Annual of the University of Mining and Geology "St. Ivan Rilski"
vol.44-45, part III, Mechanization, electrification and automation in mines, Sofia, 2002, pp. 81-85

GENERATING ACCURATE PLANS OF THIRD-DEGREE MODELS WITH IMPROVED FACTOR OF MULTICOLINEARITY

Diana Decheva

University of Mining and Geology
"St Ivan Rilski"
Sofia 1700 Bulgaria

Hristo Iontchev

University of Chemical Technology and Metallurgy
Sofia 1700, Bulgaria

ABSTRACT

Known plans of experiment for assessing third-degree models exhibit well expressed multicolinearity, i. e. a linear relationship between the columns of the information matrix. Their implementation impairs the processing of experimental data, and in many cases leads to obtaining biased estimates for the coefficients.

Algorithms and programmes that minimize two indirect criteria, namely the sum of extradiagonal elements of the covariance matrix, or the maximal one of those elements, are proposed for finding plans of a low factor of multicolinearity. Using these algorithms and programmes, plans of experiment for cubic regressions with two, three, or four factors have been generated. In all cases the value of the maximum variance inflation factor of experiment plans obtained has been improved up to two or three times.

Keywords: optimal planning of experiment, multicolinearity, cubic regression.

PROBLEMS OF OPTIMAL EXPERIMENT PLANNING FOR CUBIC REGRESSION

Besides a complete polynomial of second degree the polynomial models of third degree with *m* controlling factors also involve a term in one of the following forms or a combination of several such terms.

$$A = b_0 + \sum_{i=1}^{m} b_i x_i + \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} b_{ij} x_i x_j + \sum_{i=1}^{m} b_{ii} x_i^2 \qquad (1)$$

$$B = \sum_{i=1}^{m-2} \sum_{j=i+1}^{m-1} \sum_{l=j+1}^{m} b_{ijl} x_i x_j x_l \qquad (2)$$

$$C = \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} b_{ijj} x_i x_j^2 + \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} b_{iij} x_i^2 x_j \qquad (3)$$

$$D = \sum_{i=1}^{m} b_{iii} x_i^3 \qquad (4)$$

There exist many approaches to searching optimal plans for models of that type (Vuchkov, Krug et al., 1971; Vuchkov, Iontchev et al., 1978; Merzhanova and Nikitina, 1979; Iontchev, 1991). Criteria for *D*- or *G*-optimality are mostly used in those approaches, and the plans obtained have very good values of D_{eef} or G_{eef}. However, their applications are connected with problems emerging in the statistical processing of experimental data. This is due to the violation of one of the pre-requisites of using the method of least squares for assessing the coefficients in regression models, namely the requirement for lack of multicolinearity (i. e. no linear relationship between the columns of extended matrix *F* of the experiment plan should be present).

Several criteria calculated from the elements of matrix $\overset{o}{G} = \overset{o}{F}{}^T \overset{o}{F}$ have been proposed for the assessment of multicolinearity. A survey of those criteria has been compiled by Mitkov and Minkov (1993). $\overset{o}{F}$ designates the standardized matrix of the experiment plan, the elements of which are determined in accordance with the relationship:

$$\overset{o}{f}_{ji-1} = \frac{f_{ji} - \bar{f}_i}{\sqrt{\sum_{j=1}^{N} \left(f_{ji} - \bar{f}_i \right)^2}} \; , \; \mathrm{i} = 2, \mathrm{k} \, ; \; \mathrm{j} = 1, \mathrm{N} \qquad (5)$$

where *k* is the number of coefficients in the regression model being assessed, *N* the number of trials in the experiment plan, and $\bar{f}_i$ the arithmetic mean of the i^{th} column of *F*.

The most frequently used criterion for multicolinearity is the variance inflation factor or VIF. It is a vector consisting of the diagonal elements of standardized covariance matrix $\overset{o}{C} = \overset{o}{G}{}^{-1}$ It is assumed (Belsley, Kuh et al., 1980; Hocking, 1983) that a multicolinearity is present when the maximal element of VIF is greater than 3 to 5.

Table 1 shows data for the maximal VIF values of some known plans of experiment for models of complete third degree.

Table 1

Type of the experiment plan	Proposed by	m	N	Max VIF
1	2	3	4	5
Plan of 4 levels		2	16	12.69
Orthogonal non-compositional	Razdorskiy, Chaliy et al. (1973)	3	40	34.04
	Denisov and Popov (1976)	3	32	67.50
Discrete quasi-D-optimal	Vuchkov, Krug et al. (1971)	3	40	19.83
Non-saturated consecutive D-optimal	Vuchkov, Iontchev et al. (1978)	3	20	23.48

There is also an expressed multicolinearity in plans obtained for a non-complete third degree (Merzhanova and Nikitina, 1979; Iontchev and Stoianov, 1998).

Applying the method of least squares in the presence of multicolinearity leads to instability of coefficient estimates. In such a case it is recommended to process the experimental data by using the method of principal components, ridge regression analysis, regression analysis on characteristic roots, regression analysis with generalized reversal. A common disadvantage of the estimates obtained by these methods is their biasing which is the more considerable the stronger the mutual relationship between columns in matrix *F* is expressed.

Problems discussed above impose the development of algorithms and programmes for generating plans of experiment for cubic regression that have a low factor of multicolinearity. In such a way the needed pre-conditions will be created for finding more accurate estimates for coefficients in the equations being sought after.

INDIRECT CRITERIA FOR MULTICOLINEARITY

In the most general case, procedures of searching optimal experiment plans are reduced to improving iteratively a characteristic of an initial plan by consecutively eliminating and adding points to it.

For the assigned task of searching plans of a low multicolinearity factor it would be logical that the criterion of optimality be connected with minimization of the maximal VIF value. However, its direct application is limited by the large number of necessary computational operations. Every modification of the current plan (adding or eliminating a point) requires new standardization of *F*, forming and reversing matrix *G*. Although there are recurrent relationships for the first two operations, reversing the matrix implies considerable computational losses.

To synthesize algorithms of satisfactory speed of performance it is convenient to use parameters connected with the non-standardized covariance matrix *C*. Re-calculating its elements, when there is a change in the number of trials in the plan, can be easily realized by using the relationships proposed by Galil and Kiefer (1980). Substituting an indirect criterion for the basic one will be possible if only there is a correlation between them. To verify this hypothesis for various types of experiment plans for a sample of volume 10 the following has been found:

- the estimates for correlation coefficient $\hat{r}_1$ between maxVIF and the sum of absolute values of the extradiagonal elements in *C*;
- the estimates for correlation coefficient $\hat{r}_2$ between maxVIF and the extradiagonal element of maximal absolute value in *C*;
- the estimates for correlation coefficient $\hat{r}_3$ between maxVIF and the extradiagonal element of maximal absolute value in *C*;
- the calculated values of Student's *t*-criterion t_i , $i = 1,...,3$.

Results for four of the variants examined are shown in Table 2. For the first three of them the model is in the form $\hat{y} = A + D$, and for the last one $\hat{y} = A + B + D$.

Table 2.

Characteristics	Variant			
	1	2	3	4
1	2	3	4	5
m	2	3	3	4
k	8	13	13	23
N	12	13	19	28
$\hat{r}_1$	0.896	0.945	0.898	0.786
$\hat{r}_2$	0.988	0.996	0.991	0.996
$\hat{r}_3$	0.980	0.996	0.980	0.985
t_1	8.555	8.190	5.769	3.594
t_2	26.799	33.767	20.660	29.694
t_3	21.104	30.318	14.120	16.293

At a significance level of 0.05 the tabular value of the *t*-criterion is 2.306. In all cases examined it is smaller than the calculated one. This allows to assume the hypothesis for the presence of a linear relationship between the investigated variables, and to build up algorithms using indirect criteria for multicolinearity. Here, it should be taken into account that it is not possible to realize "free of charge" decreasing of the multicolinearity, and the plans obtained will have reduced parameter values for D_{eef} .

ALGORITHMS FOR GENERATING PLANS WITH A LOW FACTOR OF MULTICOLINEARITY

Two indirect criteria for searching optimal plans ζ^* with a low factor of multicolinearity have been used: a minimum of the sum of the absolute values of extradiagonal elements in matrix *C*, and a minimum of the extradiagonal element of highest module value in the same matrix.

$$\sum_{i=1}^{k-1}\sum_{j=1}^{k-1}\left|c_{ij}\left(\zeta^*\right)\right|=\min_{\zeta}\sum_{i=1}^{k-1}\sum_{j=1}^{k-1}\left|c_{ij}\left(\zeta\right)\right| \quad (6)$$

$$\max_{i,j}\left|c_{ij}\left(\zeta^*\right)\right|=\min_{\zeta}\max_{i,j}\left|c_{ij}\left(\zeta\right)\right| \quad (7)$$

They have been selected based on the following considerations:

- finding a minimal value 0 for the first criterion leads to generating an orthogonal plan, which means non-correlation of the estimates for regression coefficients and easy processing of experimental data;
- the coefficient of correlation between the second criterion and the maximal VIF is of the highest value.

Algorithm MMC1

From a random initial plan of *N*-1 points that point shall be eliminated, which minimizes the criterion selected. A random point from the points of number *L* forming the set of candidates is added to the *N*-point plan obtained. If this leads to a decrease in the criterion value it is assumed that a point has been successfully replaced. This procedure is continued until unsuccessful replacements of number *L* are carried out.

Algorithm MMC2

It applies a criterion of optimality (6) and in a generalized form realizes the following sequence of operations:

1. An initial plan of *N*+1 points is generated, the points being randomly selected from the set of candidates.
2. Eliminating consecutively one point at a time from the initial plan leads to obtaining *N*+1 plans, each of them consisting of trials of number *N*.
3. Sums S_i, i = 1, 2, ..., *N*+1, of the absolute values of extradiagonal elements in the covariance matrices are computed for all plans obtained at step 2.
4. A k^{th} plan for which $S_k \le S_i$, i = 1, 2, ..., *N*+1 is defined. It is assumed to be the best one for the time being, and the value of S_k is assigned to the minimal sum of extradiagonal elements S_{min}.
5. A check for depleting the set of candidates is performed. If consecutive unsuccessful attempts for adding all candidate points have been made, then the best plan obtained so far is assumed to be the one that has been sought after. Otherwise, the programme goes to step 6.
6. A random point from the set of candidates is added to the best plan found so far, and a plan of *N*+1 points is obtained.
7. Eliminating consecutively one point at a time from the plan formed at step 6 leads to obtaining *N*+1 plans, each of them consisting of trials of number *N*. The sum S_i, i = 1, 2, ..., *N*+1, of the absolute values of extradiagonal elements in the covariance matrix is computed for each of those plans.
8. A p^{th} plan, for which $S_p \le S_i$, i = 1, 2, ..., *N*+1 is defined.
9. If $S_p < S_{min}$, then the p^{th} plan is assumed to be the best at that moment of the search procedure, the value of S_p is assigned to S_{min}, and the algorithm continues performing from step 6. If $S_p \ge S_{min}$, then the programme goes to step 5.

Algorithm MMC3 has a structure analogous to that of MMC2 but the optimality criterion it uses is (7).

The algorithms shown have been realized as programmes in FORTRAN 77. To increase the speed of performance the programming solutions are based on:

- using only the supradiagonal elements of the covariance matrix for the matrix is a symmetrical one, and
- applying a recurrent computation of the effectiveness criterion value.

ANALYSIS OF RESULTS

Accurate experiment plans for the cases shown in Table 3 have been searched for by using the algorithms represented for the two selected criteria.

Table 3.

№	Type of model	m	k	N	L
1	2	3	4	5	6
1	$\hat{y}=A+D$	2	8	12	441
2		3	13	13	9261
3		3	13	19	9261
4		4	19	24	6561
5	$\hat{y}=A+B+D$	3	14	15	9261
6		4	23	28	6561
7	$\hat{y}=A+C$	3	16	21	9261
8	$\hat{y}=A+B+C+D$	3	20	25	9261

The characteristics of plans generated with procedure FDOP proposed by Iontchev (1991) have been used as a basis for comparing the results obtained.

In algorithm MMC1 the decision for replacing a point is made from the value of characteristics calculated for a plan of N+1 points. It turns out this being an essential problem in searching as the minimal value of the criterion obtained on the basis of N+1 points does nor guarantee a minimum for the criterion obtained from a plan of N points. For the algorithm considered the number of replaced points is relatively small, a tendency towards involving points symmetric to those existing in the current plan is observed, and the resulting plan depends to a considerable degree upon the initial one. For these reasons, the use of algorithm MMC1 is inefficient, irrespective of the fact that it finds plans of reduced multicolinearity.

Data for the maximal VIF of the best plans obtained through procedures FDOP, MMC2, and MMC3 are given in Table 4.

Table 4.

№ of plan	MaxVIF for plans obtained through:		
	FDOP	MMC2	MMC3
1	2	3	4
1	11.1954	5.3504	4.7934
2	15.8430	8.0715	7.2006
3	15.1251	5.8280	6.6307
4	16.5743	7.8128	7.7146
5	22.1535	9.0887	6.1818
6	27.7123	9.9930	5.9266
7	10.0000	4.7857	4.2968
8	28.3218	15.8197	11.6665

Fig. 1 shows the averaged parameters for maxVIF. Data from 10 successive realizations have been used for each of the eight required experiment plans. Plots corresponding to results obtained through procedures FDOP, MMC2, and MMC3 are designated by 1, 2, and 3, respectively.

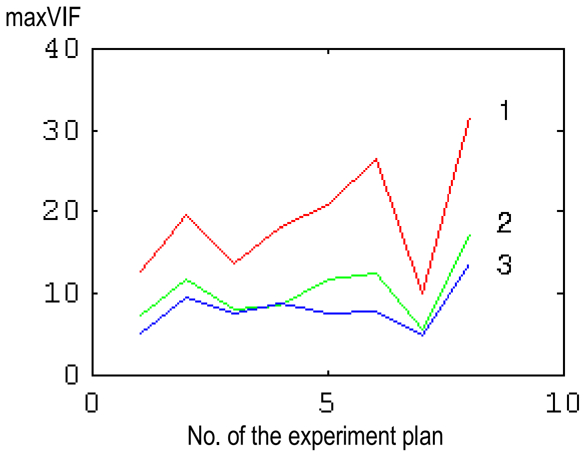

Figure 1. Variation of the mean value of the maximal variance inflation factor.

The number of iterations depends on the initial random plan. That is why the average number of iterations can be used for comparing the speeds of convergence of the algorithms proposed. Data for experiment plan No. 1 are shown in Table 5.

Table 5.

Parameters	Algorithms	
	MMC2	MMC3
1	2	3
Average number of iterations	1520	1560
Maximal number of iterations	2475	2430
Minimal number of iterations	724	805
Average number of successful iterations	42	48

Moreover, there is no much difference between the numbers of iterations needed for generating the rest of the plans when using algorithms MMC2 or MMC3. This allows to conclude that they are characterized by practically comparable speeds of convergence.

A simulation programme has been created in the MATLAB environment for the purpose of checking the properties of experiment plans generated. This programme is characterized by the following:

1. For all points of the plan examined it determines multiple times the value at the output of a plant described by a model of complete third deegree in the presence of a standard white noise. The real values of the coefficients are given in column 2 of Table 6. The ratio between the standard deviations of noise and the useful signal is 7 percent.
2. Using the method of least squares it determines the estimates of coefficients in the regression equation.
3. It calculates the variance of regression coefficients.

Results obtained for plan No. 8 are shown in Table 6.

It is obvious that for plans generated by using MMC2 and MMC3 the maximal value of coefficient variance has been reduced more than twice compared to that of the respective D-optimal plan.

The following conclusion can be deduced from the analysis of results obtained:

- using procedures MMC2 and MMC3 leads to obtaining plans of maximum value for the variance inflation factor being 2 to 3 times lower than that of respective D-optimal plans, which determines a considerable decrease in the variance of coefficient estimates for the regression equation;
- using criterion (7) generally leads to finding plans with lower values of maxVIF.

Table 6.

Symbolic designations and real values of coefficients		Coefficient variance when using:		
		FDOP	MMC2	MMC3
1	2	3	4	5
b0	3.0	0.2868	0.2299	0.1680
b1	-2.0	0.8462	0.4314	1.0682
b2	4.0	3.3604	0.8103	0.5105
b3	6.0	1.2387	0.6983	0.5164
b12	-4.0	0.0809	0.3071	0.1047
b13	2.5	0.1053	0.2905	0.1391
b23	-3.7	0.1293	0.1322	0.1446
b123	9.0	0.1043	0.0619	0.4879
b11	-12.0	0.2012	0.3108	0.2235
b22	-4.0	0.0724	0.2730	0.0892
b33	-3.0	0.1825	0.2694	0.4403
b111	-1.5	0.4016	0.3748	1.0697
b222	2.0	3.3216	1.3945	0.4629
b333	6.0	1.4865	0.3574	0.5643
b112	4.0	0.8494	0.4110	0.9025
b113	-3.0	0.5768	0.5376	0.9957
b221	6.5	0.7121	0.6688	0.3853
b223	4.4	0.2715	0.9387	0.3070
b331	-2.6	0.3759	1.3532	0.9951
b332	3.3	0.2184	1.0997	0.4293

At the same time it should be remembered that reducing the multicolinearity leads to lower D-effectiveness of the plans as well. Fig. 2 shows the averaged values of D_{eef} for the eight experiment plans obtained through MMC2 (plot 1) or MMC3 (plot 2).

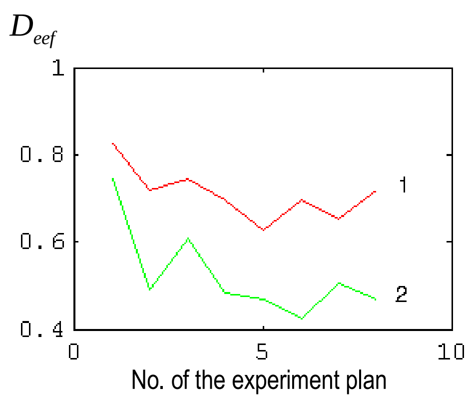

Figure 2. Variation of the mean value of D_{eef}.

Investigating the algorithms MMC2 and MMC3 and their respective programming implementations has shown that they can be used for generating plans for cubic regression, while the experimenter will have to choose the criterion to be used depending on the requirements for proximity of the plan obtained to the D-optimal one.

REFERENCES

Vuchkov, I. N., H. A. Iontchev et al. 1978. - *Catalogue of Consecutively Generated Plans*. Ministry of Education Publishing House, Sofia, . (In Bulgarian.)

Vuchkov, I. N., G. K. Krug et al. 1971. - *D-Optimal Plans for Cubic Regression*. Zavodskaya Laboratoriya, No. 7, . (In Russian.).

Iontchev, H. A. 1991. *Methods for Designing Optimal Compositional and Multi-Response Consecutive Experiment Plans*. Doctor of Technical Sciences Dissertation Thesis. STU, Sofia, . (In Bulgarian.).

Mitkov, A, D. Minkov. 1993. - *Statistical Methods for Investigating and Optimizing Agricultural Equipment.* Zemizdat State Co., Sofia, . (In Bulgarian.).

Denisov, V. I., A. A. Popov. 1976. - *A-E-Optimal and Orthogonal Plans of Regression Experiments for Polynomial Models.* Scientific Board in Cybernetics, Moscow, (In Russian.).

Merzhanova, R. F., E. P. Nikitina. 1979. - *Catalogue of Third-Degree Plans*. Published by Moscow State University, Moscow, . (In Russian.).

Razdorskiy, V. V., V. D. Chaliy et al. 1973. - *Obtaining Plant Model in the Form of Complete Third-Degree Polynomial.* In Engineering and Mathematical Methods in Physics and Cybernetics, Vol. 3, Atomizdat, Moscow, . (In Russian.)

Belsley, D. A., E. Kuh, R. E. Welsch. 1980. Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. N.Y., John Wiley.

Galil Z., Kiefer, J., 1980. Time and Space-saving Computer Methods, Related to Mitchell's DETMAX, for Finding D-optimum Designs, Technometrics, v 22.

Iontchev, H. A., K. Stoianov. 1998. Catalogue of response surface designs. University of Chemical Technology and Metallurgy, Sofia, .

Hocking, R. R. 1983. Developments in Linear Regression Methodology, Technometrics, v. 25, Nr. 3.

Recommended for publication by Department of Mine Automation, Faculty of Mining Electromechanics