Listy Biometryczne - Biometrical Letters
Vol. 33 (1996), No. 1, 33-43

Exploring interaction in two-way tables reconsidered

L.C.A. Corsten[1]

Retired from Department of Mathematics, Agricultural University, Wageningen, The Netherlands

SUMMARY

Reconsidering a clustering procedure by Corsten and Denis (1990) to obtain a parsimonious description of interaction in a two-way table the author introduces a different stopping rule based on the comparison of the mean square for negligible interactions with an external variance estimate. A more attractive idea of explaining interaction might be the data-analytic identification of exceptionally high deviations from additivity. They are discovered in a selection process of indicator vectors, in principle forward (and perhaps secondarily backward). Termination of the selection is based on the comparison of the current residual mean square with an external variance estimate (if absent with a substitute presented in the 1990 note). In a numerical example the latter proposal turns out to be more parsimonious and specific than the former.

KEY WORDS: interaction, two-way table, clustering, stopping rule, mean square for negligible interactions, individual extra-ordinary deviations, outliers, indicator vectors, forward selection, relative value of residuals.

1. Introduction

In a previous note by Corsten and Denis (1990) a clustering procedure was presented in order to obtain a parsimonious description of interaction in a two-way table. Such a table consists of I rows and J columns, rows and columns both corresponding to I and J levels of purely qualitative and unstructured levels of two factors, e.g. different genotypes of an agricultural crop and different locations, respectively. The purpose of that procedure was to find groups of rows and groups of columns such that interaction present was substantially explained by the interaction between those groups only. The simultaneous clustering procedure of rows and columns was an agglomerative hierarchical one in each step of which either two row classes or two column classes were merged depending on which pair had currently the smallest proximity measure

[1] Current address: Ritzemabosweg 20, 6703 AX Wageningen, The Netherlands

m. That measure is the mean square for interaction in the $2 \times k$ or $h \times 2$ submatrix of the data matrix corresponding to the pair of row or column classes at hand, h and k being the current number of row or column classes, respectively. Each m is equal to the sum of squares for interaction in such a subtable divided by the appropriate dimension $k-1$ or $h-1$. At step i the interaction sum of squares considered to be negligible equals $d_i m_i$. The total interaction sum of squares considered to be negligible or sacrificed at step i will be $S_i = d_1 m_1 + \ldots + d_i m_i$, and is hoped to grow with minimal speed in proportion to the total dimension $D_i = d_1 + \ldots + d_i$. Without a stopping rule the growth of S_i would end with the final contribution to the total sum of squares for interaction due to the proximity measure of the last two row and two column classes with dimension equal to 1. Stopping at the stage where there are H groups of rows and K groups of columns implies retaining all the $I+J-1$ parameters corresponding to an additive model, augmented with $(H-1)(K-1)$ parameters for interactions between the designated groups of genotypes and groups of locations.

In the following sections stopping rules will be reviewed as well as complemented with a new one, which will be applied to the numerical example of the previous note after the correction of an annoying error in it. Next, a different method of tracing interaction will be presented followed by its application to the same example.

2. Stopping rules

A stopping rule embedded in a simultaneous F-test procedure was formulated by: stop just before S_i will exceed a critical value $c(\alpha)$ defined as

$$c(\alpha)/(ns^2) = F(n, f, \alpha)$$

where n is the dimension of the total interaction space $(I-1)(J-1)$, s^2 is an external and thus independent estimator of residual variance distributed as $\sigma^2 \chi_f^2 / f$ obtainable e.g. from replicate measurements, and $F(n, f, \alpha)$ the upper α-point of the statistic $F(n, f)$ with the parameter n for its numerator and the parameter f for its denominator. Thus

$$c(\alpha) = ns^2 F(n, f, \alpha).$$

Since the choice of α is arbitrary and subjective, it was suggested instead to consider the successive critical levels in testing the absence of interaction, i.e.

$$P_i = Pr\{F(n, f) > S_i/(ns^2)\},$$

and as P_i is non-increasing in i, to stop just before P_i will be too small, or rather before a drastic drop will occur, or if this does not occur at all, then just before the largest decrease.

An essentially different stopping criterion follows from the consideration of successive mean squares for all interactions deemed to be negligible in the course of the selection process, i.e. S_i/D_i. This ratio will generally increase as i increases. When this mean square grows larger than the external variance estimate s^2, one has the indication that S_i is no more free from true interaction. On the one hand, the sum S_i is the cumulation of minimized interaction sums of squares, on the other hand, it can be considered as the square of the perpendicular from the vector of observations to the space spanned by the $(I+J-1)$-dimensional space of main effects and the space of interactions between the current number h of row groups and the current number k of column groups as they have been chosen by the clustering procedure, i.e. the square of the orthogonal projection of the observational vector to a space of dimension $(I-1)(J-1)-(h-1)(k-1)$. The ratio of this square to that dimension should not exceed the external variance estimate. Otherwise, this internal residual variance estimate would be too large due to contamination with non-negligible interaction. Hence the new stopping rule prescribes to stop just before S_i/D_i will exceed s^2.

3. Numerical example revised

As has been discovered by several readers of the previous note as well as by ourselves, the data matrix of the numerical example given as Table 1 on page 210 has not been completely presented according to our intentions. It was our purpose to rearrange genotypes and locations such that feasible dendrograms for row and column groupings would appear, and that groupings would occur only between adjacent rows or columns. Unfortunately, the whole column marked 3 appeared by inexplicable causes in a wrong position. The third column should have appeared in the utmost right position instead. In other words, the columns as they were printed should have had the superscripts 1,2,7,3,4,5,6 instead of 1,2,3,4,5,6,7, respectively. The correct table appears here as Table 1. Fortunately, there are no consequences from this error for the dendrograms in Figure 1 on page 210 of our note nor for the corresponding text printed on page 211. In particular, the groupings of genotypes and also of locations mentioned in the last two paragraphs of page 211 apply to the present revised Table 1.

Independently of this error, we now cast doubts on the preference, expressed in the last paragraph of page 211 of our previous note, for a stop on the basis of the successive values of P_i at $i = 20$. The largest decrease of P_i shown at the bottom of page 211 in Table 2, although not very spectacular in comparison to the other ones is ·170 occurring between $i = 16$ and $i = 17$, P_i decreasing from ·591 to ·421, i.e. passing the value ·5, the transition between large and small critical levels. So we now prefer to stop at $i = 16$ leading to $H = 6$ and $K = 5$, over a stop at $i = 20$ motivated by the arbitrary choice of α about equal to ·05 for a simultaneous test. The grouping

Table 1. Average yields (kg per are) of 20 genotypes of corn at 7 locations according to Denis and Vincourt (1982)

Genotype	Location						
	1	2	3	4	5	6	7
1	59.8	61.0	58.8	64.4	62.7*	53.4*	75.6
2	64.5	70.3	60.4	73.2	78.3	70.8	81.5
3	59.5	68.2	60.0	72.3	76.9	71.5	79.9
4	65.1	71.9	58.5	71.9	83.2	71.5	82.1
5	64.2	68.2	60.1	74.2	85.4	78.4	81.2
6	56.4	65.2	58.0	68.0	80.6	73.0	79.6
7	63.5	65.6	60.7	71.6	73.2	69.4	74.3
8	58.3	65.7	60.0	74.6	73.7	66.1	72.7
9	61.9	66.1	62.9	74.5	75.6	74.4	83.4
10	58.9	64.8	67.7*+	71.5	72.0	70.0	80.9
11	57.2	64.1	56.2	75.4	72.4	65.7	81.2
12	58.0	66.1	62.2	74.5	82.0	70.0	85.5
13	62.0	71.8	62.8	71.4	77.0	75.6	69.1*
14	51.6	62.5	55.7	59.6	71.8	67.3	53.7*
15	62.9	64.8	61.2	64.5	77.9	65.5	67.2*
16	60.2	63.2	60.1	74.9	86.0*+	71.7	73.1
17	55.4	63.3	58.3	73.8	76.9	65.0	66.9*
18	53.7*	68.1	64.1	76.7	90.2*+	72.5	71.5*
19	54.5	67.3	60.2	82.2*+	81.0	73.1	76.7
20	56.1	59.4	62.5	70.4	85.9*+	65.4	76.3

corresponding to this preference is indicated with horizontal and vertical separation lines in the present Table 1.

Application of the new stopping rule with the external variance estimate equal to 8.59 and relatively stable due to $f = 266$ shows that S_i/D_i is equal to 8.25 for $i = 16$, and to 8.84 for $i = 17$. So it is decided to stop at $i = 16$, a confirmation of the decision to stop just before the largest decrease of P_i. Both stopping criteria lead to a description of interaction by 20 independent parameters.

4. A different view on interaction

Although several users of the technique presented in our previous note are quite happy with the resulting parsimony of interaction description, some geneticists have interest in a more specific sort of interaction than that between groups of genotypes and groups of locations. They are eager to discover specific combinations of a genotype and a location which show exceptionally high deviations, positive or negative, from additivity of genotype and location effects. Such combinations may be very advantageous or desirable under particular conditions.

An approach to identify such combinations may run as follows. One tries to extend the regression model for additive row and column effects with a set of additional indicator vectors. Each indicator vector has 1 at one special row column combination, and zeroes elsewhere.

In order to select possible candidates of such indicator vectors of which there are in principle IJ, possibly leading to linear dependence, one may firstly have recourse to a stepwise selection procedure as for instance is provided by a SAS package for personal computers. This program forces $I+J-1$ regressors for additivity permanently into the model by an INCLUDE option. Further it makes repeatedly one step forward by choosing that indicator vector which has the largest F value, the ratio of the sum of squares explained by the additional regressor over the mean residual square after entering that regressor into the model. It should be noted that each step forward implies a large number of recalculations of the regressors for additive effects. During the process the program ascertains whether any of the previously entered regressors could rather be removed from the model on the basis of a t-test. For this double forward and backward process one may set a significance level to a two-sided t-test for any indicator vector to be removed, and another similar one for any indicator vector to be entered. Indicator vectors which are possibly linearly dependent on the set of current model vectors will be removed automatically.

Without a stopping rule such a selection process with rather liberal significance levels, e.g. ·100 for forward selection, and ·110 for backward removal, may continue until practically all linearly independent indicator vectors will have been used. In contrast to what might be hoped for, the residual mean square for the successive models will not stabilize, but will often continue to decrease.

In analogy with the stopping criterion in the previous procedure requiring that the increasing mean square for negligible interactions should not exceed the external variance estimate, it is now proposed to stop the preliminary selection process just before the decreasing residual mean square is going to drop below the external variance estimate; otherwise, the regression model would certainly be too extensive, and too many indicator vectors would be wrongly included into it.

Incidentally, it is noted that each inclusion of an indicator vector into a regres-

sion model is closely related to a test whether the corresponding observation is an outlier or an aberrant observation from the current model. Each inclusion makes the corresponding residual vanish.

5. Numerical example reconsidered

Application of the selection process to the numerical example in the present Table 1 without a stopping rule showed that no indicator vector had to be removed except one at a very late stage; but this one was re-entered immediately after its removal. After an artificial stop was set at step number 100, all 100 vectors selected had critical level smaller than ·100 at the inclusion step. Table 2 shows the first 20 steps of this preliminary selection process, turning out to be equivalent to a merely forward process. The external variance estimate 8.59 is smaller than the residual mean square 8.62 at step 13, and larger than that at step 14. So the preliminary selection stopped at step 13.

In order to check that the selection was not too strongly dependent on the order or the step number of the process it was examined which singleton, pair, triplet, quadruplet, up to 12-tuple from the 13 selected vectors was the best one as expressed by the squared multiple correlation coefficient between the observation vector and the regression model for additivity extended with the relevant subset. For each k-tuple ($k = 1, 2, \ldots, 12$) the first k selected vectors turned out to be the best, except for $k = 3$ where (1,2,4) was the best triple, and (1,2,3) was the second best. Hence, the investigation of this 13-tuple will be continued, in that particular order.

Next, in the regression model, including the 13 selected indicator regressors, joint regression provides the simultaneous estimates of the regression coefficients $\delta_1, \delta_2, \ldots, \delta_{13}$ as well as the corresponding critical levels by a two-sided t-test. The results concerning those regression coefficients are given in Table 3, and are supplemented with additive row effects 65.78; 71.29; 69.76; 72.03; 73.10; 68.69; 68.33; 67.30; 71.20; 68.20; 67.46; 71.19; 71.72; 63.03; 67.75; 68.56; 67.07; 72.47; 69.20; 66.38, for rows 1 up to 20, and column effects −9.37; −3.15; 8.93; 2.42; 8.16; 1.17; 9.70, for columns 1 up to 7, respectively. The residual mean square equals 8.622. The value of F with 13 and 266 degrees of freedom for jointly testing the nullity of $\delta_1, \delta_2, \ldots, \delta_{13}$ is 11.08, as can be deduced from Table 2.

The following observations concerning this table may be made. Note that the sum of row effect, column effect and δ_i if it does not vanish is equal to the observation; e.g. $y(18;1) = 53.7$ equals 72.47 −9.37 −9.40 indeed. A striking feature is the relatively high value (between ·759 and ·819) of the so-called tolerance of the indicator vectors, i.e. one minus the squared multiple correlation coefficient of such an indicator with the set of the other $37 = 25 + 12$ regressors, 25 being the sum of the dimensions of

Table 2. Stepwise selection of the first 20 indicator vectors extending the additive model

Step number	Row;column # of observation	Residual sum of squares	Critical level 2-sided t-test	Residual mean square
0		2108.41		18.50
1	14; 7	1883.15	0.0004	16.67
2	18; 5	1762.70	0.0066	15.74
3	1; 6	1648.69	0.0066	14.85
4	1; 5	1523.32	0.0033	13.85
5	20; 5	1431.29	0.0093	13.13
6	13; 7	1342.77	0.0088	12.43
7	19; 4	1255.82	0.0076	11.74
8	16; 5	1188.64	0.0160	11.21
9	15; 7	1122.00	0.0141	10.69
10	17; 7	1053.12	0.0104	10.13
11	18; 7	994.22	0.0152	9.65
12	18; 1	928.65	0.0085	9.10
13	10; 3	870.80	0.0110	8.62
14	14; 4	827.30	0.0239	8.27
15	15; 4	781.43	0.0178	7.89
16	19; 1	745.65	0.0325	7.61
17	20; 3	710.08	0.0304	7.32
18	16; 7	676.48	0.0309	7.05
19	11; 4	647.47	0.0418	6.82
20	8; 7	619.22	0.0411	6.59

row and column contrasts. This indicates that the deviations will be estimated with small correlations among each other. Since the square of the centred indicator vector equals unity minus the reciprocal of the number of observations the tolerance should be multiplied with that square for establishing the squared length of the perpendicular from the non-centred indicator to the space spanned by all the 38 remaining regressors. The reciprocal of the latter squared length is the variance factor of σ^2 or s^2 for the relevant deviation estimator, as it has been used for establishing the critical level of the relevant t-statistic. The critical levels are considerably smaller than those in Table 2, except the last one which is equal to that in Table 2 as was expected. From the size of the critical levels, which all would have been smaller than ·01 if one had used the external variance estimate, as well as from the preceding convincing F-value, we conclude that the 13 designated observations are exceptional for an additive

Table 3. Deviations from additivity estimated by linear regression with critical levels and indicator tolerances

Subscript	Row;column #	Deviation	Critical level	Tolerance
1	14; 7	-19.03	0.0001	0.809
2	18; 5	9.57	0.0056	0.759
3	1; 6	-13.55	0.0001	0.797
4	1; 5	-11.24	0.0010	0.789
5	20; 5	11.36	0.0008	0.812
6	13; 7	-12.32	0.0003	0.809
7	19; 4	10.57	0.0016	0.819
8	16; 5	9.28	0.0055	0.812
9	15; 7	-10.25	0.0023	0.809
10	17; 7	-9.87	0.0033	0.809
11	18; 7	-10.67	0.0022	0.756
12	18; 1	-9.40	0.0063	0.765
13	*10; 3*	*8.43*	*0.0110*	*0.819*

model. These exceptional observations are marked with a star in Table 1. The five observations with a positive deviation are marked with an additional + sign.

It is interesting to note that 8 of the 13 exceptional observations appeared in three groups, two of size three and one of size two in columns 5 and 7 within classes formed by the clustering process of rows and columns as shown in Table 1; moreover, within each of these three groups the signs of the aberrations agree. This phenomenon supports the results of the previous clustering. On the other hand, the fact that this new modelling of interaction requires 13 instead of 20 independent parameters may be a good reason for a preference of the latter, also since it localizes the interaction not vaguely, but explicitly in 8 cells of the classification generated by the previous clustering of rows and columns. An additional advantage of the latter method of tracing interaction is the fact that it does not require orthogonality of rows and columns.

It should be mentioned that not much can be learned about the latter type of interaction from looking at the residuals with respect to the additive model adjustment. A Q-Q-plot of the ordered absolute values of those residuals versus half-normal deviates shows clearly irregular behaviour only for the three or four largest values. The residuals and their size give no reliable indication as to the observations which are possibly extraordinary. Only the five absolutely largest residuals have the corresponding indicator subscripts in Tables 2 and 3 in the same order. From there onwards the appearance of an indicator vector becomes unpredictable. For instance,

the observation in cell (15; 1) with the fifth largest residual 6.12 was not a candidate for being an outlier even after 33 steps of the selection procedure, while the observation in cell (15; 7) with residual −6.02 was the ninth candidate with a final deviation estimate of −10.25. When looking at the course of the residual per observation during the adjustment of 14 successive models one observes sometimes a steady decrease in absolute value, sometimes a steady increase in absolute value, and many times hardly any change of importance. In particular, there was no change of sign at all, except at the obligatory appearance of a zero residual.

For completeness sake it should be recalled that no exact meaning should be attached to probability statements in the present selection process. It was our purpose to compare quantitatively the consequences of successive steps in this data-analytic procedure which led to a highly parsimonious and informative model for interaction with a residual variance estimate about equal to an external one. If such an external estimate would not be available it is recommended to use the procedure sketched in section 5 (page 213) of the previous note.

6. Why not otherwise, a postscript

Referees suggested that a backward selection process might perhaps give safer results, and would more closely resemble the procedure of our note consisting of the removal of negligible sums of squares of interaction components. But in an additive model extended with a full set of IJ indicators each of those indicators is equally eligible for the first removal, and, in addition, without any contribution to explained interaction sum of squares. A similar situation holds for all indicators with respect to observations in the same row or the same column as the first one, the most logical choice for reaching linear independence among the retained indicators. So we find ourselves in a train of arbitrary removals none of which is justified, and thus in contrast to common sense. These removals are irreparable after reaching linear independence of indicators left. Hence we preferred forward selection from sheer necessity.

The danger of strong dependence among estimated deviations which had also been warned against, is contradicted by our observations concerning tolerances in Table 3, and the consequences of their size. Moreover, the matrix $\mathbf{X}'\mathbf{X}$ of rank 38 concerning the additive model extended with the 13 selected indicator variables has a reasonable condition number, i.e. the ratio of the largest and smallest eigenvalue equal to 73.21.

Another suggestion to use simultaneous inference according to Bradu and Gabriel (1974) in the selection of highly interacting genotype-location combinations must be directed toward residuals with respect to the additive model, and not to interactions in smaller or larger two by two tables, since these were already used in our previous

note and in sections 1 and 2 of the present paper. But this author cannot get further than the rejection of the nullity of the first deviation (and largest residual) with respect to an additive model at critical Bonferroni level $\cdot 0004 \times 140 = \cdot 056$, to be read from Table 2. How to handle ensuing residuals in this vein is still obscure to him. Of course, the F-statistic for testing simultaneously the nullity of the 13 selected deviations is overwhelmingly large, but how to account for the fact that selection took place is unclear. It is hoped however that a reasonable data analytic approach has been reached.

Finally, the new approach by Peña and Yohai (1995) consisting of finding relatively large components in the eigenvectors of a matrix $\mathbf{M}$ proportional to $\mathbf{EDHDE}$ where $\mathbf{H}$ is the (rank 26) hat matrix $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'$ of the additive model, $\mathbf{E}$ is the diagonal matrix of residuals with respect to that model and $\mathbf{D}$ is diagonal with elements $(1-h_{ii})^{-1}$, seems to indicate that our first four selected observations are outlying as well. That no further exceptional observations have been discovered is perhaps due to the fact that residuals with respect to additivity form the only data dependent input of this procedure, and not the consequences of the necessary adaptations of the model.

Acknowledgement

The author is most grateful to Mr. Albert Otten of the Department of Mathematics of the Wageningen Agricultural University for his effective computational assistance. He is greatly indebted for the positive comments of Dr. J.B. Denis of INRA, Versailles, France as well as Professor T. Caliński, Poznań, Poland.

References

Bradu, D. and Gabriel, K.R. (1974), Simultaneous inference on interactions in two-way analysis of variance. *Journal American Statistical Association* **69**, 428-436.

Corsten, L.C.A. and Denis, J.B. (1990), Structuring interaction in two-way tables by clustering. *Biometrics* **46**, 207-215.

Peña, D. and Yohai, V.J. (1995), The detection of influential subsets in linear regression by using an influence matrix. *Journal Royal Statistical Society*, Series B **57**, 145-156 and 611.

Received 12 April 1996; revised 11 May 1996

Badanie interakcji w tablicach dwukierunkowych

STRESZCZENIE

Praca dotyczy metody skupiania służącej do otrzymywania oszczędnego, w sensie liczby użytych parametrów, opisu interakcji w tablicach dwukierunkowych, opisanej przez Corstena i Denisa (1990). Autor wprowadza regułę zatrzymania bazującą na porównaniu średniego kwadratu dla nieistotnych interakcji z niezależną oceną wariancji. Przedstawia też sposób wyjaśniania interakcji poprzez identyfikację szczególnie dużych odchyleń od addytywności. Odchylenia te są wykrywane w procesie selekcji w przód lub wstecz. Zamieszczony przykład numeryczny pokazuje, że obecna propozycja pozwala uzyskać bardziej oszczędny opis interakcji.

SŁOWA KLUCZOWE: interakcja, tablica dwukierunkowa, skupianie, reguła zatrzymania, średni kwadrat dla nieistotnych interakcji, odchylenia, obserwacje odstające, wektory wskaźnikowe, selekcja w przód, względna wartość reszt.