ZASTOSOWANIA MATEMATYKI
APPLICATIONES MATHEMATICAE
XIV, 2 (1974)

ALGORITHM 34

ANNA BARTKOWIAK (Wrocław)

USE OF ORTHOGONAL COMPONENTS IN TESTS FOR TREND

1. Procedure declaration. The between-class-variation of the observed variable y is subdivided into orthogonal components, one degree of freedom each, associated with linear, quadratic, cubic etc. trends of y. The components are extracted as long as they appear statistically significant. Expected values of y and coefficients of regression can also be computed, if required.

Data:

q — number of classes,
maxq — greatest number of components to be extracted,
$w[1:q]$ — weights of observations in different classes of x,
$x[1:q]$ — values of the independent (optionally continuous) variable x at which the dependent variable y is observed,
$y[1:q]$ — values of response variable y observed in different classes of x,
bcv — between-class-variation,
sey — residual variance,
df — degrees of freedom for *sey*,
alfa — significance level,
v — Boolean variable; if $v \equiv$ **true**, the residual variance is modified by the use of local procedure *correctms* (see Section 2, case B).

Results:

ifault — error indicator, determined by the global procedure *orthonw* (see description of procedure *orthonw* [1]).

Other results of procedure *poltrenddec* are obtained by procedures *printpoltd*, *printey* and *printcoef* with headings described below. The call of these procedures is conditioned by the Boolean variables *vdd*, *ve* and *vc*. The headings of the procedures mentioned above are the following:

a. **procedure** *printpoltd*(*j*, *ss*, *f*, *alfa*, *percent*, *sw*);
value *sw*;
integer *j*, *sw*;
real *ss*, *f*, *alfa*, *percent*;

The procedure is used as follows:
1. if *sw* = 0: printing the headings;
2. if *sw* = 1: printing
   j — actual degree of approximation,
   ss — sum of squares due to the regression term of the *j*-th degree,
   f — value of the *F*-Snedecor test function (test of significance for the coefficient b_j),
   alfa — probability of *f* ($P\{x > f\}$),
   percent — percent of the total between-class-variation due to the regression term of the *j*-th degree;
3. if *sw* = 2: printing summary results.

b. **procedure** *printey*(*i*, *eyi*);
value *i*, *eyi*;
integer *i*;
real *eyi*;
This procedure is called *q* times for each degree of approximation. This gives the expected values of *y* at the *i*-th value of *x* for *i* = 1, ..., *q*.

c. **procedure** *printcoef*(*j*, *a*);
value *j*; **integer** *j*; **array** *a*;
This procedure prints
   j — the actual degree of approximation,
   a[*0* : *j*] — the coefficients of polynomial regression.

Procedures *printpoltd*, *printey* and *printcoef* should be supplied by the user.

d. **real procedure** *Ftest*(*fc*, *df1*, *df2*, *maxn*);
value *f*, *df1*, *df2*, *maxn*;
real *fc*;
integer *df1*, *df2*, *maxn*;
Ftest gives the probability in the *F*-Snedecor test.
Data:
   fc — calculated value of the *F*-Snedecor test function,
   df1 — number of degrees of freedom for the numerator of the *F*-formula,
   df2 — number of degrees of freedom for the denominator of the *F*-formula,
   maxn — if *df1* or *df2* are greater than *maxn*, normal approximations are used.
The body of this procedure can be found in [2].

e. **procedure** *orthonw*(*p*, *n*, *w*, *x*, *pl*, *pa*, *ifault*);
value *p*, *n*;

integer $p, n, ifault$;
array w, x, pl, pa;

This procedure generates the values of orthogonal polynomials, orthogonal on a given set of independent variables. It can be found in [1].

2. Method used. Let us given a response variable y with values $y_1, y_2, \ldots, y_q$ observed at q (not necessarily different) levels of a variable x, i.e. $x_1, x_2, \ldots, x_q$, with weights $w_1, w_2, \ldots, w_q$.

Let us also given the between-class-variation

$$bcv = \sum_{i=1}^{q} w_i y_i^2 - \frac{\left(\sum_{i=1}^{q} w_i y_i\right)^2}{\sum_{i=1}^{q} w_i} \tag{1}$$

and an estimate of the residual variance *sey* with *dfsey* degrees of freedom. The residual variance *sey* can, in some cases, coincide with $bcv/(q-1)$. Therefore, we describe two cases.

Case A. The recorded values of y represent values averaged from other observations, and we get an estimate of the residual variance *sey* with *df* degrees of freedom, which is not based on the between-class-variation *bcv*;

Case B. The array y contains values not averaged and the estimate of *sey* is based on the between-class-variation *bcv*.

Further, we assume a relationship between the variables x and y which can be approximated by a polynomial regression of the form

$$y(x) = a_0 + a_1 x^1 + \ldots + a_k x^k.$$

The direct application of least-squares method leads to a system of equations which with the increase of k are known to become ill-conditioned. This can be avoided by introducing the regression

$$y(x) = b_0 + b_1 p_1(x) + \ldots + b_k p_k(x),$$

where $p_j(x)$ $(j = 0, 1, \ldots, k)$ are polynomials of the j-th degree. The last form of regression is more suitable for numerical calculations, especially for greater values of k. It provides straightforward orthogonal components of the variance

$$SS_{\text{trend}\,j} = \frac{\left[\sum_{i=1}^{q} w_i y_i p_j(x_i)\right]^2}{\sum_{i=1}^{q} w_i [p_j(x_i)]^2} \qquad (j = 1, 2, \ldots, maxq). \tag{2}$$

```
procedure poltrenddec(q,maxq,w,x,y,bcv,sey,dfsey,alfa,
 ifault,v,vdd,ve,vc,Ftest,printey,printpoltd,printcoef,
 orthonw);
 value q,maxq,sey,dfsey,bcv,alfa,v,vdd,ve,vc;
 integer maxq,q,dfsey,ifault;
 real bcv,alfa,sey;
 array w,x,y;
 Boolean v,vdd,ve,vc;
 procedure printey,printpoltd,printcoef,orthonw;
 real procedure Ftest;
 begin
  integer i,j,k,n,fault;
  real s,s1,t,c1,c2,c3,sum,z,zmk,coef;
  array pt[1:q+6],pl,ey[1:q],b,bb,a[0:maxq];
  procedure correctms(x,y,i,j);
   integer i,j;
   real x,y;
   begin
    integer l;
    l:=i-j;
    x:=if l>0 then (x×i-y)/l else 0.0;
    i:=l
   end correctms;
  procedure bcalc(l);
   value l;
   integer l;
   begin
    if l=0
     then
     begin
```

```
 b[0]:=bb[0]:=p1[1];
 for i:=0 sten 1 until maxq do
  a[i]:=0.0
end l=0
else
 if l=1
  then
  begin
   bb[0]:=-pt[q+5];
   bb[1]:=pt[q+4]
  end l=1
  else
  begin
   c1:=pt[q+6];
   c2:=pt[q+5];
   z:=b[0];
   c3:=pt[q+4];
   b[0]:=t:=bb[0];
   bb[0]:=-c2×t-c1×z;
   j:=l-2;
   for i:=1 sten 1 until j do
    begin
     z:=c3×t-c1×b[i];
     b[i]:=t:=bb[i];
     bb[i]:=z-c2×t
    end i;
   z:=t;
   b[l-1]:=t:=bb[l-1];
   bb[l-1]:=c3×z-c2×t;
   bb[l]:=c3×t
```

```
    end l>1;
  for i:=0 sten 1 until l do
   a[i]:=a[i]+coef×bb[i]
 end bcalc;
n:=fault:=0:
k:=q-1:
zmk:=bcv/k;
s:=s1:=0.0;
bcv:=1/bcv;
orthonw(n,q,w,x,pl,pt,fault);
if vdd
 then printpoltd(0,1.0,1.0,1.0,1.0,0);
if vc∨ve
 then
 begin
  coef:=0.0;
  for i:=1 sten 1 until q do
   coef:=coef+w[i]×y[i];
  coef:=coef×pl[1];
  bcalc(n);
  z:=a[0];
  if ve
   then
   for i:=1 sten 1 until q do
    begin
     ey[i]:=z;
     printey(i,z)
    end i;
  if vc
   then printcoef(n,a)
```

```
   end vc\/ve;
ettrend:
  n:=n+1;
  orthonw(n,q,w,x,pl,pt,fault);
  if fault=0Vfault=2
   then
   begin
    sum:=0.0;
    for i:=1 step 1 until q do
     sum:=sum+y[i]×pl[i]×w[i];
    coef:=sum;
    sum:=sum×sum;
    correctms(zmk,sum,k,1);
    if v
     then correctms(sey,sum,dfsey,1);
    z:=if sey>0.0 then sum/sey else 0.0;
    if z>999.99
     then
     begin
      if z>999999.99
       then maxq:=n;
      z:=999.99
     end;
    if z=0.0
     then maxq:=n;
    s:=s+sum;
    t:=100.0×sum×bcv;
    s1:=s1+t;
    if vdd
     then printpoltd(n,sum,z,Ftest(z,1,dfsey,80),t,1);
```

```
    if ve
     then
     for i:=1 sten 1 until q do
      begin
       ey[i]:=z:=ey[i]+coef*pl[i];
       printey(i,z)
      end ve;
    if vc
     then
     begin
      bcalc(n);
      printcoef(n,a)
     end vc;
    if s1>99.99Vn≥(q-1)Vn≥maxq
     then go to ktrend;
    if v
     then go to ettrend;
    if Ftest(zmk/sey,k,dfsey,80)<alfa
     then go to ettrend
   end fault=0Vfault=2;
ktrend:
  ifault:=fault;
  if vdd
   then printpoltd(n,s,z,z,s1,2)
 end poltrenddec
```

The values $p_j(x_i)$ are calculated by the use of procedure *orthonw* [1]. The trend components are extracted from the between-class-variation step by step for $p = 1, 2, \ldots$ Each component has one degree of freedom. The j-th component compared with the polynomial regression of degree $j-1$ represents the amount by which the variance *bcv* can be reduced while introducing an additional polynomial regression term of the j-th degree. In other words, the j-th component represents the increase in predictability that would accrue for the sample data using a j-th degree equation instead of an equation of degree $j-1$.

The test of significance, for the j-th component, is

$$F = \frac{SS_{\text{trend}\,j}}{sey}$$

with 1 degree of freedom for the numerator and with *df* degrees of freedom for the denominator.

In case A the value of *sey* remains constant for all j.

In case B we assume that

$$y = b_0 + b_1 p_1(x) + \ldots + b_{j-1} p_{j-1}(x) + b_j p_j(x)$$

and test the hypothesis $H : b_j = 0$ which is equivalent to the assumption

$$y = b_0 + b_1 p_1(x) + \ldots + b_{j-1} p_{j-1}(x).$$

The test of significance for the j-th component is

$$F = \frac{SS_{\text{trend}\,j}}{Sey}, \tag{3}$$

where

$$S^{ey} = \frac{\sum_{i=1}^{q} (y_i - \bar{y})^2 - \sum_{k=1}^{j-1} SS_{\text{trend}\,k}}{q-j}.$$

The component $j+1$ is extracted when the following conditions hold:

a. the sum of trend components already extracted is less than 99.9 % of the initial between-class-variation;

b. $j < q-1$;

c. $j < maxq$;

d. the residual

$$R = bcv - \sum_{i=1}^{j} SS_{\text{trend}\,i}$$

is statistically significant (in case A only);

e. the test statistic F described by formula (3) has a value less than 999 999.99.

Condition e is introduced in case B, where the between-class-variation is practically exhausted and the remainder *Sey* should have the value zero.

Dependent on the Boolean variable *vdd*, a call of procedure *printpoltd* is executed, giving

j — actual degree of a polynomial;

sum — square of the j-th trend component as given by formula (2);

z — value of F evaluated by the use of *Ftest*; to ensure the value z not to be too big, z is censored by

$$z = \begin{cases} z & \text{for } z \leqslant 999.99, \\ 999.99 & \text{for } z > 999.99; \end{cases}$$

pc — amount of the variance *bcv* associated with the polynomial relationship of the j-th degree — in percents of the total variance *bcv*;

sw — integer associated with the optional form of print-out.

If *ve* is **true**, expected values of the variable y for given values of x are calculated by the formula

$$Ey_i^{(j)} = Ey_i^{(j-1)} + b_j p_j(x_i) \quad (i = 1, \ldots, q).$$

The expected values are stored in the local array $Ey[1:q]$. The calculations for the j-th degree ($j = 0, 1, \ldots$) are followed by a call of procedure *printey*.

If *vc* is **true**, the regression coefficients of equation (1) are calculated. The calculations are carried out for each degree by procedure *printcoef*.

3. Certification. Let us given the responses y,

$$y_1 = -708,\ y_2 = -85,\ y_3 = 3,\ y_4 = 8,\ y_5 = 750,\ y_6 = 3107,\ y_7 = 942,$$

observed along the scale of x,

$$x_1 = -3,\ x_2 = -2,\ x_3 = 0,\ x_4 = 1,\ x_5 = 3,\ x_6 = 4,\ x_7 = 5,$$

with corresponding weights

$$w_1 = 2,\ w_2 = 3,\ w_3 = 3,\ w_4 = 1,\ w_5 = 2,\ w_6 = 2,\ w_7 = 1.$$

(It is easy to verify that the values y and x satisfy the relation $y = 3x^5 + 2x^2 + 3$.)

By the use of formula (1) we get *bcv* = 93 209 220.

By the use of *poltrenddec* with $v \equiv vdd \equiv ve \equiv vc \equiv$ **true** we get, on the Odra 1204 computer, the following results:

The total variation of y, i.e. *bcv*, is decomposed into five trend components. With the fifth degree the total between-class-variation is exhausted, so F_c (F calculated) is set equal to 0.

degree of polynomial	*SS*	F_c	$P(F > F_c)$	% (percent of total *SS*)
1	51 696 000	14.94	.0022	55.46
2	22 006 900	12.41	.0048	23.61
3	17 684 500	97.07	.0000	18.97
4	1 567 580	55.49	.0000	1.68
5	254 235	.00	1.0000	.27
total	93 209 215			100.00

Expected values of y

degree of polynomial	Ey_1	Ey_2	Ey_3	Ey_4	Ey_5	Ey_6	Ey_7
0	1106.3	1106.3	1106.3	1106.3	1106.3	1106.3	1106.3
1	−1443.4	−729.5	698.33	1412.3	2840.1	3554.0	4267.9
2	240.36	−564.65	−808.42	−247.16	2241.6	4169.1	6552.1
3	−1006.5	281.94	−87.718	−585.81	691.55	3627.0	8867.0
4	−660.38	−164.16	−123.13	−148.23	577.09	3314.6	9296.8
5	−708.00	−85.000	3.0000	8.0000	750.00	3107.0	9428.0

Coefficients of regression

degree of polynomial	b_0	b_1	b_2	b_3	b_4	b_5
0	1106.29					
1	698.332	713.918				
2	−808.416	333.543	227.712			
3	−87.7184	−587.004	−7.75071	96.6681		
4	123.128	−197.753	−131.011	44.8887	12.5225	
5	3.00000	.00000	2.00000	−0.00000	−0.00000	3.00000

4. Examples of application.

EXAMPLE 1. In an experiment scheduled as a two-way analysis of variance, the observed responses of y at different levels of factors A and B with $p = 2$ and $q = 3$ levels are the following:

TABLE 1. Values of the character y observed at different levels of factors A and B

A \ B	$B1$ $x_1 = 1$	$B2$ $x_2 = 2$	$B3$ $x_3 = 3$
$A1$	5	0	2
$A2$	3	2	1
total	8	2	3
mean	4	1	1.5
weights of mean values	2	2	2

Let different levels of factor B represent subsequent points on a continuous x-scale corresponding to $x_1 = 1$, $x_2 = 2$ and $x_3 = 3$. We wish to investigate the general relationship between the response variable y and different levels of factor B. The column means of table 1 will be the starting point for the analysis.

The between-class-variance, by the use of formula (1), is

$$bcv = 38.50 - 28.16 = 10.34$$

with 2 degrees of freedom.

The residual sum of squares is

$$SS_{\text{res}} = \sum_{i=1}^{p}\sum_{j=1}^{q}(y_{ij} - y_{..})^2 - \sum_{i=1}^{p} q(y_{i.} - y_{..})^2 - \sum_{j=1}^{q} p(y_{.j} - y_{..})^2$$

$$= 14.83 - 0.17 - 10.33 = 4.33$$

with $df = (p-1)(q-1) = 2$ degrees of freedom.

Hence the residual variance is

$$sey = SS_{\text{res}}/df = 2.17.$$

Since there are three different values of x, the observed mean column values can be approximated at least by a linear relationship. The call of *poltrenddec* with $maxq = 1$, $v \equiv$ **false**, $vdd \equiv$ **true**, $ve \equiv$ **true** and $vc \equiv$ **true** works for $j = 0$ and $j = 1$.

For $j = 0$, we get

$$p_0(x_i) = \frac{1}{\sqrt{\sum_{i=1}^{q} w_i}} = 0.408 \qquad (i = 1, 2, 3).$$

The expected values of y are $Ey_1^0 = Ey_2^0 = Ey_3^0 = 2.17$, and the coefficient of zero-order regression is $a_0 = y_{..} = 2.17$.

The next call, for $j = 1$, evaluates the first order polynomials:

$$p_1(x_1 = 1) = -0.5000,\ p_1(x_2 = 2) = 0.000,\ p_1(x_3 = 3) = 0.500.$$

The amount of variation, due to linear trend, is

$$SS_{\text{trend 1}} = \frac{(2\cdot 4(-0.5)+2\cdot 1\cdot 0+2\cdot 1.5\cdot 0.5)^2}{1} = 6.25,$$

i.e. 60.48 % of the total between-class-variation.

The F-criterion for the significance test is $Fc = 6.25/2.17 = 2.88$, and the probability $P(F > Fc)$, calculated by the use of *Ftest*, is $P = 0.2315$; hence the extracted trend component is statistically non-significant. Since *maxq* is attained, no further trend components can be extracted.

EXAMPLE 2. Let us given the following values of x and y:

x	0	1	2	0	1	2
y	0	1	2	2	3	4

Each value has the same weight ($w[i] = 1$, $i = 1, \ldots, 6$).

The total variation of y is

$$bcv = \sum_{i=1}^{q} (y_i - \bar{y})^2 = 10.0.$$

It can be subdivided into variation caused by regression (the straight line $y = x + 1$) and the residual (deviation from regression).

Entering procedure *poltrenddec* with $q = 6$, *maxq* $= 5$, *sey* $= bcv/5$, *df* $= 5$, *alfa* $= 0.05$, $v \equiv vdd \equiv$ **true** and $ve \equiv vc \equiv$ **false**, we get the following trend decomposition:

degree of polynomial	*SS*	*F*	*alfa*	%
1	4.0000	2.67	.1778	40.00
2	.0000	.00	1.0000	.00
total	4.0000			40.00

ifault $= 3$, total between-class-variation $= 10.000$.

Since there are only three different values of x, procedure *orthonw* works only for $q = 1, 2$. Entering the procedure with $q = 3$, the error indicator *ifault* is set equal to 3 and further calculations are not executed.

References

[1] A. Bartkowiak, *Construction of polynomial values orthogonal on a given set of a one-dimensional variable*, this fascicle, p. 327-333.

[2] J. Morris, *Algorithm 346, F-test probabilities*, Comm. ACM 12 (1969), p. 184-185; see also: *Funkcja F-test*, Algol procedures library for the Odra 1204 computer, part 4, Elwro, Wrocław 1971.

COMPUTING CENTRE
UNIVERSITY OF WROCŁAW
50-384 WROCŁAW

Received on 10. 3. 1973

ALGORYTM 34

ANNA BARTKOWIAK (Wrocław)

ANALIZA TRENDU METODĄ WIELOMIANÓW ORTOGONALNYCH

STRESZCZENIE

Procedura *poltrenddec* (*polynomial trend decomposition*) wydziela ze zmienności międzyklasowej składniki wyrażające trend stopnia 1°, 2° itd. Wydzielanie składników odbywa się dopóty, dopóki pozostała reszta zmienności międzyklasowej pozostaje istotna.

W zależności od zmiennych boolowskich *vc* i *ve*, procedura oblicza również współczynniki regresji wielomianowej oraz oczekiwane wartości badanej zmiennej, wynikające z odpowiedniego równania aproksymacyjnego.

Dane:

q — liczba klas,
maxq — największa liczba wydzielonych składników,
w[*1* : *q*] — wagi obserwacji w różnych klasach na *x*,
x[*1* : *q*] — wartości niezależnej zmiennej *x* (która może być ciągła), dla których dane są obserwacje zmiennej zależnej *y*,
y[*1* : *q*] — wartości zmiennej *y* zaobserwowane w różnych klasach na *x*,
bcv — zmienność międzyklasowa,
sey — wariancja resztowa,
df — liczba stopni swobody dla *sey*,
alfa — poziom istotności,
v — zmienna boolowska; gdy *v* ≡ **true**, wariancja resztowa jest modyfikowana za pomocą procedury lokalnej *correctms* (patrz paragraf 2, przypadek B).

Wyniki:

ifault — wskaźnik błędu wzięty z procedury globalnej *orthonw* (patrz [1]).

Pozostałe wyniki procedury *poltrenddec* otrzymuje się za pomocą procedur *printpoltd*, *printey* oraz *printcoef*.