and c and ***eps*** the same as in Example 1, we obtain ***linres*** $= 0.04\,130$, wherefrom we can easily calculate the square of multiple correlation coefficient

$$R^2_{y(1,2)} = \frac{0.12\,692 - 0.04\,130}{0.12\,692}$$

(see Rao [3], formula 4g.1.11).

To calculate the multiple correlation coefficient $R^2_{y(1,2,3)}$ after having calculated $R^2_{y(1,2)}$, we have to call ***linres*** once more with the values $n = 4$, ***q1*** $=$ ***q2*** $= 3$, and c, ***last*** and ***eps*** as in Example 1. We get as the result the value ***linres*** $= 0.02\,783$, wherefrom we calculate

$$R^2_{y(1,2,3)} = \frac{0.12\,692 - 0.02\,783}{0.12\,692},$$

following the same formula as previously.

The square of the partial multiple correlation coefficient expressing the reduction of variance by the variable x_3 after eliminating the association by x_1, x_2 is easily calculated by the formula

$$R^2_{y(3)(1,2)} = \frac{R^2_{y(1,2,3)} - R^2_{y(1,2)}}{1 - R^2_{y(1,2)}}$$

(see Rao [3], formula 4g.2.2).

Example 3. Mahalanobis distance D^2, calculated on Fisher's data on *Iris versicolor*, quoted by Rao [3], p. 480.

The pooled covariance matrix for two species of *Iris versicolor* calculated for four characteristics is the following:

$$C = \begin{bmatrix} 0.195\,340 & 0.092\,200 & 0.099\,626 & 0.033\,055 \\ 0.092\,200 & 0.121\,079 & 0.047\,175 & 0.025\,251 \\ 0.099\,626 & 0.047\,175 & 0.125\,488 & 0.039\,586 \\ 0.033\,055 & 0.025\,251 & 0.039\,586 & 0.025\,106 \end{bmatrix}.$$

The differences between sample means based on 50 observations for each of two species are

$$d = [0.930,\ -0.658,\ 2.789,\ 1.080].$$

Calling *linres* with the values

$$n = 5, \quad q1 = 1, \quad q2 = 4,$$

$$c = [0.195\,340,\ 0.092\,200,\ 0.121\,079,\ 0.099\,626,\ 0.047\,175,\ 0.125\,488,$$
$$0.033\,055, 0.025\,251, 0.039\,586, 0.025\,106, 0.930, -0.658, 2.789, 1.080, 0],$$

$$eps = 10^{-10}, \quad ind = [\text{arbitrary}], \quad last = 0,$$

we get *linres* $= -102.8428$. Hence the Mahalanobis distance D^2 between the species under consideration is $D^2 = 102.8428$. (The value D^2 reported by Rao is $D^2 = 103.2119$. We checked the computations by applying the direct formula $D^2 = \boldsymbol{b}\boldsymbol{C}^{-1}\boldsymbol{b}$, and evaluating the inverse $\boldsymbol{C}^{-1}$ with the aid of *cholinversion2* [2], and got the same value $D^2 = 102.8428$.)

Example 4. Hotelling's T^2, calculated on data on verbal and performance scores, quoted by Morrison [1], p. 122.

The mean values for two scores under investigation based on measurements of 101 men and women are $\overline{\boldsymbol{x}} = [55.24,\ 34.97]$. The sample covariance matrix of the scores is

$$\boldsymbol{C} = \begin{bmatrix} 210.54 & 126.99 \\ 126.99 & 119.68 \end{bmatrix}.$$

We wish to test the hypothesis that the observations came from a population with mean vector $\boldsymbol{\mu} = [60,\ 50]$. The test statistic is $T^2 = n(\boldsymbol{\mu}-\overline{\boldsymbol{x}})'\boldsymbol{C}^{-1}(\boldsymbol{\mu}-\overline{\boldsymbol{x}})$. Calling *linres* with

$$n = 3, \quad q1 = 1, \quad q2 = 2,$$

$$c = [210.54,\ 126.99,\ 119.68,\ 4.76,\ 15.03,\ 0.0],$$

$$eps = 10^{-10}, \quad ind = [\text{arbitrary}], \quad last = 0,$$

we get *linres* $= -3.5390$. Hence $T^2 = 101 \times 3.5390 = 357.4390$. (The value reported by Morrison is $T^2 = 357.43$.)

5. Additional remarks. Procedure *linres* was checked by calculating the residual sum of squares by definition, i.e. calculating the inverse $\boldsymbol{C}^{-1}$ by the use of *cholinversion2* [1], a procedure operating also on the one-dimensional simulation of the lower triangle of the matrix $\boldsymbol{C}$, and then multiplying $\boldsymbol{b}'\boldsymbol{C}^{-1}\boldsymbol{b}$. The gain in time of run on the Odra 1204 computer is the following:

Time (in sec.) needed by *linres* and *cholinversion2*

number of characteristics	$p = 4$	$p = 9$	$p = 19$
cholinversion2	0.226	1.473	10.618
linres	0.118	0.664	4.577

References

[1] D. F. Morrison, *Multivariate statistical methods*, McGraw Hill, New York 1967.

[2] R. S. Martin, G. Peters and J. H. Wikinson, *Symmetric decomposition of a positive definite matrix*, Numerische Mathematik 7 (1965), p. 362-383.

[3] C. R. Rao, *Linear statistical inference and its applications*, Wiley, New York 1965.

INSTITUTE OF INFORMATICS
UNIVERSITY OF WROCŁAW
50-384 WROCŁAW

Received on 21. 12. 1974

ALGORYTM 44

ANNA BARTKOWIAK (Wrocław)

SKRÓCONY SPOSÓB OBLICZANIA ODLEGŁOŚCI MAHALANOBISA LUB ZMIENNOŚCI RESZTOWEJ W LINIOWYM MODELU REGRESYJNYM

STRESZCZENIE

Procedura *linres* oblicza dla danego wektora $\boldsymbol{b}$ i symetrycznej nieujemnie określonej macierzy $\boldsymbol{C}$ wartość wyrażenia $z = y - \boldsymbol{b}'\boldsymbol{C}^{-1}\boldsymbol{b}$, występującego we wzorach na zmienność resztową w liniowym modelu regresyjnym, w wielokrotnych lub cząstkowych współczynnikach korelacji, w statystyce T^2 Hotellinga i odległości D^2 Mahalanobisa.

Procedura wykorzystuje dolny trójkąt macierzy $\boldsymbol{C}$, która jest na ogół macierzą kowariancji między badanymi zmiennymi. Nie odwraca się całej macierzy $\boldsymbol{C}$, ale wykonuje się tylko te obliczenia, które są niezbędne do wyznaczenia wartości z. Wiersze macierzy $\boldsymbol{C}$, przedstawiające cechy (prawie) liniowo zależne od pozostałych, są automatycznie opuszczane, o czym informuje tablica *ind*.

Procedura *linres* może być wywoływana kilkakrotnie, przy czym każde następne wywołanie może dołączać nowe zmienne (wiersze) do obliczanej charakterystyki, korzystając z obliczeń poprzedniego wywołania. W takim przypadku należy zagwarantować, żeby wartości *last* i *c* nie uległy zmianie między kolejnymi wywoływaniami *linres* oraz żeby wartości *last* i *q1* były ściśle rosnące.

Dane:

n — rozmiar macierzy $\boldsymbol{C}$ (liczba zmiennych objaśniających występujących w równaniu regresji plus zmienna wynikowa; w przypadku obliczania T^2 lub D^2 — liczba rozpatrywanych zmiennych plus jeden);

$q1 \leqslant q2$ — numery zmiennych, dla których ma być obliczana charakterystyka z;

$c[1: n \times (n+1) \div 2]$ — tablica zawierająca dolny trójkąt macierzy $\boldsymbol{C}$, wektor $\boldsymbol{b}$ oraz dodatkową liczbę y, wyrażającą zmienność całkowitą przy obliczaniu zmienności resztowej lub przyrównaną do zera w przypadku obliczania T^2 lub D^2;

eps — mała liczba uzależniona od maszynowej dokładności (zero maszynowe);

last — liczba kontrolująca właściwą kolejność wywoływania *linres*; przy pierwszym wywołaniu *last* musi mieć nadaną wartość 0, po wykonaniu obliczeń otrzymuje wartość *q2*.

Wyniki:

ind — tablica wskazująca numery zmiennych, na podstawie których została obliczona wielkość z; $ind[i] = 1$, jeśli zmienna o numerze i została wprowadzona do wielkości z; $ind[i] = 0$, jeśli wiersz o numerze i okazał się (prawie) liniowo zależny od poprzednio wprowadzonych;

last — otrzymuje wartość *q2* (patrz dane);

linres — wartość $y - \boldsymbol{b}'\boldsymbol{C}^{-1}\boldsymbol{b}$.

Procedura *linres*, w porównaniu z czasem obliczeń wymaganym przy obliczaniu wyrażenia z za pomocą odwracania macierzy $\boldsymbol{C}$, działa przeciętnie dwa razy szybciej, umożliwia ponadto otrzymywanie wyników pośrednich dla mniejszej liczby zmiennych.