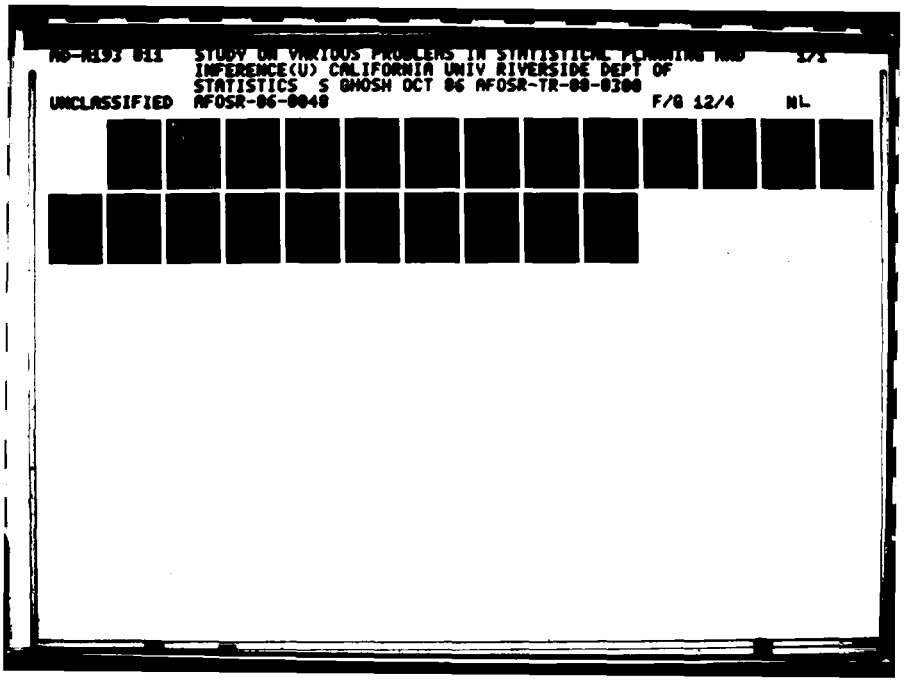
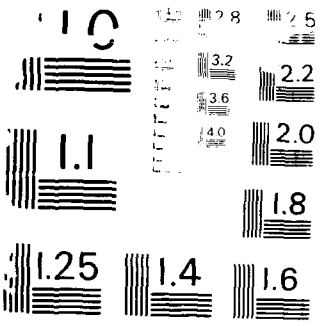

AD-A193 811 STUDY ON VARIOUS PROBLEMS IN STATISTICAL PLANNING AND 1/1
INFERENCE(U) CALIFORNIA UNIV RIVERSIDE DEPT OF
STATISTICS S GHOSH OCT 86 AFOSR-TR-88-0300
UNCLASSIFIED AFOSR-86-0048 F/G 12/4 NL



TEST CHART

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY: UNCLASSIFIED	3. DISTRIBUTION/AVAILABILITY OF REPORT: Approved for public release, distribution unlimited
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S)	5. MONITORING ORGANIZATION REPORT NUMBER(S): AFOSR-TR- 88-[illegible]
6a. NAME OF PERFORMING ORGANIZATION: University of Califor	6b. OFFICE SYMBOL (If applica): SD
6c. ADDRESS (City, State and ZIP Code): Department of Statistics, University of California, Riverside, CA 92521	7a. NAME OF MONITORING ORGANIZATION: AFOSR
	7b. ADDRESS (City, State and ZIP Code): BLDG #410, Bolling AFB, DC 20332-6448
8a. NAME OF FUNDING/SPONSORING ORGANIZATION: AFOSR	8b. OFFICE SYMBOL (If applicable): NM
8c. ADDRESS (City, State and ZIP Code): BLDG #410, Bolling AFB, DC 20332-6448	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER: AFOSR- 86-0048
	10. SOURCE OF FUNDING NOS.

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT NO
61102F	2304	A[illegible]	

11. TITLE (Include Security Classification): Study on various problems in statistical planning and inference

12. PERSONAL AUTHOR(S): Subir Ghosh

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Yr., Mo., Day)	15. PAGE COUNT
Final	FROM Dec 86 TO Nov 87	[illegible]	

16. SUPPLEMENTARY NOTATION

17. COSATI CODES		
FIELD	GROUP	SUB. GR

18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

Two new measures were proposed to indentify influential sets of oberservations at the design stage in view of prediction and fitting. Three measures of dispersion effects at different levels of factors in factorial experments were introduced. Research continues on characterring designs to enable one to measure and compare ~~displision~~ effects of levels of factors. Six papers were published and submitted for publication during this period .

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT	21. ABSTRACT SECURITY CLASSIFICATION
UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT	UNCLASSIFIED

22a. NAME OF RESPONSIBLE INDIVIDUAL	22b. TELEPHONE NUMBER (Include Area Code)	22c. OFFICE SYMBOL
Brian W. Woodruff, Maj	202-767-5028	

DD FORM 1473, 83 APR — EDITION OF 1 JAN 73 IS OBSOLETE.

UNCLASSIFIED

FINAL
[redacted] TECHNICAL REPORT AFOSR-TR- 88-

Project Title: Study on various problems in statistical planning and inference.

Principal Investigator: Subir Ghosh
Department of Statistics
University of California
Riverside, CA 92521

Program Manager: Major Brian Woodruff
Department of Mathematical and
Information Sciences
Air Force Office of Scientific Research
Bolling Air Force Base, DC 20332-6448

GRANT NO. AFOSR-86-0048

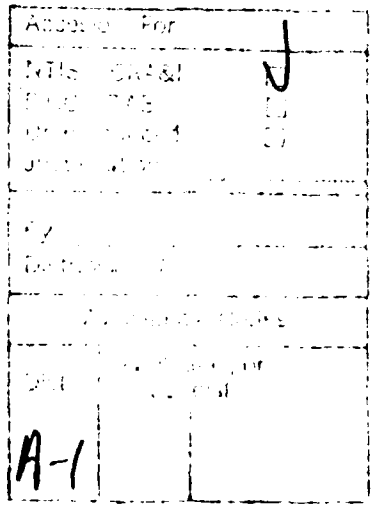


Table of Contents

0. Summary

The following researches are done under the Grant No. AFOSR-86-0048 during the year 1986-87.

(1) Two new measures are proposed to identify influential sets of observations at the design stage in view of prediction and fitting. A relationship is established between one of proposed measures and the Cook's measure at the inference stage.

(2) The problem of measuring dispersion effects at different levels of factors in factorial experiment is very important in quality control studies. Assuming that for the fitted model to the data there is no significant lack of fit, we proposed three measures of dispersion effects at levels 0 and 1 of factors in a 2^m factorial experiment. All three of them are relevant in replicated factorial experiments and two of them are applicable to unreplicated factorial experiments. We observe that the measures of dispersion effects, based on residuals obtained by the least squares fit of the model to the data, at levels 0 and 1 of a factor are correlated in most situations. We introduce a method of adjusting residuals and then propose measures based on residuals and adjusted residuals.

(3) This research is in progress. We characterize designs which enable us to measure and compare dispersion effects of levels of factors. We obtain some important results in this area.

1. Research Done

1.1 Two Methods of Identifying Influential Sets of Observations

The assessment of influence of a set of observations in the analysis of data is important not only at the inference stage but also at the planning (or design) stage. A set of observations under a design is said to be influential if the set affects not only the fitting of the model to the data but also the prediction in terms of the fitted model. In the problem of identifying sets of t (a positive integer) influential observations, we assume the underlying design is robust against the unavailability of any t observations [Ghosh (1979)]. To explain this concept we consider the standard linear model

$$E(\underline{y}) = X\underline{\beta}, \tag{1}$$

$$V(\underline{y}) = \sigma^2 I, \tag{2}$$

$$\text{Rank } X = p, \tag{3}$$

where $\underline{y}$ (Nx1) is a vector of observations, X(Nxp) is a known matrix, $\underline{\beta}$(px1) is a vector of fixed unknown parameters and σ^2 is a constant which may or may not be known. Let d be the underlying design corresponding to $\underline{y}$. The design d is assumed to be robust against the unavailability of any t observations in the sense that the parameters in $\underline{\beta}$ are still unbiasedly estimable when any t observations in $\underline{y}$ are unavailable. There are $\binom{N}{t}$ possible sets of t observations. The idea of robustness of designs against unavailability of data is fundamental in measuring the influence of a set of observations. We measure the influence of a set of t observations by assuming the observations in the set unavailable and then assessing the model fitted with the remaining (N-t) observations.

First Method

We propose the first measure in terms of precise prediction of t unavailable observations. A set of t observations is the most influential if the model fitted with the remaining (N-t) observations does the worst job in predicting t unavailable observations. We denote the ith set of t observations in $\underline{y}$ by $\underline{y}_2^{(i)}$; and the remaining observations in $\underline{y}$ by $\underline{y}_1^{(i)}$; the corresponding submatrices of X by $X_2^{(i)}$ and $X_1^{(i)}$; the resulting design when t observations in the ith set are unavailable by $d^{(i)}$, $i=1,\ldots,\binom{N}{t}$. The least squares estimators of $\underline{\beta}$ under d and $d^{(i)}$ are $\hat{\underline{\beta}}_d = (X'X)^{-1}X'\underline{y}$ and $\hat{\underline{\beta}}_{d(i)} = (X_1^{(i)'}X_1^{(i)})^{-1}X_1^{(i)'}\underline{y}_1^{(i)}$. We write the fitted values of $\underline{y}$ under d and $d^{(i)}$ as $\hat{\underline{y}}_d = X\hat{\underline{\beta}}_d$ and $\underline{y}_{d(i)} = X\hat{\underline{\beta}}_{d(i)}$. When t observations in the ith are unavailable, the predicted values of unavailable observations $\underline{y}_2^{(i)}$ from available observations are the elements in $\hat{\underline{y}}_2^{(i)} = X_2^{(i)}\hat{\underline{\beta}}_{d(i)}$. The reliability of these estimators can be judged by $V(\underline{y}_2^{(i)}) = \sigma^2 X_2^{(i)}(X_1^{(i)'}X_1^{(i)})^{-1}X_2^{(i)'}$. The first measure of influence of $\underline{y}_2^{(i)}$ is defined as

$$I_1(\underline{y}_2^{(i)}) = \text{Trace } V(\hat{\underline{y}}_2^{(i)}). \qquad (4)$$

The smallest value of $I_1(\underline{y}_2^{(i)})$, $i=1,\ldots,\binom{N}{t}$, for i=u, indicates that the uth set of t observations is the least influential in terms of precise prediction of unavailable observations. On the otherhand the largest value of $I_1(\underline{y}_2^{(i)})$, $i=1,\ldots,\binom{N}{t}$, for i=w, indicates the wth set of t observations is the most influential.

We denote the ith observation in $\underline{y}$ by y_i and the ith row in X by $\underline{x}_i'$, $i=1,\ldots,N$.

Theorem 1 For any design

$$\sigma^{-2} I_1(\underline{y}_2^{(i)}) \geq \sum_{i \in \{i_1,\ldots,i_t\}} \frac{\underline{x}_i (X'X)^{-1} \underline{x}_i}{1 - \underline{x}_i' (X'X)^{-1} \underline{x}_i} , \qquad (5)$$

where the $i_1,\ldots,i_t$ rows of X are rows of $X_2^{(i)}$.

Theorem 2 If for a design, the individual observations are equally influential then

$$I_1(y_i) = \frac{p\sigma^2}{(N-p)} . \qquad (6)$$

Theorem 3 If for a design, the individual observations are equally influential, then

$$I_1(\underline{y}_2^{(i)}) \geq \frac{p\sigma^2 t}{(N-p)} . \qquad (7)$$

From (6) and (7), we observe that for a design with equally influential individual observations $I_1(\underline{y}_2^{(i)}) \geq t\, I_1(y_i)$.

Second Method

The vector of least squares fitted values of N observations is uncorrelated with a complete set of orthonormal linear function of $\underline{y}$ with zero expectations. This fact implies the optimum property "Minimum Variance" of the vector of the least squares fitted value as an unbiased estimator of its expectation. Assuming a set of t observations unavailable, the least squares fitted values of the remaining (N-t) observations are correlated with a complete set of orthonormal linear functions of $\underline{y}$(Nx1) with zero expectation and therefore the covariance matrix is not a null matrix. Further the covariance matrix is away from the null matrix, the more influence has the set of t observations on the least squares fitted values of the remaining (N-t) observations.

Let Z ((N-p)xN) be a matrix such that Rank Z = (N-p), ZX = 0 and ZZ' = I. It can be seen that $Cov(\hat{\underline{y}}_d, Z\underline{y}) = 0$. This implies that $\hat{\underline{y}}_d$ has the minimum variance within the class of all unbiased estimators of $E(\hat{\underline{y}}_d)$ under (1-3). When t observations are unavailable, the least squares fitted values are $\hat{\underline{y}}_1^{(i)} = X_1^{(i)} \hat{\underline{\beta}}_{d(i)}$. We denote the sub-matrices of Z corresponding to $X_1^{(i)}$ and $X_2^{(i)}$ by $Z_1^{(i)}$ and $Z_2^{(i)}$. It follows that $Cov(\hat{\underline{y}}_1^{(i)}, Z\underline{y}) = \sigma^2[X^{(i)}(X_1^{(i)'}X_1^{(i)})^{-1}X_1^{(i)'}Z_1^{(i)'}]$. The further $Cov(\hat{\underline{y}}_1^{(i)}, Z\underline{y})$ is away from the null matrix, the more influential is the set of t observations $\underline{y}_2^{(i)}$. We thus have the second measure of influence as

$$I_2(\underline{y}_2^{(i)}) = \sigma^2 \text{ [Sum of Squares of elements in } Cov(\hat{\underline{y}}_1^{(i)}, Z\underline{y})]. \quad (8)$$

The largest value of $I_2(\underline{y}_2^{(i)})$, for i=w, indicates that the wth set of t observations is the most influential. The following results show some similarities between two measures of influence $I_1(\underline{y}_2^{(i)})$ and $I_2(\underline{y}_2^{(i)})$.

Theorem 4 The following is true for $i=1,\ldots,\binom{N}{t}$,

$$V(Z_1^{(i)}\hat{\underline{y}}_1^{(i)}) = V(Z_2^{(i)}\hat{\underline{y}}_2^{(i)}). \quad (9)$$

Theorem 5 The following is true.

$$I_2(\underline{y}_2^{(i)}) = \text{Trace } V(Z_2^{(i)}\hat{\underline{y}}_2^{(i)}). \quad (10)$$

The equations (4), (8) and (10) display the similarity between two measures of influence $I_1(\underline{y}_2^{(i)})$ and $I_2(\underline{y}_2^{(i)})$. Although the matrix Z is not unique, it can be checked that $I_2(\underline{y}_2^{(i)})$ is unique for all choices of the matrix Z.

Relationship with Cook's distance

Cook (1977) proposed a distance function between $\hat{\underline{y}}_d$ and $\hat{\underline{y}}_d(i)$, popular as Cook's distance, at the inference stage as

$$D_i = \frac{(\hat{\underline{y}}_{d(i)} - \hat{\underline{y}}_d)'(\hat{\underline{y}}_{d(i)} - \hat{\underline{y}}_d)}{p\, S_d^2}, \qquad (11)$$

where $(N-p)S_d^2 = (\underline{y}-\hat{\underline{y}}_d)'(\underline{y}-\hat{\underline{y}}_d)$. The Cook's distance D_i measures the degree of influence of t observations in the ith set on the estimation of $\underline{\beta}$. The following result shows that the first measure of influence $I_1(\underline{y}_2^{(i)})$ is in fact related to D_i.

Theorem 6 From (4) and (11), we have

$$E(p\, S_d^2\, D_i) = I_1(\underline{y}_2^{(i)}). \qquad (12)$$

Examples are presented in Ghosh (1987) to illustrate applicability of the two proposed methods.

References

Cook, R. D. (1977). Detection of influential observations in linear regression. Technometrics, 22, 495-508.

Ghosh, S. (1979). On robustness of designs against incomplete data, Sankhyā, B, Pts. 3 and 4, 204-208.

Ghosh, S. (1987). On two methods of identifying influential sets of observations. (Submitted to Statistics and Probability Letters.)

1.2 Comparing Dispersion Effects At Various Levels Of Factors In Factorial Experiments.

We consider a 2^m factorial experiment under a completely randomized design. Let T(n×m) be the design. The rows of T denote treatments and the columns denote factors. The design T is called an inner array for m controlled factors. For various level combinations of noise factors (outer array), we get replicated observations for every treatment in T (see Taguchi and Wu 1985). In the experiment, we take r ($\geq$ 1) observations for every treatment. The case r = 1 is called the unreplicated experiment and the case r > 1 is called the replicated experiment. Again, for simplicity equal replication is considered for the replicated experiment and the idea is easily extendable to unequal replications. Let y_{ij} be the jth observation for the ith treatment, $\bar{y}_i$ be the mean of all observations for the treatment i, i=1,...,n and j=1,...r, and (N = nr) be the total number of observations. The standard linear model for the experiment is

$$E(\underline{y}) = X\underline{\beta}, \tag{1}$$

$$V(\underline{y}) = \sigma^2 I, \tag{2}$$

$$\text{Rank } X = p, \tag{3}$$

where $\underline{y}$(N×1) is the vector of observations, $\underline{\beta}$(p×1) is the vector of factorial effects considered in the experiment, X(N×p) is a known matrix that depends on the design T and σ^2 is an unknown constant. We denote $H = X(X'X)^{-1}X'$ and $R = (I-H)$. The vectors $H\underline{y}$ and $R\underline{y}$ are the vector of least squares fitted values and the vector of residuals, respectively. The fitted values for all observations corresponding to the ith treatment are identical and is denoted by $\hat{y}_i$, i=1,...,n. Suppose that for the

fitted model to the data there is no significant lack of fit. The sum of squares of error is $SSE = \sum_{i=1}^{n} \sum_{j=1}^{r} (y_{ij} - \hat{y}_i)^2$, the mean square of error is $MSE = (SSE/(N-p))$, the sum of squares of pure error is $SSPE = \sum_{i=1}^{n} \sum_{j=1}^{r} (y_{ij} - \bar{y}_i)^2$ and the mean square of pure error is $MSPE = (SSPE/n(r-1))$. Note that both MSE and MSPE are measures of error variance σ^2. We now take MSE and MSPE as descriptive measures of noise. We then express MSPE as the weighted average of $(MSPE)_1$ and $(MSPE)_0$, where $(MSPE)_u$ is called the contribution of the level u (u = 0,1) of the factor to MSPE. We do the same for MSE. Different levels of a factor may contribute differently to MSE and MSPE. In general the contributions of levels of a factor to noise (measured by MSPE or MSE) are called the dispersion effects of levels of the factor.

We take a single factor out of m factors and develop the methods of measuring dispersion effects at the level 0 and 1 of the chosen factor. We do not introduce any notation for the chosen factor. This is to keep our presentation neat and clean. We define for i=1,...,n,

$$\delta_i = \begin{cases} 1 & \text{if the level of the factor in the ith treatment is 1,} \\ 0 & \text{if the level of the factor in the ith treatment is 0.} \end{cases}$$

Let D_1 (N×N) be a diagonal matrix with n sets of diagonal elements and the elements in the ith (i=1,...,n) set are equal to δ_i. We define $D_0 = I - D_1$. It can be seen that $D_1 D_0 = 0$ and both D_1 and D_0 are idempotent matrices. We have

$$SSE = \underline{y}' \, R \, D_1 \, R \, \underline{y} + \underline{y}' \, R \, D_0 \, R \, \underline{y}$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{r} \delta_i (y_{ij} - \hat{y}_i)^2 + \sum_{i=1}^{n} \sum_{j=1}^{r} (1-\delta_i)(y_{ij} - \hat{y}_i)^2,$$

$$SSPE = \sum_{i=1}^{n} \sum_{j=1}^{r} \delta_i (y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^{n} \sum_{j=1}^{r} (1-\delta_i)(y_{ij} - \bar{y}_i)^2.$$

The first set of measures of dispersion effects of levels of the factor are

$$S_1^2(1) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{r} \delta_i (y_{ij} - \bar{y}_i)^2}{\left(\sum_{i=1}^{n} \delta_i \right) (r-1)},$$

$$S_0^2(1) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{r} (1-\delta_i)(y_{ij} - \bar{y}_i)^2}{\left(\sum_{i=1}^{n} (1-\delta_i) \right) (r-1)}, \qquad (4)$$

at the levels 1 and 0, respectively.

We have

$$MSPE = \left(\frac{\sum_{i=1}^{n} \delta_i}{n} \right) S_1^2(1) + \left(\frac{\sum_{i=1}^{n} (1-\delta_i)}{n} \right) S_0^2(1).$$

Thus $S_1^2(1)$ and $S_0^2(1)$ are regarded as $(MSPE)_1$ and $(MSPE)_0$ in the notation of the previous section. If $S_1^2(1) > S_0^2(1)$, we then say that the level 0 of the factor has less contribution to MSPE and therefore would be preferred to the level 1 in view of stability against noise factors.

We denote Rank R D_1 R = V_1 and Rank R D_0 R = V_0. We now present the the second set of measures of dispersion effects of levels of the factor as

$$S_1^2(2) = (\underline{y}' \; R \; D_1 \; R \; \underline{y})/V_1 \; ,$$

$$S_0^2(2) = (\underline{y}' \; R \; D_0 \; R \; \underline{y})/V_0 \; , \qquad (5)$$

at the levels 1 and 0, respectively. We have

$$MSE = \left(\frac{V_1}{(N-p)}\right) S_1^2(2) + \left(\frac{V_0}{(N-p)}\right) S_0^2(2).$$

We now investigate the situation where $S_u^2(1) = S_u^2(2)$, u = 0,1. In other words, we like to characterize designs for which $\hat{y}_i = \bar{y}_i$. We denote the row of the matrix X corresponding to the treatment i by $\underline{x}_i'(1\times p)$. Note that for each i, i=1,...,n, the row $\underline{x}_i'$ is repeated r times in X. Let X*(n×p) be a matrix whose ith row is $\underline{x}_i'$. Notice that rows of X* are infact distinct rows of X. We have X'X = r(X*'X*).

Theorem 1. For i=1,...,n, $\hat{y}_i = \bar{y}_i$ if and only if $X^*(X^{*\prime}X^*)^{-1} X^{*\prime} = I_n$.

Corollary. For n = p, we have $S_u^2(1) = S_u^2(2)$, u = 0,1.

We thus observe that for designs with n = p, two sets of measures are identical. The class of designs with n = p includes the known Plackett and Burman designs (see Plackett and Burman 1947). We however strongly

feel that this class of designs is very weak in view of measuring dispersion effects, particularly for the condition that there is no significant lack of fit.

We denote

$$\underline{y} = \begin{pmatrix} \underline{y}_1 \\ \cdots \\ \underline{y}_0 \end{pmatrix}, \quad R = \begin{pmatrix} r_1 \\ \cdots \\ r_0 \end{pmatrix} = \begin{pmatrix} R_{11} & R_{10} \\ R_{01} & R_{11} \end{pmatrix}, \quad X = \begin{pmatrix} X_1 \\ \cdots \\ X_0 \end{pmatrix}, \quad \hat{\underline{y}} = \begin{pmatrix} \hat{\underline{y}}_1 \\ \cdots \\ \underline{y}_0 \end{pmatrix},$$

where $\underline{y}_u$ is the vector of all observations corresponding to treatments with $\delta_i = u$, $u = 0,1$ for the chosen factor. Two vectors of residuals $r_1 \underline{y}$ and $r_0 \underline{y}$ at the levels 1 and 0 of the factor are generally correlated under the model (1-3). We now present a vector of "adjusted residuals" at the level 0 of the factor, adjusted w.r.t. $r_1 \underline{y}$ so that it is uncorrelated with $r_1 \underline{y}$. We denote

$$r_1 = \begin{pmatrix} r_{11} \\ \cdots \\ r_{12} \end{pmatrix}, \quad R_{11} = \begin{pmatrix} R_{111} & R_{112} \\ R'_{112} & R_{113} \end{pmatrix}, \quad R_{10} = \begin{pmatrix} R_{101} \\ R_{102} \end{pmatrix},$$

where R_{111} $(V_1 \times V_1)$ with its rank V_1, r_{11} $(V_1 \times N)$ with its rank V_1. In fact we have $R_{111} = r_{11} r'_{11}$. We now write

$$r_{0a} = r_0 - R'_{101} R^{-1}_{111} r_{11}. \tag{6}$$

It can be seen that Rank $r_{0a} = [(N-p)-V_1] = V_{0a}$ (say) and furthermore, $\mathrm{Cov}(r_{11} \underline{y}, r_{0a} \underline{y}) = 0$. We call $r_{0a} \underline{y}$ the vector of "adjusted residuals" at the level 0 of the factor, adjusted w.r.t. the residuals at the level 1 of the factor. We denote

$$r_{0a} = \begin{pmatrix} r_{0a1} \\ r_{0a2} \end{pmatrix},$$

where $r_{0a1}\ (V_{0a}\times N)$ with its rank V_{0a}. We now have the sum of squares of the sets of linear functions $r_{11}\,\underline{y}$ and $r_{0a1}\,\underline{y}$ [see Scheffé 1959] as

$$SS(r_{11}\,\underline{y}) = \underline{y}'r_{11}\,[r_{11}r_{11}']^{-1}r_{11}\,\underline{y}\ ,$$

$$SS(r_{0a1}\,\underline{y}) = \underline{y}'r_{0a1}'\,[r_{0a1}r_{0a1}']^{-1}r_{0a1}\,\underline{y}\ , \qquad (7)$$

with d.f. V_1 and V_{0a}, respectively. We present the measures of dispersion and adjusted dispersion effects of levels of the factor

$$S_1^2(3) = [SS(r_{11}\,\underline{y})/V_1]\ ,$$

$$S_{0a}^2(3) = [SS(r_{0a1}\,\underline{y})/V_{0a}]\ , \qquad (8)$$

at the levels 1 and 0 (adjusted for level 1), respectively. We have

$$MSE = \left(\frac{V_1}{(N-p)}\right)\ S_1^2(3) + \left(\frac{V_{0a}}{(N-p)}\right)\ S_{0a}^2(3). \qquad (9)$$

Following the above approach we find the vector $r_{1a}\,\underline{y}$ of adjusted residuals at the level 1 of the factor adjusted w.r.t. $r_0\,\underline{y}$ so that it is uncorrelated with $r_0\,\underline{y}$. Let $r_{1a1}(V_{1a}\times N)$ be a submatrix of r_{1a} such that Rank r_{1a1} = Rank r_{1a} = V_{1a}, $r_{01}(V_0\times N)$ be a submatrix of r_0 with Rank r_{01} = Rank r_0 = V_0. We again present the measures of dispersion and adjusted dispersion effects of levels of the factor

$$S_{1a}^2(3) = [SS(r_{1a1}\,\underline{y})/V_{1a}]\ ,$$

$$S_0^2(3) = [SS(r_{01}\,\underline{y})/V_0]\ , \qquad (10)$$

at the levels 1 (adjusted for the level 0) and 0, respectively. We have

$$MSE = \left(\frac{V_{1a}}{(N-p)}\right)\ S_{1a}^2(3) + \left(\frac{V_0}{(N-p)}\right)\ S_0^2(3)\ ,\ (N-p) = V_{1a} + V_0 = V_1 + V_{0a}. \qquad (11)$$

Theorem 2. The following results are true.

1. $V_{1a} \geq \left(\sum_{i=1}^{n} \delta_i\right)(r-1)$, $V_{0a} \geq \left(\sum_{i=1}^{n} (1-\delta_i)\right)(r-1)$,

ii. $V_{1a}S^2_{1a}(3) \geq \left(\sum_{i=1}^{n} \delta_i\right)(r-1)S^2_1(1)$,

$V_{0a}S^2_{0a}(3) \geq \left(\sum_{i=1}^{n} (1-\delta_i)\right)(r-1)S^2_0(1)$,

iii. If $V_{1a} = \left(\sum_{i=1}^{n} \delta_i\right)(r-1)$ then $S^2_{1a}(3) = S^2_1(1)$,

iv. If $V_{0a} = \left(\sum_{i=1}^{n} (1-\delta_i)\right)(r-1)$ then $S^2_{0a}(3) = S^2_0(1)$.

We now study the measures in two extreme situations: (i) $r_1\,\underline{y}$ and $r_0\,\underline{y}$ are uncorrelated, i.e., $R_{10} = 0$, (ii) $r_1\,\underline{y}$ and $r_0\,\underline{y}$ are completely correlated, i.e., $r_{01} = D\,r_{11}$ for some matrix D.

Theorem 3. Consider the situation $R_{10} = 0$. Then $S^2_u(3) = S^2_u(2) = S^2_{ua}(3)$, u=0,1.

Theorem 4. If $r_{01} = D\,r_{11}$ then we have $r_{0a} = 0$, $V_{0a} = 0$ and $SS(r_{0a}\underline{y}) = 0$.

Theorem 3 tells that in case $R_{10} = 0$ there is no need for the adjustment of residuals. Theorem 4 tells that in case $r_0\,\underline{y}$ is linearly dependent on dependent on $r_1\,\underline{y}$ then the level 1 of the factor makes all contribution to SSE and the level 0 does not make any additional contribution to SSE. In case $V_{0a} = V_{1a} = 0$, we have $V_0 = V_1 = (N-p)$, $r_{01} = D\,r_{11}$ and D is nonsingular. This is a situation where the levels 0 and 1 have equal dispersion effects because of the design influence. It follows from Theorem 2 that for $r > 1$, V_0 and V_1 are both nonzero. (We assume naturally that there is at least one $\delta_i = 1$ and at least one $(1-\delta_i) = 1$.) For the case $r = 1$, at least one of V_{0a} and V_{1a} could be zero or both of them could be nonzero. We now consider the important situation when

both V_{0a} and V_{1a} are nonzero. If $S_1^2(3) >$ Maximum $\left(S_0^2(3), S_{0a}^2(3)\right)$, then the level 0 of the factor has an advantage edge over the level 1. On the other hand if $S_1^2(3) <$ Minimum$\left(S_0^2(3), S_{0a}^2(3)\right)$, the level 1 of the factor is superior to the level 0 in terms of smaller dispersion effects.

We now state some properties of the descriptive measures proposed in Section 3 under the model (1-3). We first observe that the measures $S_1^2(1)$ and $S_0^2(1)$ do not depend on the fitted model and all other measures depend on the fitted model. The measures $S_1^2(1)$ and $S_0^2(1)$ are always uncorrelated under the model (1-3). The measures $S_1^2(2)$ and $S_0^2(2)$ may however be correlated. Two sets of linear functions of observations $D_1 R \underline{y}$ and $D_0 R \underline{y}$ are uncorrelated if and only if $D_1 R D_0 = 0$. Therefore if $D_1 R D_0 = 0$ then $S_1^2(2)$ and $S_0^2(2)$ are uncorrelated. We have the following results:

Theorem 5. Suppose $\underline{y} \sim N(x\underline{\beta}, \sigma^2 I)$. A necessary and sufficient condition that

(1) $\dfrac{\underline{y}' R D_1 R \underline{y}}{\sigma^2} \sim$ Central X^2 with d.f. = Trace R_{11},

(2) $\dfrac{\underline{y}' R D_0 R \underline{y}}{\sigma^2} \sim$ Central X^2 with d.f. = Trace R_{00},

(3) and furthermore, (1) and (2) are statistically independent, is that $R_{10} = 0$

Notice that $D_1 R D_0 = 0$ if and only if $R_{10} = 0$. Moreover, $V_1 + V_0 = (N-p)$ if $R_{10} = 0$ and $V_1 + V_0$ could be greater than $(N-p)$ if $R_{10} \neq 0$. We question the use of estimators $S_1^2(2)$ and $S_0^2(2)$ for comparison unless $R_{10} = 0$. We of course realize that the condition $R_{10} = 0$ is too stringent to satisfy even for one out of m factors.

Theorem 6. The following results are true.

a. $\sum_{i=1}^{n}\sum_{j=1}^{r}\delta_i(y_{ij}-\hat{y}_i) = \sum_{i=1}^{n}\sum_{j=1}^{r}(1-\delta_i)(y_{ij}-\hat{y}_i) = 0,$

b. If for the factor $R_{10} = 0$, then

b.1. R_{11} and R_{00} are idempotent matrices,

b.2. $(\underline{y}_u-\hat{\underline{y}}_u) = R_{uu}\,\underline{y}_u$, $u = 0,1$,

b.3. $X_u' R_{uu} = 0$, $u = 0,1$,

b.4 $\sum_{i=1}^{n}\sum_{j=1}^{r}\delta_i\hat{y}_i(y_{ij}-\hat{y}_i) = \sum_{i=1}^{n}\sum_{j=1}^{r}(1-\delta_i)\hat{y}_i(y_{ij}-\hat{y}_i) = 0.$

The measures $S_u^2(3)$ and $S^2_{(1-u)a}(3)$, $u = 0,1$, are always uncorrelated.

The reason for adjusting residuals is to obtain uncorrelated dispersion effects.

References

Box, G. E. P. and Meyer, R. D. (1986). Dispersion effects from fractional designs. Technometrics, 28, 19-27.

Ghosh, S. (1986). Non-orthogonal designs for measuring dispersion effects in sequential factor screening experiments using search linear models (To appear in Communications in Statistics, Issue A12, No. 10, 1987.)

Taguchi, G. and Wu, Y. (1985). Introduction to Off-line quality control. Central Japan Quality Control Association, Tokyo.

2. Research in Progress

2.1 Characterizations of Designs In Measuring Dispersion Effects In Factorial Experiments.

In the problem of comparison of dispersion effects at levels 0 and 1 of a factor, we observe that certain designs are very good and the other designs are not so good. Among orthogonal designs there are designs which do not serve any purpose but there are orthogonal designs which are powerful. The balanced designs can be chosen to be efficient for this purpose. Some results for which I am proud of, are obtained. As mentioned before, this research is in progress, the details will come out in the near future technical report.

3. New Discoveries

Two new measures are proposed to identify influential sets of observations at the design stage in view of prediction and fitting. In the problem of measuring dispersion effects at different levels of factors in factorial experiment, a method of adjusting residuals is introduced and then measures are proposed based on residuals and adjusted residuals.

4. Publications

We present the list of published, accepted, and submitted papers under the Grant AFOSR-86-0048.

Ghosh, S. (1986). On a new graphical method of determining the connectedness in three dimensional designs. Sankhya, 48, Ser. B, Pt. 2, pp. 207-215.

Ghosh, S. (1987). Influential nonnegligible parameters under the search linear model. Commun. Statist. - Theory Meth., 16(4), 1013-1025.

Ghosh, S. and Zhang, X. D. (1986). Two new series of Search designs for 3^m factorial experiments. (To appear in November 1987 issue of Utilitas Mathematica.)

Ghosh, S. (1986). Non-orthogonal designs for measuring dispersion effects in sequential factor screening experiments using search linear models. (To appear in Communications in Statistics, Issue A12, No. 10, 1987.)

Ghosh, S. (1987). On two methods of identifying influential sets of observations (Submitted to Statistics and Probability Letters.)

Ghosh, S. and Lagergren, E. (1987). Comparing dispersion effects at various levels of factors in factorial experiments. (Submitted to Technometrics.)

5. Conferences Attended

The following is a list of conferences I attended during the year 1986-1987 (December 1986 - November 1987).

(1) Conference on the analysis of the unbalanced mixed model, April 6-10, 1987, University of Florida, Gainesville, Florida.

(2) Foundations and philosophy of probability and statistics, An International Symposium in Honor of I. J. Good, May 25-26, 1987. Eastern Region Institute of Mathematical Statistics meeting May 27-29. Virginia Tech., Blacksburg, VA.

I presented the contributed paper "Two methods of identifying influential sets of observations".

(3) Group Invariance Applications in Statistics NSF/CBMS Regional Conference, June 15-19, 1987, The University of Michigan, Ann Arbor.

(4) Joint Statistical Meetings American Statistical Association, Biometric Societies, Institute of Mathematical Society, August 17-20, 1987. San Francisco.

I presented the contributed paper (joint with X. D. Zhang) "Two new series of search designs for 3^m factorial experiments".

6. Interactions

Mr. Xiao Di Zhang, completed his Ph.D work under my direction in the Winter of 1987. The title of his thesis is "Search designs with applications to off-line quality control". Mr Eric S. Lagergren is working for his Ph.D. thesis under my direction in the area of measuring dispersion effects. Mr. Hamid Namini is working for his Ph.D. thesis under my direction in the area of robustness of designs against the unavailability of data. Ms. Jo Mahoney of UC, Irvine and Hughes Aircraft Company, is working for his Ph.D. thesis under my direction in the area of deleted factorial designs in incomplete blocks.

END

DATE

FILMED

DTIC

JULY 88