REGRESSION METHOD WITH DUMMY VARIABLES IN LINEAR MODELS FOR UNBALANCED DATA

By

RAM PRASAD GOSWAMI

Thesis submitted to the Haryana Agricultural University in partial fulfilment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

STATISTICS

College of Basic Sciences and Humanities

Hissar

1983

D E D I C A T E D

T O M Y

MOTHER SMT. DIWANO DEVI

IN MEMORY OF MY REVEREND FATHER

LATE SHRI LONG GIRI, WHO, BEING

A FARMER, ALWAYS INSPIRED ME FOR

HIGHER STUDIES.

CERTIFICATE - I

This is to certify that this thesis entitled, "Regression Method with Dummy Variables in Linear Models for Unbalanced Data" submitted for the degree of Doctor of Philosophy, in the subject of Statistics of the Haryana Agricultural University, is a bonafide research work carried out by Ram Prasad Goswami under my supervision and that no part of this thesis has been submitted for any other degree.

The assistance and help received during the course of investigation have been fully acknowledged.

29/X/83

(Dr.Umed Singh)
Major Advisor.

CERTIFICATE - II

This is to certify that the thesis entitled, "Regression Method with Dummy Variables in Linear Models for Unbalanced Data" submitted by Ram Prasad Goswami to the Haryana Agricultural University in partial fulfilment of the requirements for the degree of Doctor of Philosophy in the subject of Statistics has been approved by the Student's Advisory Committee after an oral examination on the same, in collaboration with an External Examiner.

10/2/84
Major Advisor

10-2-84
External Examiner

Head of the Department

Dean, Post-Graduate Studies.

ACKNOWLEDGEMENTS

I wish to express my deep sense of profound gratitude and affection to Dr.Umed Singh, Associate Professor of Statistics, Haryana Agricultural University, Hissar and Chairman of my advisory committee for his inspiring guidance, invaluable counsel and persistent encouragement throughout the investigation and preparation of this manuscript.

I am highly thankful to Dr.A.S.Arya, Associate Professor of Statistics; Prof.J.B.Chowdhury, Dean,College of Basic Sciences and Humanities; Dr.S.C.Chopra, Professor, Department of Animal Breeding; Dr.A.C.Gangwar, Head, Deptt. of Agricultural Economics; members of my advisory committee; for their benevolent help, valuable suggestions and constructive criticism. My sincere thanks are also due to Dr.P.D.Puri, Associate Professor of Statistics for getting conducted my preliminary examination (oral) and guided the work in the absence of Dr.Umed Singh when he was abroad on foreign assignment in U.S.A.

I am also thankful to Dr.I.J.Singh, Professor and Head, Deptt.of Agricultural Economics; Dr.F.S.Chaudhary, Dr.N.P.Singh, Dr.S.R.Chaudhary, Dr.S.S.Khirwar and Dr.R.N.Pandey,who always boosted my morale and spirit to finish the research work as early as possible.

continued...

Thanks are also due to Dr.O.P.Srivastava, Professor and Head, Deptt.of Statistics, Haryana Agricultural University, Hissar, for providing necessary research facilities in the department. I express my thanks to Sarvashri K.C.Goel, S.D. Batra and other staff for their help in programming and data analysis.

I shall be failing in my duty if I do not mention continuous help, kind affection and persistance insistance from Mrs.Sudha Dahiya. What I owe to her is too precious and I fail to find proper words at my command to express feelings.

Most sincere thanks are due to my colleagues, friends, both staff and students, in the Department of Animal Breeding, for their cooperation and help. I wish my thanks to Shri J.C. Malhotra for his secretarial help.

Last but not least I place on record my heartfelt thanks to my sons, Anil, Arun and Amit for creating the congenial atmosphere for working at home. I fail to find adequate words to express my affection and love to my wife Mrs.Shakuntala for patience and understanding.

(R. P. GOSWAMI)

Hissar
Dated: 29.X.83.

CONTENTS

CHAPTER - I

INTRODUCTION

In the analysis of linear models for designed experiments with balanced data (i.e. equal subclass numbers), there is a general agreement on the appropriate analysis of variance (ANOVA) table; or, stated differently there is a general concensus on the hypothesis tested under such headings as main effects and interaction. The analysis of variance (ANOVA) has been used most frequently in statistical methods for nearly half a century. Inspite of this popularity and great volume of literature available on this topic, one may find that still there is a disagreement on the appropriate analysis of unbalanced data, although all analysis are based on the statistically valid concepts of least squares. There are two basic reasons for the current controversy, first, the description of the linear model on which the analysis is based and second, the variety of computational methods which are used to perform the analysis.

Attempts at analyzing designs with unequal subclass numbers are generally based on the extensions of the methods for balanced data. In fact, the situation should have been opposite. Designs with unbalanced data have

their own analysis of variance techniques and those for balanced data are merely special cases of the techniques for unbalanced data. There are number of possible ways of generalizing the balanced analysis. Unfortunately, they do not, in general, lead to unique results. The article, by Francis (1973) compared the results of four computer programmes and discussed the rationale for making a choice. All programmes yielded different results. He concluded that programme BMDIOV (BMD X 64) (Dixon,1970), provided the most suitable analysis out of four computer programmes. This lack of agreement between the different analysis has led to much confusion and prompted a number of articles by various workers in the recent statistical literature.

Kutner (1974) related various sum of squares obtained by a number of computer programmes to the corresponding linear hypothesis being tested about the stated model.

The R( ) notation or reduction in sum of squares has been used by several workers as an aid in calculating the various entries in the analysis of variance table. While there is only one R( ), there are two different and distinct ways. First technique is to apply the R( ) to the non-full rank model (Searle, 1971, 1972). The

another technique used by Harvey (1968), Overall and Klett (1972) and Carlson and Timm (1974) is to apply the R( ) to the reparametrized full rank model. Two forms of linear models were used viz., the classified fixed effects β-model and cell means or μ-model. In general non-full rank, fixed effect β-model can be defined as:

$$\underline{Y} = \underline{X}\underline{\beta} + \underline{e} \qquad \ldots \qquad (1.1)$$

where $\underline{Y}$ is an nxl vector of observations, $\underline{e}$ is an nxl vector of errors which are assumed normally independently distributed with mean vector $\underline{o}$ and variance $\sigma^2\underline{I}_p$, $\underline{\beta}$ is a pxl vector of unknown parameters and $\underline{X}$ is a nxp design matrix of rank $q < p$. The elements of $\underline{X}$ are assumed to be known and are either zero or one. The analysis of such linear models including point and interval estimation and test of hypothesis is well known in case of balanced data. The analysis of β-model focuses on inferences about $\underline{\beta}$ with particular emphasis on test of linear hypothesis. The analysis of such linear models for unbalanced data, e.g., unequal cell numbers, n_{ij}, is not well understood since this model is overparametrized that is if $q < p$ then there are only q independent parameters. As a consequence, the individual parameters have no meaning. This overparametrization allows for the imposition of non-estimable condition on $\underline{\beta}$. In β-model, we review often confusing

concepts of estimable conditions, testable hypothesis, reparametrization and interpretation of the parameters.

The non-estimable conditions

$$\underline{K}\beta = \underline{o} \qquad . \quad . \quad . \qquad (1.2)$$

are imposed to solve the model (1.1). Here $\underline{K}$ has rank (p-q) and rows of $\underline{K}$ are not in the row space of $\underline{X}$. The choice of particular $\underline{K}$ gives meaning to the components of β and allows for the interpretation of the hypothesis about β . Unfortunately, the choice of $\underline{K}$ often makes the interpretation difficult with the result that there is much confusion as to the meaning of certain hypothesis.

Another basic model is cell means or μ-model which is an alternative form of the model (1.1). Expressing the model in matrix form analogous to (1.1), the model is

$$\underline{Y} = \underline{W}\,\underline{\mu} + \underline{e} \qquad . \quad . \quad . \qquad (1.3)$$

subject to the condition

$$\underline{G}\,\underline{\mu} = \underline{O} \qquad . \quad . \quad . \qquad (1.4)$$

where $\underline{Y}$ is an nxl vector of observations ($n = \sum_{i=1}^{p} n_i$) and $\underline{e}$ is vector of errors which is $\sim N(\underline{O}, \sigma^2 \underline{I}_n)$, $\underline{\mu}$ is a pxl vector of means μi, i = 1,2, ... p corresponding to the populations sampled. $\underline{W}$ = Diag ($\underline{\mathcal{J}ni}$), where $\underline{\mathcal{J}ni}$ is the vector of n_i ones indicating the number of observations on the i^{th} population (i^{th} cell) or $\underline{W}$ is a matrix of ones and

zeros such that each row of $\underline{W}$ has only one one and each column has as many ones as there are observations from the corresponding populations. The matrix $\underline{G}$ is a known qxp matrix of rank q and defines assumed linear relations among means. It is emphasized that actually (1.4) represents assumptions made about the means of the populations sampled as opposed to conditions imposed on the parameters in model (1.1) to achieve full rank. Here the concept of constrained and unconstrained model is introduced.

Consider the two way classification model

$$Y_{ijk} = \mu_{ij} + e_{ijk} \quad . \quad . \quad . \qquad (1.6)$$

where $i = 1,2, \ldots ,a$, $j = 1,2, \ldots ,b$

and $k = 1,2, \ldots ,n_{ij}$

The general condition is

$$\mu_{ij} - \mu_{ij'} - \mu_{i'j} + \mu_{i'j'} = 0 \quad . \quad . \quad . \qquad (1.7)$$

for all i, i', j and j', μ is a vector of μ_{ij} of length ab. The constraints in equation (1.7) are linearly independent and can be easily reduced to a set of (a-1) (b-1) linearly independent constraints.

The constrained and unconstrained models should be clearly spelled out. The constraints are assumed to be known linear relationship among the means based on the research workers knowledge of the experimental situation.

Suppose in a two way table that $\mu_{11} = \mu_{12} = \mu_{13}$. A common constraint is that there is no interaction. In this case when $n_{ij} = n$, the estimates of $\mu_{ij} = \bar{Y}_{i.} + \bar{Y}_{.j} - \bar{Y}_{..}$ as opposed to $\bar{Y}_{ij}$. It should be noted that such constraints are based on the prior knowledge rather than being a consequence of the data.

Most linear models can also be alternatively considered in third type of model, called regression model. This can be done by defining appropriate dummy variables in a regression model. A regression formulation often is desirable, if not mandatory when dealing with unbalanced data involving two or more factors. The general regression formulation corresponding to the over-parametrized model $Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}$ is

$$Y = \mu + \sum_{i=1}^{a-1} \alpha_i x_i + \sum_{j=1}^{b-1} \beta_j z_j + \sum_{i=1}^{a-1}\sum_{j=1}^{b-1} \gamma_{ij} x_i z_j + e \quad . \quad . \quad . \quad (1.8)$$

where μ, α_i, β_j and γ_{ij} are regression coefficients, and x_i and z_j are appropriately defined dummy variables. Here three coding schemes for the dummy variables in the general regression model (1.8) are defined.

Scheme A. Define

$$x_i = \begin{cases} 1 \text{ for level i of the row factor} \\ -1 \text{ for level a of the row factor} \\ 0 \text{ otherwise.} \end{cases}$$

and

$$z_j = \begin{cases} 1 \text{ for level } j \text{ of the column factor} \\ -1 \text{ for level } b \text{ of the column factor} \\ 0 \text{ otherwise.} \end{cases}$$

Scheme B. Define

$$x_i = \begin{cases} 1 \text{ for level } (i+1) \text{ of the row factor} \\ 0 \text{ otherwise} \end{cases}$$

and

$$z_j = \begin{cases} 1 \text{ for level } (j+1) \text{ of the column factor} \\ 0 \text{ otherwise} \end{cases}$$

Scheme C. Define

$$x_i = \begin{cases} n_{a.} \text{ for the level } i \text{ of the row factor} \\ -n_{i.} \text{ for the level } a \text{ of the row factor} \\ 0 \text{ otherwise} \end{cases}$$

and

$$z_j = \begin{cases} n_{.b} \text{ for level } j \text{ of the column factor} \\ -n_{.j} \text{ for level } b \text{ of the column factor} \\ 0 \text{ otherwise} \end{cases}$$

where $i = 1,2,\ldots,(a-1)$ and $j = 1,2,\ldots,(b-1)$.

The present study has been undertaken to investigate the analysis of unbalanced data through regression on dummy variables, calculating various reduction in sum

of squares and testing of various hypotheses associated with the sum of squares. It will be shown that various regression on dummy variable approach are not identical for the purpose of testing hypothesis and that one is not a special case of the others. In this study, these procedures will be sorted out with respect to where they agree and where they disagree and relate the associated sum of squares to the linear hypotheses being tested. A unified regression approach has been developed which yields analysis of variance techniques as a special case of the techniques for unbalanced data. This approach will remove many prevailing sources of confusions. The present investigation was undertaken to fulfil the following objectives:

i. to propose appropriate regression method of analysis of linear models for unbalanced data,

ii. to propose an appropriate index for the measure of extent of non-orthogonality present in the data, and

iii. to work out analysis procedure based on regression approach for unbalanced data including zero cell frequency.

CHAPTER - II

REVIEW OF LITERATURE

The main objective of this work is to investigate the procedure of analysis of unbalanced data through the regression on dummy variables which will remove many prevailing confusions in testing of a particular hypothesis. Great volume of research work has been carried out by several workers on the analysis of balanced data. But in the applied field of research, there are many situations where the cell frequency is not same or equal. Even some times the observations are subject to loss or damage and zero cell frequency is there. The statistical technique for analysing this type of data called unbalanced or non-orthogonal data is quite different from the technique of balanced data. The relevant literature was reviewed under the following heads:

1. Methods of estimation
2. Testing of hypothesis
3. Computer programming.

Yates (1934) described two methods of analysis of variance to estimate the effects of different factors; (i) method of fitting constants and (ii) method of weighted squares of means when there is only one fold classification.

He found that both the methods of estimation yielded identical results in case of balanced data.

Nair (1941) discussed only one method of analysis i.e. method of fitting constants in case of non-orthogonal data arranged into two fold classification. He estimated the main effects α_i and β_j on the assumption that two attributes exert their influence independently on each others. The problem of testing the significance between any two treatment effects has also been discussed along with their standard error.

Crump (1946) aim was to discuss the uses of analysis of variance and to estimate the various components of variance. The first use is that if an observed statistical variate i.e., the plot yield of varietal experiment, is assumed to be the sum of several separate effects, the variance of each effect will contribute to the total variance. The second use of analysis of variance provides estimates of these several variance components. This was the purpose to discuss the hypothesis appropriate to the two uses of analysis of variance and explain its uses to estimate variance components.

Smith (1951) described in this paper that analysis of variance table provided the solution to two or more

problems relating to the estimation as: (i) to detect and estimate the components of variance in a composite population and (ii) to detect and evaluate the significance of difference among means of subsets when cell frequency are unequal but proportionate to their marginal totals, still additive property of sum of squares holds.

Federer (1957) classified the variance covariance analysis in three categories, case I, when interaction was absent and analysis was being done as a method of fitting constants. Case II when interaction present and analysis was carried out assuming the fixed effects model using the technique of weighted squares of means and case III when interaction was present and the interaction effect and at least one of the effect, main effects of the factors represented in the interaction, assumed to be random effects. The statistical procedures for the three cases had been derived for two way and three way classification. The procedure for q-way classification with $\underline{b}$ covariate was indicated.

Raut (1960) defined a type of design called generalized nonorthogonal design. A method of intrablock analysis has been described together with involving a method for obtaining estimates of treatments using block totals. Most of the existing incomplete block designs comes out as a

particular case of this design.

The optimum properties of the popular estimates of the intra class variance in components of variance models are known if there are equal number of observations in the class. Read (1961) showed in its contrast that unbiased quadratic estimates having uniformly minimum variance do not exist if the design is not balanced. The sampling variance of the general unbiased quadratic estimate is given for a single classification assuming the components have normal distribution.

Federer (1963) studied to determine what information on the source of variation for a three way classification analysis of variance can be extracted from nested and two way classification analysis of variance. The case of unequal number of observations in the subclass was also considered. Although various analysis of variance produced variation and different sum of squares, it was not possible to extract all the information via, the analysis considered in this investigation. Thus if all the sum of squares for a higher way classification analysis of variance were desired, it would be necessary to perform the complete analysis. The purpose of this paper was to set forth the various quantities obtained from these analysis. Although the results were confined to a three way classi-

fication, extension to four way and higher way classification is straight forward.

Pearce (1963) investigated that nonorthogonal designs can usefully be classified to the pattern of matrix Ω^{-1} where Ω is the variance covariance matrix of the estimates of the treatment parameter. The pattern of Ω^{-1} determined the method of analysis and that of Ω , the potential uses of designs. Some broad classes were suggested which cover most block design and some other useful designs were considered in relation to the proposed classification. Attention was paid to row and column design also.

Mielke and Mchugh (1965) described that analysis of variance of the data in cross classifications with unequal subclass numbers, that are disproportionate, caused complexities which were not found in case of equal or proportionate subclass frequencies. A design, model and analysis of variance were presented appropriate to the estimation of treatment effects in this situation, as encumbered by two additional features, i.e. (i) a mixed effects model rather than a random or fixed effects model, (ii) a finite random effect population rather than finite. A numerical example was also taken to illustrate the

application of the theory.

Draper and Smith (1966) discussed the regression on dummy variables but at an introductory level. Federer and Zelen (1966) revealed that the sum of squares in the general unequal numbers analysis of variance for n-way or n factor-classification might be obtained from standard regression theory. It was felt that when number of factors were large, computing formula using normal equation became very cumbersome and difficult for solving. In this paper, the authors were able to set forth relatively simple computational procedures for obtaining sums of squares in analysis of variance for any effect eliminating all other effects.

Harvey (1968) discussed various methods of estimation in case of unequal subclass numbers using least square analysis or fitting constants method in one way, two way and heirarchial data for non-full rank and reparamaterized full rank model.

Overall and Spiegel (1969) presented three methods of analysis of the model which might be appropriate under different circumstances as follows:

Method 1: Complete least squares:- This method completely resembles with the method of weighted squares of means.

Method 2: Experimental design method:- This method is almost equal to the method of fitting constants.

Method 3: A priori ordering method:- It is a special case of fitting constants method.

Rees (1969) discussed the analysis of variance of some experimental design in which there may be several levels of splitting of plots and where treatments at any level may be non-orthogonal to blocks. The construction of designs based on Kronecker product method as used by Kurkizon and Zelan (1962) is discussed. The method of analysis by Nelder (1965, a, b) is also outlined.

Searle (1970) obtained a general expression for the elements of information matrix of maximum likelihood estimators of variance components derived from unbalanced data of any mixed model. This general expression is used to obtain explicit result for two way nested classification in random model. Searle (1971) discussed the general procedures of estimation and testing of hypotheses for linear statistical models and showed their application for unbalanced data (i.e., unequal subclass numbers data) to ascertain specific models often arised in research and survey work. The main objective was to describe the linear models techniques for analysing unbalanced data. The case

of nonfull rank model and reparametrized full rank model both are described utilizing generalized inverse matrix and giving a verified procedure for testing any testable linear hypothesis.

Overall and Klett (1972) described the regression theory to apply the R( ) notation to the reparametrized full rank model in multivariate analysis.

Hartwell and Gayler (1973) objective was to examine the method of unweighted means with regard to estimating the variance components in two way classification with unbalanced observation (with one or more cell have no observation, zero cell frequency). Two procedures M and $\bar{M}$ based upon unweighted means were developed to estimate the variance components by (i) estimating observations for the missing cells, (ii) by computing mean squares by the method of unweighted mean and (iii) equate these mean squares to their expectation and solving for variance components.

Federer and Paik (1974) solved four problems associated with the use of Zelen's calculus of factorial in the statistical analysis of nonorthogonal n-way classification data. These were the situations for which (i) some effect parameters are equated to zero, (ii) some subclass

contains no observations, (iii) experimental values of mean square under fixed, mixed and random model are desired, (iv) the expected value of sum of squares for single degree's of freedom contrast or of several single degree of freedom contrast. The various analysis are first described for non-orthogonal two way classification and generalised to n-way classification. In order to make a unified approach to the four problems, authors made use of vector of subclass means and of an orthogonal transformation of the levels of any given factor into single degree of freedom contrast which allowed to relate the concept of fractional replication from complete factorial to any n-way classification with missing subclass to illustrate the bias of the estimate of effect parameters through the aliasis matrix.

The key feature of Neter and Wasserman (1974) was its unified approach to the application of linear statistical models in regression analysis of variance and experimental design. They discussed about the indicator variable covering both dependent and independent indicator variables. Computrized selection procedures for obtaining a best set of independent variables was described. A wide variety of case examples were presented both to illustrate the great diversity of application of linear statistical models and to show how to choose the analysis procedures for different

problems.

Hocking and Speed (1975) in their article compared non-full rank model vs full rank model by considering many examples. They observed that although nonfull rank model had certain appeal in terms of describing the experiment, it had led to considerable confusion in the analysis. In reparametrized full rank model, many confusions had been eliminated, particularly in the areas of hypothesis testing. Scientist user can have the hypothesis of his own interest in a simple manner. Authors concluded that analysis through reparametrized full rank model would remove much of the confusions in the teaching of applied statistics and hence provided a better understanding among data analysis. The case of $n_{ij} = o$ in cell frequencies was also discussed by illustrating the example and tested different hypothesis.

Searle (1976) defined as 'messy data', when each sub most cell of the classification did not have the same number of observations which included the possibility of the cells might have no data at all which might be called unbalanced, nonorthogonal or unequal subclass numbers data including the possibility of empty cells. Author had discussed the procedure for calculating the R( ) notation

for $R(\mu)$, $R(\alpha/\mu)$, $R(\beta/\mu,\alpha)$, $R(\gamma/\mu,\alpha,\beta)$ and $R(\beta)$ etc. in two way classification model. Two different methods have been described where one is correct method and another is incorrect. The difference between these procedure was that the principle of calculating $R(\underline{b})$ from a model containing $\underline{b}$ was violated. The method of Yates weighted squares of means was also described.

Speed and Hocking (1976) used R( ) notation to define $R(\underline{\beta})$ and $R(\underline{\beta}^{*})$ as an aid for calculating various entries in the ANOVA table. In this article authors tried to emphasize the difference between two procedures i.e. nonfull rank model and second reparametrized full rank model. Both the methods are not identical and one is not a special case of the other. Some illustration were used where these methods agree and where disagree. When $R(\underline{\beta})$ was given, the procedure used must be indicated and the non-estimable conditions used should also be indicated if reparametrized full rank model was used. The R( ) notation should not be used to dictate what hypotheses are to be tested, but once the hypotheses to be tested were decided upon, it might be used to obtain the appropriate sum of squares. In addition the paper considered a number of related points i.e., R( ) notation did not indicate the actual hypotheses being tested and in the light of this

point many authors were misinterpreting what is being tested.

Rao's (1946) main aim was to bring out the generality of the method of least squares of which all tests of linear hypothesis came out as a special case. A unified approach to the problem of testing of linear hypothesis involved in a variety of cases had been put forward. As an application of this method for appropriate linear hypothesis and the analysis of variance and covariance in biological experiments had been considered and general theory of statistical regression was discussed.

Graybil (1961) discussed the statistical concept and testing of various hypothesis for those workers who do not have much mathematical background. The author was concerned only with the mathematical treatment of statistical models and no attempt was made to justify any model for a given real world situation. Throughout the study emphasis was made on the power of the test and on the calculation of width of the confidence interval.

Elston and Bush (1964) described the sum of squares appropriate for testing the main effects depending upon how these main effects were defined when interaction was assumed in the model. Some times in a model all the parameters were not defined but it was still possible to test

certain hypothesis. In this paper it was pointed what hypothesis about the main effects could be tested to suggest a method for programming on a computer to obtain sum of squares appropriate for testing any testable hypothesis in case of nonorthogonal data. Model 1 of analysis of variance in which all effects were fixed except μ, in two way classification was considered in detail.

Gosslac and Lucas (1965) studied the disturbance to the level of significance of additive sum of squares method of analysis in disproportionate data for several patterns of subclass numbers in two way classification. These methods yielded too many significant result for main effects under the null hypothesis, although this disturbance was judged to be moderate for the method of unweighted means. Factors to remove the bias in the method of expected subclass numbers were given in the article. A procedure for computing the power of exact method similar to the computation for equal sub class numbers and approximate procedure for determining the power of the method of unweighted means were described.

Ballas and Webester (1966) illustrated in their article, the effect of non-independent numerator on 'F' test in an analysis of variance. Symmetrical balanced incomplete block designs were considered with blocks as

random effect and no interaction. The joint density of adjusted sum of squares was determined under the null hypothesis of no treatment effect.

Carlson and Timm (1974) did link a discussion of the method to the hypothesis being tested but other offering heuristic interpretation and abstract rationalization, only added to the confusion. They used method 2 to apply the R( ) notation to the reparametrized full rank model.

Kutner (1974) related various sum of squares obtained by a number of computer programmes to the corresponding linear hypothesis being tested about the stated model. These articles were followed by some comments by Nelder (1974) in which he indicated that experiments with unbalanced data are least understood of the data structure and that the associated analysis was the worst taught in statistics courses. He appealed to the statistician to return to the analysis advocated by Yates (1934).

Hocking's et al. (1978) purpose was to describe the hypothesis commonly tested in linear models with unbalanced data including the case of zero cell frequency. The intention was to demonstrate the hypothesis under test when standard computer routines are applied to some of

the common models. The analysis were urged to study these hypothesis to see if they were reasonable before submitting the data for evaluation. Ideally the computer programmes being used should be sufficiently flexible to allow the specification of any linear hypothesis rather than restricting to user to one or more of those cited here. Clearly which placed a burden on the proportions of programmes but if the alternative was testing of inappropriate hypothesis, it seemed that additional effort was justified. In this paper, the hypothesis associated with the R( ) notation for general sets of conditions were discussed in terms of means of the observed populations. The discussion was restricted to two way model. Examples were included to illustrate the hypothesis tested with missing cells.

The objective of Speed et al. (1978) was to review existing methods for analyzing experimental design models with unbalanced data and to relate them to existing computer programmes. The methods were distinguished by the hypothesis associated with the sum of squares which were generated. The choice of method should be based on the appropriateness of the hypothesis rather than computational convenience of the orthogonality of the quadratic forms. The sum of squares were described by using R( ) notation as applied

to the overparametrized linear model but the hypotheses are stated in terms of the full rank cell means model. Hypothesis H_1 and H_5 seemed to be reasonable measure of main effects, measuring the effect of one factor when averaged over the other factors. In the presence of significant interaction, these hypothesis might not be of general interest and more specialized hypothesis might be considered.

Rao's (1955) aim was to provide some general computational techniques required in the analysis of multiple classified data when there were unequal numbers in the cells and when there were more than one character under study. Tests in the analysis of dispersion were obtained by a generalization of the corresponding tests in the analysis of variance applied to a single variable. This technique due to Fisher (1939) consisted in replacing the multiple variables by a linear compound for which variance ratio test was constructed. The compounding coefficients were chosen to maximise the ratio.

Freeman and Jeffers (1962) considered various methods of estimating means and standard errors for non-orthogonal data designs from the point of view of their generality and suitability for computer programming. The

use of variance covariance matrix was shown to have a number of advantages and method was described for design with two or three way classification. The method of analysis for three way classification, no two of which were orthogonal was new. Programmes had been written for the Ferrentic Regasus Computer for two and three way classification. All analysis of nonorthogonal designs had some common features. They observed that treatment means needed some adjustment to take account of the lack of orthogonality.

Francis (1973) compared several canned programmes to perform routine analysis of variance in two way factorial design with unequal number of observations in the cells and discussed the rationale for making a choice. When several widely available canned computer programme were used to perform a routine analysis of variance, the results were found misleading or some time completely wrong. Out of four computer programmes BMD 10 V (BMD x 64)(DIXON, 1970), Biomed series proved to be accurate, unambiguous versatile, well documented, statistically attractive, efficient and inexpensive.

Hammerle (1974) developed a method for computing an exact non-orthogonal analysis of variance using cell means. This was accomplished without forming or using

computer storage for matrix $\underline{Xo}$ or $\underline{Xo}'\underline{Xo}$ or an orthogonal transformation of $\underline{Xo}$ where $\underline{Xo}$ is nxp non-orthogonal design matrix. This method was found a convergent iterative method which utilizes balanced analysis of variance estimates and residuals iteratively in solving the relevant normal equations and conducting test of hypothesis. A monotonocity property of the method was derived to minimise iteration for nonsignificant factors interaction in hypothesis testing. A non-orthogonal analysis of variance for a fixed effects model might be computed by applying regression technique to indicator variables. Kutner(1974) related the sum of squares obtained by a number of computer programmes to the corresponding linear hypothesis being tested about the stated model. Golhar and Skillings(1976) did link a discussion of the methods to the hypothesis being tested, but others offering heuristic interpretations and abstract rationalizations only added to the confusion.

Under the present investigation, various regression methods with dummy variables will be undertaken for unbalanced data using three coding schemes. A unified regression approach will be developed which gives analysis of variance as a special case of the techniques for unbalanced data. This will remove many prevailing confusions. The measure of extent of non-orthogonality will also be studied.

CHAPTER - III

LINEAR STATISTICAL MODELS

In any analysis of variance it is very essential that the mathematical model underlying the analysis and the assumptions made in using the model should be well defined. The analysis of such linear models is well known and supported by rigorous mathematical theory by several workers e.g. Graybil (1961), Carlson and Timm (1964), Harvey (1968), Searle (1971, 1972) and Overall and Klett (1972). Here we are not making any contribution to theory but our attempt is to fill up some apparent gaps between theory and practice. In this chapter we will describe and discuss various linear statistical models which are commonly used and are generally useful to describe the experimental situations. Two types of linear models are discussed:

1. Classified fixed effects β-model
2. Cell means model or μ-model.

3.1 Classified Fixed Effects β-Model

The classified fixed effects β-model is defined as

$$\underline{Y} = \underline{X}\,\underline{\beta} + \underline{e} \quad . \quad . \quad . \qquad (3.1)$$

where

$\underline{Y}$ is an nxl vector of observations

$\underline{e}$ is an nxl vector of random errors.

which are assumed to be independent and are distributed normally with mean vector $\underline{0}$ and variance $\sigma^2\underline{I}$

$\underline{\beta}$ is a pxl vector of unknown parameters

s.t $\underline{\beta} = (\beta_1, \beta_2, \ldots \ , \beta_p)$

$\underline{X}$ is an nxp design matrix of rank q where $q < p$. The elements of $\underline{X}$ matrix are assumed to be known and are either 0's or one's.

3.1.1 Algebraic form of the β-model

It will be interesting to spell out several different models in algebraic form to emphasize the simplicity of β-model.

(a) The one way classification

In one way classification, the model is

$Y_{ij} = \mu + \alpha_i + e_{ij}$

$i = 1, 2, \ldots \ , a$

$j = 1, 2, \ldots \ , n_i$

where μ is the overall population mean when equal frequencies exist among the classes of the factors.

α_i is the effect of the i^{th} level of the factor A expressed as a deviation from the overall mean μ,

Y_{ij} is the j^{th} observation in the i^{th} level of A class.

e_{ij} is the random error assumed to be independent and distributed N $(0, \sigma^2)$.

(b) Two way classification model with interaction

The general linear mathematical model for two way classification with interaction is

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}$$

$i = 1, 2, \ldots, a$

$j = 1, 2, \ldots, b$

$k = 0, 1, 2, \ldots, n_{ij}$

where, μ and α_i are same as above.

β_j effect of j^{th} level of factor B expressed as the deviation from the population mean μ

γ_{ij} the interaction effect of j^{th} level of factor B and i^{th} level of factor A or effect of the ij^{th} AB subclass, after effects of factor A and factor B have been removed.

e_{ijk} random errors assumed to be $\sim$ NID $(0, \sigma^2)$

The mathematical model is same regardless whether the α_i and/or β_j are fixed or random effects. If both are fixed, the interaction effect are also fixed. When all effects except μ are fixed, the model is referred to a Model I of Eisenhart or the fixed effects model and if all the effects except μ, are random, it is referred to Model II of Eisenhart or the random model. The model is regarded as mixed model when one set of effects either α_i's or β_j's is fixed and other is random. In the mixed model, interaction effects are random effects.

(c) Two way classification without interaction

The mathematical linear model is as follows:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + e_{ijk}$$

$$i = 1, 2, \ldots, a$$

$$j = 1, 2, \ldots, b$$

$$k = 0, 1, 2, \ldots, n_{ij}$$

All notations are same as described in (b).

(d) The two fold nested design

Although there are many specific designs for which linear mathematical models are standard, but it may differ

considerably from the standard models for a given set of data. Most of the animal breeding data with disproportionate class numbers are classified in two more nested classification. The specific model for nested design is:

$$Y_{ijk} = \mu + \alpha_i + \gamma_{ij} + e_{ijk}$$

$$i = 1, 2, \ldots, a$$

$$j = 1, 2, \ldots, b_i$$

$$k = 1, 2, \ldots, n_{ij}$$

All the effects follow the same notations as above except γ_{ij} effect which is the effect of j^{th} level of factor B within i^{th} level of factor/class A. This model does not include a term for the interaction.

3.1.2 Features of the β-model

The essential features of the β-model are defined as follows:

i) The parameters in the β-model are not clearly understood and in fact, these are not uniquely defined without additional information.

ii) All linear functions of $\underline{\beta}$ are not estimable. The concept of estimability is often misunderstood.

iii) Estimation of estimable functions is usually achieved by imposing nonestimable conditions or restrictions on the parameters which gives the source of confusion in the statistical design with unequal cell frequencies.

iv) Only hypothesis on estimable functions of β make any sense.

v) The hypothesis are oftenly not clearly stated or if stated, not easy to understand.

3.1.3 Analysis of the β-model

Suppose the β-model is

$$\underline{Y} = \underline{X}\underline{\beta} + \underline{e}$$

with usual notations as defined earlier.

The reduction in sum of squares due to fitting such a model (either by least square method under normality assumption) is

$$R(\beta) = \hat{\underline{\beta}}' \underline{X}' \underline{Y} \quad \ldots \quad (3.1)$$

where $\hat{\underline{\beta}} = \underline{S}^{-1} \underline{X}' \underline{Y}$ and $\underline{S} = (\underline{X}' \underline{X})$

S.S.Error = Total S.S. - $R(\beta)$ ..... S.S.E.

ANOVA Table

Source of variation	d.f.	S.S.	M.S.	'F'
Due to Reg.(Fitting constants)	p-1	$R(\beta)$	MSH	$\frac{MSH}{MSE}$
Error	N-p	S.S.E.	MSE	
Total	N-1			

M.S.H. due to hypothesis is tested against M.S.E. by usual 'F' test with (p-1), and (N-p) degree of freedom.

3.2 Cell means model or μ-model

In μ-model it is assumed that the observations have been taken on each of the p normal populations (p cells) with the same variance σ^2 and, μ_i, different cell means. These means be linearly related. Symbolically μ-model can be described as follows:-

$$\underline{Y} = \underline{W}\,\underline{\mu} + \underline{e} \qquad . \quad . \quad . \qquad (3.2.1)$$

subject to the restriction.

$$\underline{G}\,\underline{\mu} = \underline{0} \qquad . \quad . \quad . \qquad (3.2.2)$$

where,

$\underline{Y}$ is an nx1 vector of all observation in all cells S.t, $n = \sum_{i=1}^{p} n_i$.

$\underset{pxl}{\underline{\mu}}$ is the pxl vector of cell means where μ_i, $i = 1, 2, \ldots, p$ corresponding to the p populations sampled.

$\underset{nxp}{\underline{W}}$ is a diagonal matrix of J_{ni}, $i = 1, 2, \ldots, p$ where J_{ni} is a vector of n_i's one's indicating the number of observations in the i^{th} populations.

$\underset{nx1}{\underline{e}}$ is a random error which is distributed independently and normally with mean vector $\underline{0}$ and common variance $\sigma^2 I$.

$\underset{qxp}{\underline{G}}$ is the matrix of order qxp with rank q and represents known linear relations about the cell means.

These restrictions defined by $\underline{G\mu} = \underline{0}$ may not be always present. In general the assumption of no interaction in the model is assumed. For illustration, the restricted full rank model, two way classification without interaction can be written as

$$Y_{ijk} = \mu_{ij} + e_{ijk} \quad . \quad . \quad . \qquad (3.2.3)$$

subject to the restriction

$$\mu_{ij} - \mu'_{ij} - \mu_{ij'} + \mu''_{ij} = 0 \quad . \quad . \quad . \qquad (3.2.4)$$

where

$i = 1, 2, \ldots, a$

$j = 1, 2, \ldots, b$

$k = 0, 1, 2, \ldots, n_{ij}$

The condition in equation (3.2.4) represents assumption about the means μ_{ij} which is based on the experience and knowledge of the research worker. Several authors e.g. Searle (1971) and Neter and Wasserman (1974) have discussed

the full rank model but only in the restricted case that is the model (3.2.1) without the restriction (3.2.2).

3.2.1 Algebraic form of the μ-model

The μ-model can be defined in several familiar forms.

(a) The one way classification model

$$Y_{ij} = \mu_i + e_{ij}$$

where, $i = 1, 2, \ldots, a$

$j = 1, 2, \ldots, n_i$

Y_{ij} is the j^{th} observation in the i^{th} population and are assumed to be independently normally distributed with mean μ_i and common variance σ^2 and that an estimate of σ^2 can be obtained by pooling together all single degrees of freedom from each cell i.e.

$$\hat{\sigma}^2 = \sum_{i=1}^{a} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})^2 \Big/ \sum_{i=1}^{a} (n_i - 1)$$

μ_i is the mean response on the i^{th} population

n_i denote the number of observations in the i^{th} population

e_{ij} is the independent random error $\sim$ as $N(0, \sigma^2)$

(b) Two way classification model with interaction

The model can be defined as

$$Y_{ijk} = \mu_{ij} + e_{ijk}$$

$i = 1, 2, \ldots, a$

$j = 1, 2, \ldots, b$

$k = 0, 1, 2, \ldots, n_{ij}$

where

Y_{ijk} = k^{th} observation on the $(i, j)^{th}$ treatment combination

μ_{ij} = Mean response on the $(i, j)^{th}$ treatment combination

e_{ijk} = random independent error distributed as $N(0, \sigma^2)$.

(c) Two way classification model without interaction

The linear model can be defined as

$$Y_{ijk} = \mu_{ij} + e_{ijk}$$

$i = 1, 2, \ldots, a$

$j = 1, 2, \ldots, b$

$k = 0, 1, 2, \ldots, n_{ij}$

such that

$$\mu_{ij} - \mu'_{ij} - \mu_{ij}' + \mu''_{ij} = 0$$

All notation are same as in (b).

(d) Two fold nested model

The linear model is as follows:

$$Y_{ijk} = \mu_{ij} + e_{ijk}$$

$$i = 1, 2, \ldots, a$$

$$j = 1, 2, \ldots, b$$

$$k = 1, 2, \ldots, n_{ij}$$

Here μ_{ij} is the mean response of j^{th} level of factor B within the i^{th} level of factor A.

3.2.2 Features of the μ-model

i) It is conceptually easy to understand. The parameters, namely, the cell means, are meaningful and easy to interpret.

ii) There is no problem with the concept of estimability. Any cell mean can be estimated if that population is observed. In addition, if the model is constrained, we may be able to obtain estimates even if the population is not directly observed.

iii) Any linear hypothesis on the observed cell means can be tested.

3.2.3. The analysis of μ-model

i) Constrained case

Let us suppose the μ-model is

$$\underline{Y} = \underline{W}\,\underline{\mu} + \underline{e}$$

$$\text{s.t. } \underline{G}\,\underline{\mu} = \underline{0}$$

With all notations and symbols as used earlier. The analysis of the μ-model follows from classical linear model theory and is simple. On the same principle, the minimum variance unbiased estimate of $\underline{\mu}$ is

$$\hat{\underline{\mu}} = \underline{A}\ (\underline{W}'\underline{W})^{-1}\ \underline{W}'\underline{Y} \qquad \ldots \qquad (3.2.3.1)$$

where

$$\underline{A} = I - (W'W)^{-1}\ G'\left[G(W'W)^{-1}\ G'\right]^{-1}G \ \ldots\ (3.2.3.2)$$

and

$$\text{Var}\ (\hat{\underline{\mu}}) = \underline{A}\ (\underline{W}'\underline{W})^{-1}\ \underline{A}'\sigma^2$$

Also $\hat{\mu} \backsim$ NID $(\mu,\ \text{var}\ (\hat{\mu}))$ Further an unbiased estimate of σ^2 is given by

$$\hat{\sigma}^2 = (\underline{Y} - \underline{W}\hat{\underline{\mu}})'\ (\underline{Y}-\underline{W}\hat{\underline{\mu}})\,/\,n-p+s$$

$$= Y'\left[I-W\ A\ (W'W)^{-1}\ A'\ W'\right]Y\ /\ n-p+s \quad (3.2.3.3)$$

and

$$\frac{\hat{\sigma}^2}{\sigma^2} \backsim \chi^2_{n-p+s}.$$

ii) Unconstrained case

When $\underline{G} = \underline{0}$

In this case $\underline{A} = \underline{I}$ and

$$\hat{\underline{\mu}} = (W' W)^{-1} \underline{W}' \underline{Y} \quad . \quad . \quad . \quad (3.2.3.4)$$

$\hat{\underline{\mu}}$ is the vector of sample means e.g. $\hat{\mu}_{ij} = \bar{Y}_{ij}$.

iii) Test of hypothesis

For testing linear hypothesis on the vector μ of the form

H_o: $\underline{H}\underline{\mu} = \underline{0}$ vs

H_A: $\underline{H}\underline{\mu} \neq \underline{0}$

where $\underline{H}$ is of qxp of rank q, the appropriate test statistic is

$$F = \frac{SSH}{q \hat{\sigma}^2} \quad . \quad . \quad . \quad (3.2.3.5)$$

where, SSH is given by

$$SSH = (\underline{H}\hat{\underline{\mu}})' \left[\underline{HA} (\underline{W}' \underline{W})^{-1} \underline{A}' \underline{H}'\right]^{-1} \underline{H}\hat{\underline{\mu}} \ldots (3.2.3.5)$$

It is noted that $\underline{SSH}$ is a non-central χ^2_q, λ, where $2\sigma^2\lambda = (\underline{H}\underline{\mu})' \left[\underline{HA}\, \sigma^2 (W' W)^{-1} \underline{A}' \underline{H}'\right]^{-1} \underline{H}\underline{\mu}$. Further SSH and $\hat{\sigma}^2$ are independent. Hence the 'F' ratio in (3.2.3.5) is distributed as $F_{q;\ n-p+s,\ \lambda}$.

3.3 Comparison between β-model and μ-model

The cell means model is not new and is actually often used as the basis for β-model.

$$\underline{Y} = \underline{X}\beta + \underline{e}$$

Let $\underline{X}$ is the order of nxk and its rank is $\leq k$. Formally the relation between the two models can be defined as:

i. $\mu = p\beta$

ii. $\underline{X}' = \underline{W}p$

iii. $\underline{G}p = 0$

iv. Rank (G) = p-k

and p are the number of population sampled.

We may write the following examples in β-model equivalents to the relation $\mu = p\beta$. The relations are as follows:

i. $\mu_i = \mu + \alpha_i$ one way classification

ii. $\mu_{ij} = \mu + \alpha_i + \beta_j$ two way classification

iii. $\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$ two way classification with interaction

iv. $\mu_{ij} = \mu + \alpha_i + \gamma_{ij}$ two way nested.

To summarise, this discussion is not intended to

conclude that μ-model is now "in" and the β-model is "out". Indeed we feel that both are useful. It does appear that the μ-model is conceptually simpler and particularly useful in consulting. The β-model has some appeal since it serves to describe the nature of the experiment. The β-model also has computational advantages in the constrained case. The effective dimension, in two way classification without interaction, in μ model is ab while the actual dimension is a+b-1 as in β-model. Of course, these are computational advantage for μ-model that could be programmed to take advantage of this reduced dimensionality.

CHAPTER - IV

THE R( ) NOTATION

By considering more complex model than one way classification, it will lead us to compare the adequacy of different models for the same set of data. In the analysis of variance table, SSE is calculated by subtracting the SSR from SST. SSR is a measure of the variation in Y accounted for by that model. Comparison of different models in terms of a given set of data is made by comparing the different values of SSR which are obtained by fitting different models for the same set of data. This SSR is referred as the reduction in sum of squares, denoted by R( ), with the contents of the brackets indicating the model fitted, for example, in fitting the model

$$Y_{ij} = \mu + \alpha_i + e_{ij}$$

the reduction in sum of squares is, $R(\mu,\alpha)$ where μ and α indicate a model which has parameters μ and those of an α-factor. Similarly the reduction in sum of squares due to fitting the model

$$Y_{ijk} = \mu + \alpha_i + \beta_j + e_{ijk}$$

is denoted by $R(\mu, \alpha, \beta)$. The reduction in sum of squares due to fitting the nested model

$$Y_{ijk} = \mu + \alpha_i + \beta_{ij} + e_{ijk}$$

is denoted by $R(\mu, \alpha, \beta : \alpha)$ which indicates that β-factor is nested within α-factor. This letter R is referred only for 'Reduction' in sum of squares and not for residual.

This R( ) notation has been used by several authors as an aid for calculating the various entries in ANOVA table. The R( ) notation provides a convenient means for describing the hypothesis tested and the computational procedure used to obtain the sum of squares associated with a hypothesis. This R( ) notation suffers some deficiencies.

The first one is that Speed and Hocking (1976) defined two different and distinct ways in which R( ) notation can be defined for unbalanced data. The definition of R( ) notation is different even for balanced data. The one described by Searle (1971, 1972) is well defined but does not indicate utility for testing hypothesis. He defined it to find out the solutions of normal equations for the original non full rank linear model but cautioned that it is solely for finding the solutions of normal equations

and in no way, these solutions can be taken as an estimate of the population parameter vector. Moreover, the parameters in the original non full rank model are neither defined, nor these are individually estimable and testable and the second as described by Harvey (1968), Overall and Spiegel (1969), Winer (1971), Overall and Klett (1972) and Carlson and Timm (1974), is more flexible but the sums of squares obtained depend on how the β-model is derived from the μ-model and procedure is to apply the R( ) notation to the reparametrized full rank model.

The μ-model and over parametrized β-model, when derived from the μ-model, results in nonestimable restrictions on the parameters in the overparametrized model which are considered as the integral part of the model. When the β-model is considered along with the nonestimable restrictions, all the parameters in the model are uniquely defined and are linear functions of μ_{ij}'s.

Another serious draw back of the R( ) notation is that it does not indicate the actual hypothesis being tested rather it suggests such statements like testing α effects after fixing β effects, ignoring γ effects. Precise statements of the hypothesis are more helpful for the interpretation of the results.

Given the β-model

$$\underline{Y} = \underline{X}\underline{\beta} + \underline{e}$$

the R( ) notation is defined as

$$R(\underline{\beta}) = \hat{\underline{\beta}}' \underline{X}' \underline{Y} \quad . \quad . \quad . \quad (4.1)$$

where $\hat{\underline{\beta}}$ is any solution to

$$\underline{X}' \underline{X} \hat{\underline{\beta}} = \underline{X}' \underline{Y} \quad . \quad . \quad . \quad (4.2)$$

These are the normal equations for solving the population parameter $\underline{\beta}$.

Let $\underline{X}$ and $\underline{\beta}$ be confirmable partitioned into

$$\underline{X} = (\underline{X}_1, \underline{X}_2)$$

$$\underline{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$$ respectively and

$$\underline{Y} = \underline{X}_2 \underline{\beta}_2 + \underline{e} \quad . \quad . \quad . \quad (4.3)$$

model is considered thus

$$R(\underline{\beta}_2) = \hat{\underline{\beta}}_2' \underline{X}_2' \underline{Y} \quad . \quad . \quad . \quad (4.4)$$

where $\hat{\underline{\beta}}_2$ is any solutions to the normal equations

$$\underline{X}_2' \underline{X}_2 \hat{\underline{\beta}}_2 = \underline{X}_2' \underline{Y} \quad . \quad . \quad . \quad (4.5)$$

In addition it can be defined that

$$\begin{aligned} R(\underline{\beta}_1 / \underline{\beta}_2) &= R(\underline{\beta}) - R(\underline{\beta}_2) \\ &= R(\underline{\beta}_1, \underline{\beta}_2) - R(\underline{\beta}_2) \quad . \quad . \quad . \quad (4.6) \end{aligned}$$

which is the reduction in sum of squares due to fitting $\underline{\beta}_1$ after fixing $\underline{\beta}_2$. It is seen from the above that

although R( ) notation is well defined but it depends on whether it is applied to non full rank model or full rank model or reparametrized full rank model. In addition, if R( ) notation is applied to the reparametrized full rank model, the results will depend upon the set of nonestimable conditions used to achieve the full rank model.

4.1 Application of R( ) Notation.

4.1.1 Method 1

Suppose non-full rank model is

$$\underline{Y} = \underline{X}\underline{\beta} + \underline{e} \quad ,$$

i) Calculate $R(\underline{\beta}) = \hat{\underline{\beta}}\,\underline{X}'\underline{Y}$ where $\hat{\underline{\beta}}$ is the solution of normal equation $\underline{X}'\underline{X}\hat{\underline{\beta}} = \underline{X}'\underline{Y}$

Here nonestimable conditions may be applied to solve for $\hat{\underline{\beta}}$, but the results would be independent of the choice of nonestimable conditions. Here we mean to say that $R(\underline{\beta})$ will always be same what ever nonestimable conditions be applied.

ii) Assume the model $\underline{Y} = \underline{X}_2\underline{\beta}_2 + \underline{e}$ where (X_1, X_2) and $\begin{pmatrix} \underline{\beta}_1 \\ \underline{\beta}_2 \end{pmatrix}$ are confirmable partition of $\underline{X}$ and $\underline{\beta}$ respectively.

iii) $R(\underline{\beta}_2) = \hat{\underline{\beta}}_2'\,\underline{X}_2'\,\underline{Y}$ where $\underline{X}_2'\underline{X}_2\hat{\underline{\beta}}_2 = \underline{X}_2'\underline{Y}$

Here also the result will depend upon the assumed

nonestimable conditions to solve for $\hat{\underline{\beta}}_2$ but the results are independent of the choice of nonestimable conditions assumed in (i) for calculating $R(\underline{\beta})$.

4.1.2 Method 2

Here we shall be dealing with full rank model. R( ) notation depends upon the definitions of the parameters to describe the sum of squares in the β-model which may be different as the definitions of the parameters change . It may be noted that R( ) is always read along with the definition of the parameters resulting and most of the confusion will be clear about the different values of R( ) by using a different set of non-estimable conditions. The sets of nonestimable conditions, which are used to full rank model, define uniquely the parameters. The following procedure of applications of R( ) notation in case of reparametrized full rank model is adopted:

i) Choose a set of nonestimable condition and reparametrize the non full rank model which turns out to be

$$Y = \underline{X}^* \underline{\beta}^* + \underline{e} \quad \text{as a full rank model.}$$

It is important to note that the set of nonestimable conditions is used throughout method 2.

ii) Calculate $R(\underset{\sim}{\beta}^*) = \hat{\underset{\sim}{\beta}}^{*\prime} \underset{\sim}{X}^{*\prime} \underset{\sim}{Y}$

where $\hat{\underset{\sim}{\beta}}^* = (\underset{\sim}{S}^*)^{-1} \underset{\sim}{X}^{*\prime} \underset{\sim}{Y}$

and $\underset{\sim}{S}^* = \underset{\sim}{X}^{*\prime} \underset{\sim}{X}^*$

iii) Assume

$\underset{\sim}{\beta}^* = \begin{pmatrix} \underset{\sim}{\beta}_1^* \\ \underset{\sim}{\beta}_2^* \end{pmatrix}$ and $\underset{\sim}{X}^* = (\underset{\sim}{X}_1^*, \underset{\sim}{X}_2^*)$ be the confirmable partitions. Let

$$\underset{\sim}{Y} = \underset{\sim}{X}_2^* \underset{\sim}{\beta}_2^* + \underset{\sim}{e}$$

then

$$R(\underset{\sim}{\beta}_2^*) = \hat{\underset{\sim}{\beta}}_2^{*\prime} \underset{\sim}{X}_2^{*\prime} \underset{\sim}{M}$$

where

$$\hat{\underset{\sim}{\beta}}_2^* = (\underset{\sim}{S}_2^*)^{-1} \underset{\sim}{X}_2^{*\prime} \underset{\sim}{Y} \quad \text{and}$$

$$\underset{\sim}{S}_2^* = \underset{\sim}{X}_2^{*\prime} \underset{\sim}{X}_2^*$$

Therefore

$$R(\underset{\sim}{\beta}_1^* / \underset{\sim}{\beta}_2^*) = R(\underset{\sim}{\beta}^*) - R(\underset{\sim}{\beta}_2^*)$$

4.2 Results

i) It is observed that while $R(\underset{\sim}{\beta}) = R(\underset{\sim}{\beta}^*)$, it is not necessary that $R(\underset{\sim}{\beta}_2^*)$ is also equal to $R(\underset{\sim}{\beta}_2)$.

ii) $R(\alpha^*/\mu^*)$, $R(\beta^*/\mu^*)$, $R(\alpha^*/\mu^*,\beta^*)$, $R(\beta^*/\mu^*,\alpha^*)$ $R(\alpha^*, \beta^*/\mu^*)$, $R(\mu^*, \alpha^*, \beta^*)$ and $R(\mu^*, \alpha^*, \beta^*, \gamma^*)$ are always the same irrespective of how the β-model is derived from the μ-model.

iii) $R(\gamma^*/\mu^*)$ differs and depends on how the β-model is derived from the μ-model. It differs even for the balanced case likewise $R(\alpha^*, \gamma^*/\mu^*)$ and $R(\beta^*, \gamma^*/\mu^*)$ differ.

iv) $R(\mu^*, \alpha^*, \gamma^*) \neq R(\mu^*, \alpha^*, \beta^*, \gamma^*)$ and

$R(\mu^*, \beta^*, \gamma^*) \neq R(\mu^*, \alpha^*, \beta^*, \gamma^*)$

Speed and Hocking (1976) stated this result saying that demonstration of this point is not included since closed form solutions are not readily available and the interested reader is referred to verify this point by applying to a set of data with unequal number of observations per cell. But this explanation is inappropriate since as explained in (iii) above, firstly, it can be verified with any data balanced or unbalanced and secondly there is no reason to find closed form solutions since for a full rank model, quantities like $R(\mu^*, \alpha^*, \gamma^*)$ and $R(\mu^*, \alpha^*, \beta^*, \gamma^*)$ stand entirely for a different and distinct two models which shows that Method 1 and 2 are not identical.

v) $R(\beta_2^*)$ is always defined in terms of μ_{ij} and its value always depends upon how the β-model is derived or in other words what are the non-estimable restrictions. Speed and Hocking (1976) claimed that $R(\beta_2^*)$ is not always defined which is incorrect and inappropriate and misleading. They

mis-illustrated it by considering the model for two way classification with interaction and non-estimable restrictions.

$$\mu = \alpha_i = \beta_j = 0 \text{ for all i and j}$$

Here $R(\mu^*, \alpha^*)$, $R(\mu^*, \beta^*)$, $R(\mu^*, \alpha^*, \beta^*)$ are all identically zero by virtue of non-estimable restrictions and $R(\underset{\sim}{\beta}_2^*)$ is defined.

v) In methods 1 and 2 we can choose any set of non-estimable restrictions even if we choose the same set of non-estimable restrictions, the result will differ.

vi) In Method 1 and Method 2, the actual hypothesis being tested is not clear.

The sum of squares $R(\underset{\sim}{\beta}_2/\underset{\sim}{\beta}_1)$ and $R(\underset{\sim}{\beta}_2^*/\underset{\sim}{\beta}_1^*)$ are used in ANOVA table for testing same hypothesis about the parameters. The R( ) notation, in no way, indicate the nature of hypothesis being tested. Secondly $R(\alpha^*/\mu^*, \beta^*, \gamma^*)$ as well as $R(\beta^*/\mu^*, \alpha^*, \gamma^*)$ test different hypothesis depending upon which set of non-estimable conditions are used. Searle (1972) indicated that when using method 1, $R(\alpha/\mu, \beta, \gamma)$ and $R(\beta/\mu, \alpha, \gamma)$ are identically equal to zero. It is interesting to note that $R(\alpha^*/\mu^*, \beta^*, \gamma^*)$ and $R(\beta^*/\mu^*, \alpha^*, \gamma^*)$ correspond to the sum of squares suggested by Yates (1934) with method 2 after imposing

the non-estimable restrictions as

$$\sum_{i} \alpha_i = \sum_{j} \beta_j = \sum_{i} \gamma_{ij} = \sum_{j} \gamma_{ij} = 0$$

vii) R( ) notation is of little help in identifying the different hypothesis. $R(\alpha/\mu)$ is defined as the reduction in sum of squares for testing α after ignoring β and γ and $R(\alpha/\mu,\beta)$ is defined as the reduction in sum of square for testing α after adjusting for β but ignoring γ and $R(\alpha/\mu, \beta, \gamma)$ is the reducing in sum of squares for testing α after adjusting β and γ . While this terminology is consistant with R( ) notation, it in no way, indicates the actual hypothesis being tested.

4.3 Steps for the calculation of R( )

Suppose we are interested to test the four crosses of cows namely ½ Holstein Friesian + ½ Hariana, ½ Brown Swiss + ½ Hariana, ½ Jersey + ½ Hariana and ½ Red Dane + ½ Hariana, for their milk yield. The animals are grouped according to their body weights. There are different number of animals in each cell. Let the observations on yield presented in the following table.

Table 1. 300 days milk yield of four breeds of cows.

	Body weights				
	B_1	B_2	B_3	B_4	Total
½F+½H	Y_{111}	Y_{121}	Y_{131}	Y_{141}	$Y_{1.}$
	Y_{112}	Y_{122}	Y_{132}	Y_{142}	$(n_{1.})$
	$Y_{11}n_{11}$	$Y_{12}n_{12}$	$Y_{13}n_{13}$	$Y_{14}n_{14}$	
½B+½H	Y_{211}	Y_{221}	-	Y_{241}	$Y_{2.}$
	Y_{212}	Y_{222}	-	Y_{242}	$(n_{2.})$
	$Y_{21}n_{21}$	$Y_{22}n_{22}$	-	$Y_{24}n_{24}$	
½J+½H	-	-	-	-	
½R+½H	Y_{411}	-	-	Y_{441}	$Y_{4.}$
	Y_{412}	-	-	Y_{442}	$(n_{4.})$
	$Y_{41}n_{41}$	-	-	$Y_{44}n_{44}$	
Total	$Y_{.1.}$	$Y_{.2.}$	$Y_{.3.}$	$Y_{.4.}$	$Y_{..}$
	$(n_{.1})$	$(n_{.2})$	$(n_{.3})$	$(n_{.4})$	$(n_{..})$

Where $n_{i.} = \sum_{j} n_{ij}$ $\quad n_{.j} = \sum_{i} n_{ij}$

$n_{..} = \sum_{i}\sum_{j} n_{ij}$

$Y_{i..} = \sum_{j}\sum_{k} Y_{ijk}$ $\quad Y_{.j.} = \sum_{e}\sum_{k} Y_{ijk}$

$i = 1, 2, .., 4$ breeds

$j = 1, 2, .., 4$ body weights

$k = 1, 2 \ldots, n_{ij}$ with $n_{ij} > 0$

The above data can be analyzed as two way classification model with interaction which is defined as

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}$$

Where

Y_{ijk} is the 300 days milk yield of the k^{th} cows of j^{th} body weight group in i^{th} breed of the cows.

μ is the overall population mean

α_i is the effect of i^{th} breed

β_j is the effect of j^{th} body weight group

γ_{ij} is the interaction between breed of the cows and body weight groups

e_{ijk} is the error term $\sim$ N.I.D. (o, σ^2)

In this situation, two analysis of variance table are made for the above data. In the first analysis, reduction in sum of squares $R(\alpha/\mu)$ due to α, ignoring β and γ, and reduction in sum of squares $R(\beta/\mu,\alpha)$ due to β-factor after adjusting α, ignoring γ, are calculated where as in the 2nd analysis of variance table,

the reduction in sum of squares due to β ignoring α and γ, $R(\beta/\mu)$ and $R(\alpha/\mu,\beta)$, the reduction in sum of squares due to α after adjusting β, ignoring γ are calculated. Suppose there are s = ab different cells consisting of n.. total observations in the whole experiment. Following is the analyses of variance table.

Table 2. Analysis of variance table

Analysis 1			Analysis 2		
Source	Sum of squares	d.f.	Source	Sum of squares	d.f.
μ	$R(\mu)$	1	μ	$R(\mu)$	1
α corrected	$R(\alpha/\mu)$	(a-1)	β corrected	$R(\beta/\mu)$	(b-1)
β -adjusted	$R(\beta/\mu,\alpha)$	(b-1)	α adjusted	$R(\alpha/\mu,\beta)$	(a-1)
Interaction	$R(\gamma/\mu,\alpha,\beta)$	(s-(a+b-1)	Interaction	$R(\gamma/\mu,\alpha,\beta)$	s-(a+b-1)
S.S.E.		n..-s	Error		n..-s
Total		n..	Total		n..

Where

$$R(\mu) = \frac{Y^2_{...}}{n..}$$

$$R(\mu,\alpha) = \sum_i \frac{Y^2_{i.}}{n_{i.}}$$

$$R(\mu,\beta) = \sum_j \frac{Y^2_{.j.}}{n_{j.}}$$

Since this model involves the terms μ, α_i, β_j and γ_{ij},

the sum of squares for fitting the model can be denoted by $R(\mu, \alpha, \beta, \gamma) = \hat{\beta}' \underline{X}' \underline{Y}$

$$= \sum_i \sum_j \frac{Y_{ij.}^2}{n_{ij}}$$

Where $Y_{ij.} = \sum_{k=1}^{n_{ij}} Y_{ijk}$

The other terms needed for the analysis is $R(\mu, \alpha, \beta)$ which is the sum of squares due to fitting the model

$$Y_{ijk} = \mu + \alpha_i + \beta_j + e_{ijk}$$

Hence $R(\mu, \alpha, \beta) = R(\alpha, \beta/\mu) + R(\mu)$

$= \hat{\beta}_o' \underline{X}' \underline{Y}$ under the above model

where $\hat{\beta}_o$ are the parameter vector.

$= \frac{Y_{i.}^2}{n_{i.}} + \beta'$

Where $\beta_j = Y_{.j.} - \sum n_{ij} \bar{Y}_{i.}$ for $j = 1,2 \ldots b-1$

S.S.Total $= \sum_i \sum_j \sum_k Y_{ijk}^2$

S.S.E. $=$ S.S.Total $- R(\mu, \alpha, \beta, \gamma)$

Similarly $R(\mu, \beta, \gamma)$ is the reduction in sum of squares due to fitting the model.

$$Y_{ijk} = \mu + \beta_j + \gamma_{ij} + e_{ijk}$$

and is equal to $\sum_i \sum_j \frac{Y_{ij.}^2}{n_{ij}}$

Now the sum of squares given in the table can be calculated as under:

$$R(\alpha/\mu) = R(\mu,\alpha) - R(\mu)$$

$$R(\beta/\mu) = R(\mu,\beta) - R(\mu)$$

$$R(\beta/\mu,\alpha) = R(\mu,\alpha,\beta) - R(\mu,\alpha)$$

$$R(\alpha/\mu,\beta) = R(\mu,\alpha,\beta) - R(\mu,\beta)$$

$$R(\gamma/\mu,\alpha,\beta) = R(\mu,\alpha,\beta,\gamma) - R(\mu,\alpha,\beta)$$

$$R(\alpha,\beta/\mu) = R(\alpha/\mu) + R(\beta/\mu,\alpha)$$

$$R(\alpha/\mu,\beta,\gamma) = R(\mu,\alpha,\beta,\gamma) - R(\mu,\beta,\gamma) = 0$$

Despite the above result that $R(\alpha/\mu,\beta,\gamma) = 0$ there are ways of computing of $R(\alpha/\mu,\beta,\gamma)$ and getting a non-zero value for it. To distinguish it we call it $R(\alpha^*/\mu^*,\beta^*,\gamma^*)$, and describe its calculation by first considering the procedure for calculating $R(\mu^*,\alpha^*,\beta^*,\gamma^*)$ based on $\hat{\underline{\beta}}'\,\underline{X}'\,\underline{Y} = R(\underline{\beta})$ = inner product of the solutions vector and R.H.S. of the normal equation.

4.3.1 Correct procedure for calculating $R(\mu,\alpha,\beta,\gamma)$

i) Write down the model with μ, α, β and γ.

ii) Write down the normal equation for μ, α, β and γ.

iii) Amend the equations in (ii) to be of full rank.

iv) Solve the amended equations in (iii)

v) Calculate $R(\mu,\alpha,\beta,\gamma)$ = inner product of the solutions and R.H.S. of (ii).

The important thing is to note here that calculation of $R(\mu, \alpha, \beta, \gamma)$ starts from the above model which contain μ, α, β, and γ.

4.3.2 Correct procedure for calculating $R(\mu, \beta, \gamma)$

i) Write down the model with μ, α, β and γ.

ii) Reduce the model μ, β, γ by deleting α

iii) Write down the normal equations for this reduced model.

iv) Amend the equations in (iii) to be of full rank.

v) Solve the equations which are amended.

vi) Calculate $R(\mu, \beta, \gamma)$ = Inner product of the solution and R.H.S. of (iii).

A consequence of this procedure is that

$R(\mu, \alpha, \beta, \gamma) = R(\mu, \beta, \gamma)$.

4.3.3 An incorrect procedure for calculating $R(\mu, \beta, \gamma)$

i) Write down the model with μ, α, β, and γ

ii) Write all normal equations for the model in (i)

iii) Amend the equations in (ii) to be of full rank.

iv) Reduce the equations in (iii) by deleting α's

v) Solve the equation in (iv)

vi) Calculate $R(\mu, \beta, \gamma)$ = I.P.O. solutions and R.H.S. of (iv)

The result will be that

$$R(\mu, \beta, \gamma) \neq R(\mu, \alpha, \beta, \gamma)$$

and hence

$$R(\alpha/\mu, \beta, \gamma) = R(\mu, \alpha, \beta, \gamma) - R(\mu, \beta, \gamma)$$
$$\neq 0.$$

The difference between this procedure and earlier one is that the principle of calculating $R(\underline{\beta})$ from a model containing just $\underline{\beta}$ is violated. The model here is specified in (i) of (4.3.3) which contain $\mu, \alpha, \beta, \gamma$. The normal equations in (ii) are for the model and are amended in (iii) and then in (iv) reduction after deleting α. Here it is emphasized that this is a reductions of equations not of a model which is given in (i) of (4.3.2) which gives the correct calculation for $R(\mu, \beta, \gamma)$. So, it should be always a reduction of model so as to start the calculation of $R(\underline{\beta})$ from the correct model. This will apply for all linear models.

4.4 Examples

4.4.1 Application of Method 1

We begin by considering a simple experiment in which a scientist wants to test two high yielding varieties of wheat for yield differences and also to test the effects

of three types of fertilizers. Thereware 15 experimental plots available for conducting research in the field. The following table gives the results of experiment.

Table 3. Yield in (kg) of two varieties of wheat under three types of fertilizers with unequal number of observations per cell.

		Types of fertilizers			Total
		F_1	F_2	F_3	
Varieties	V_1	8,13,9 (3)	12 (1)	7,11 (2)	60(6)
	V_2	11,14,17 (3)	14,16 (2)	10,11 14,13 (4)	120(9)
		72(6)	42(3)	66(6)	180(15)

The figures in the parenthesis show the number of observations.

The above data are summarized in a linear model as follows:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}$$

with usual notations as described earlier

where $i = 1,2$

$j = 1,2,3$

Now various reduction in sum of squares are calculated as under:

$$R(\mu) = \left(\frac{\sum\sum\sum Y_{ijk}}{n..}\right)^2 = \frac{180^2}{15} = 2160.00$$

$$R(\alpha/\mu) = R(\mu,\alpha) - R(\mu)$$

$$= \sum_i \frac{Y_{i.}^2}{n_{i.}} - 2160$$

$$= \frac{60^2}{6} + \frac{120^2}{9} - 2160.00 = 40.00$$

$$R(\beta/\mu) = R(\mu,\beta) - R(\mu)$$

$$= \sum_{j=1}^{b} \frac{Y_{.j.}^2}{n_{.j.}} - 2160.00$$

$$= \frac{72^2}{6} + \frac{42^2}{3} + \frac{66^2}{6} - 2160 = 18.00$$

$$R(\mu,\alpha,\beta) = R(\alpha,\beta/\mu) + R(\mu) = \hat{\beta}_o' \underline{x}' \underline{y}$$

$$= 59.14 + 2160.0$$

$$= 2219.14 + \hat{\beta}_o' \underline{x}'$$

which is obtained by reducing the model after deleting γ from the model

$$= \frac{Y_{i.}^2}{n_{i.}} + \frac{Y_{.j.}^2}{n_{.j.}} + C.F.$$

$$= 40.00 + 2160 = 2200.00$$

$$R(\mu,\alpha,\beta,\gamma) = \sum_i \sum_j \frac{Y_{ij.}^2}{n_{ij}}$$

$$= \frac{30^2}{3} + 12^2 + \frac{18^2}{2} \ldots\ldots \frac{48^2}{4}$$

$$= 2220.00$$

It is emphasized here that the value of $R(\mu,\alpha,\beta,\gamma)$ will be same irrespective of any nonestimable conditions

that are imposed on the model.

Table 4. The following are the normal equations of the form

$$\underline{X}'\underline{X}\,\hat{\underline{\beta}} = \underline{X}'\underline{Y}$$

Table 4:- Normal equations in matrix form

μ	α_1	α_2	β_1	β_2	β_3	γ_{11}	γ_{12}	γ_{13}	γ_{21}	γ_{22}	γ_{23}			
15	6	9	6	3	6	3	1	2	3	2	4	$\hat{\mu}$		180
6	6	0	3	1	2	3	1	2	0	0	0	$\hat{\alpha}_1$		60
9	0	9	3	2	4	0	0	0	3	2	4	$\hat{\alpha}_2$		120
6	3	3	6	0	0	3	0	0	3	0	0	$\hat{\beta}_1$		72
3	1	2	0	3	0	0	1	0	0	2	0	$\hat{\beta}_2$	=	42
6	2	4	0	0	6	0	0	2	0	0	4	$\hat{\beta}_3$		66
3	3	0	3	0	0	3	0	0	0	0	0	$\hat{\gamma}_{11}$		30
1	1	0	0	1	0	0	1	0	0	0	0	$\hat{\gamma}_{12}$		12
2	2	0	0	0	2	0	0	2	0	0	0	$\hat{\gamma}_{13}$		18
3	0	3	3	0	0	0	0	0	3	0	0	$\hat{\gamma}_{21}$		42
2	0	2	0	2	0	0	0	0	0	2	0	$\hat{\gamma}_{22}$		30
4	0	4	0	0	4	0	0	0	0	0	4	$\hat{\gamma}_{23}$		48

(4.4.1)

For obtaining the estimates, $\hat{\underline{\beta}}$ of the parameters, we have to amend these equations to be of full rank by using the following non-estimable conditions

$$\sum_i \alpha_i = 0 \quad = \quad \sum_j \beta_j = \sum_i \gamma_{ij} = \sum_j \gamma_{ij} = 0 \quad ..(4.4.2)$$

for all i and j

When $\sum_i \alpha_i = 0$

$$\alpha_1 + \alpha_2 = 0$$

$$\alpha_2 = -\alpha_1$$

Replace α_2 by putting the value $(-\alpha_1)$

Similarly

$$\sum_j \beta_j = 0$$

$$\beta_1 + \beta_2 + \beta_3 = 0$$

thus $\beta_3 = -(\beta_1 + \beta_2)$ and replace β_3 and so on for γ's

The amended normal equations reduces to

$$\begin{bmatrix} 15 & -3 & 0 & -3 & 2 & 1 \\ -3 & 15 & 2 & 1 & 0 & -3 \\ 0 & 2 & 12 & 6 & -2 & -2 \\ -3 & 1 & 6 & 9 & -2 & -3 \\ 2 & 0 & -2 & -2 & 12 & 6 \\ 1 & -3 & -2 & -3 & 6 & 9 \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{\alpha}_1 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\gamma}_{11} \\ \hat{\gamma}_{12} \end{bmatrix} = \begin{bmatrix} 180 \\ -60 \\ 6 \\ -24 \\ 18 \\ 12 \end{bmatrix} \quad ....(4.4.3)$$

For obtaining $R(\mu, \beta, \gamma)$, set $\alpha_1 = 0$ in the above model. The amended normal equation reduces to

$$\begin{bmatrix} 15 & 0 & -3 & 2 & 1 \\ 0 & 12 & 6 & -2 & -2 \\ -3 & 6 & 9 & -2 & -3 \\ 2 & -2 & -2 & 12 & 6 \\ 1 & -2 & -3 & 6 & 9 \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\gamma}_{11} \\ \hat{\gamma}_{12} \end{bmatrix} = \begin{bmatrix} 180 \\ 60 \\ -24 \\ 18 \\ 12 \end{bmatrix} \quad \ldots\ldots(4.4.4)$$

The equation (4.4.4) gives the solution of the parameters.

$$(\hat{\mu}, \hat{\beta}_1, \hat{\beta}_2, \hat{\gamma}_{11}, \hat{\gamma}_{12}) = \left(12\tfrac{3}{7}, \tfrac{-3}{7}, 1\tfrac{13}{14}, \tfrac{-6}{7}, 1\tfrac{1}{14}\right)$$

$$R(\mu, \beta, \gamma) = \hat{\mu} Y_{1.} + \hat{\beta}_1 Y_{.1.} + \hat{\beta}_2 Y_{.2.} + \ldots\ldots$$

$$= 12\tfrac{3}{7}(180) - \tfrac{3}{7}(6) + 1\tfrac{13}{14}(-24) - \tfrac{6}{7}(18) + 1\tfrac{1}{14}(12)$$

$$= 2185\tfrac{5}{7}$$

$$R(\alpha/\mu, \beta, \gamma) = R(\mu, \alpha, \beta, \gamma) - R(\mu, \beta, \gamma)$$

$$= 2220 - 2185\tfrac{5}{7} = 34\tfrac{2}{7}$$

$R(\alpha/\mu, \beta, \gamma)$ is also obtained directly from the following formula:

$$= \sum_i \frac{\left(\sum_j \bar{y}_{ij.}\right)^2}{\sum_j \left(\frac{1}{n_{ij}}\right)} - \left[\sum_i \frac{\sum_j \bar{y}_{ij.}}{\sum_j \frac{1}{n_{ij}}}\right]^2 \Bigg/ \sum_i \frac{1}{\sum_j \frac{1}{n_{ij}}}$$

$$= \frac{(10+12+9)^2}{\frac{1}{3}+1+\frac{1}{2}} + \frac{(14+15+12)^2}{\frac{1}{3}+\frac{1}{2}+\frac{1}{4}} - \frac{\left\{\frac{10+12+9}{\frac{1}{3}+1-1\frac{1}{2}} + \frac{14+15+12}{\frac{1}{3}+\frac{1}{2}+\frac{1}{4}}\right\}^2}{\frac{1}{\frac{1}{3}+1+\frac{1}{2}} + \frac{1}{\frac{1}{3}+\frac{1}{2}+\frac{1}{4}}} = 34\tfrac{2}{7}$$

Now two analysis of variance tables are

Table 5. Analysis of variance

Analysis I			Analysis II		
Source	d.f.	S.S.	Source	d.f.	S.S.
$R(\mu)$	1	2160.00	$R(\mu)$	1	2160.00
$R(\alpha/\mu)$	1	40.00	$R(\beta/\mu)$	2	18.00
$R(\beta/\mu,\alpha)$	2	19.14	$R(\alpha/\mu,\beta)$	1	41.14
$R(\gamma/\mu,\alpha,\beta)$	2	0.86	$R(\gamma/\mu,\alpha,\beta)$	2	0.86
SSE	9	52.00	SSE	9	52.00
SST	15	2272.00	SST	15	2272.00

Equation (4.4.3) has a unique solutions which on utilizing (4.4.2) satisfies (4.4.1) but there are many other solutions of (4.4.1).

4.4.2 Method 2.

In order to illustrate method 2, we again consider an experimental situation described by a two way classification model with interaction. The model is given by

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}$$

with usual notations.

Here also we take the same numerical example described under Method 1 and use the non-estimable

conditions viz. $\sum_i \alpha_i = 0$, $\sum_j \beta_j = 0$, $\sum_i \gamma_{ij} = 0$ and $\sum_j \gamma_{ij} = 0$ to reparametrize the model and to make it a full rank model. The design matrix $\underline{X}^*$ reduces to

$$\underline{X}^* = \begin{array}{c} \begin{matrix} \hat{\mu}^* & \hat{\alpha}_1^* & \hat{\beta}_1^* & \hat{\beta}_2^* & \hat{\gamma}_{11}^* & \hat{\gamma}_{12}^* \end{matrix} \\ \begin{bmatrix} 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & 0 & -1 & 0 \\ 1 & -1 & 1 & 0 & -1 & 0 \\ 1 & -1 & 1 & 0 & -1 & 0 \\ 1 & -1 & 0 & 1 & 0 & -1 \\ 1 & -1 & 0 & 1 & 0 & -1 \\ 1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & -1 & 1 & 1 \end{bmatrix} \end{array}$$

The corresponding normal equations $\underline{X}^{*\prime} \underline{X}^* \hat{\underline{\beta}}^* = \underline{X}^{*\prime} \underline{Y}$ are as follows:

$$\begin{array}{c} \begin{matrix} \hat{\mu}^* & \hat{\alpha}_1^* & \hat{\beta}_1^* & \hat{\beta}_2^* & \hat{\gamma}_{11}^* & \hat{\gamma}_{12}^* \end{matrix} \\ \begin{bmatrix} 15 & -3 & 0 & -3 & 2 & 1 \\ -3 & 15 & 2 & 1 & 0 & -3 \\ 0 & 2 & 12 & 6 & -2 & -2 \\ -3 & 1 & 6 & 9 & -2 & -3 \\ 2 & 0 & -2 & -2 & 12 & 6 \\ 1 & -3 & -2 & -3 & 6 & 9 \end{bmatrix} \end{array} \begin{bmatrix} \hat{\mu}^* \\ \hat{\alpha}_1^* \\ \hat{\beta}_1^* \\ \hat{\beta}_2^* \\ \hat{\gamma}_{11}^* \\ \hat{\gamma}_{12}^* \end{bmatrix} = \begin{bmatrix} 180 \\ -60 \\ 6 \\ -24 \\ 18 \\ 12 \end{bmatrix}$$

The solution of the above normal equations multiplied by the corresponding element of R.H. side of the normal equations gives $R(\mu^*, \alpha^*, \beta^*, \gamma^*) = 2220.00$. Next, for obtaining $R(\mu^*, \alpha^*)$, we set

$$\beta_2^* = \beta_3^* = \gamma_{11}^* = \gamma_{12}^* = 0$$

Then, obviously the appropriate normal equations reduce to

$$\begin{matrix} & \hat{\mu}^* & \hat{\alpha}_1^* \\ \end{matrix}$$

$$\begin{bmatrix} 15 & -3 \\ -3 & 15 \end{bmatrix} \begin{bmatrix} \hat{\mu}^* \\ \hat{\alpha}_1^* \end{bmatrix} = \begin{bmatrix} 180 \\ -60 \end{bmatrix}$$ for the reduced model.

Using the solution of these normal equations we obtain,

$R(\mu^*, \alpha^*) = 2200.00$.

Hence, $R(\alpha^*/\mu^*) = 2200.00 - 2160.00 = 40.00$

Similarly by setting the appropriate parameters equal to zero, we obtain $R(\beta^*/\mu^*, \alpha^*)$ and $R(\gamma^*/\mu^*, \alpha^*, \beta^*)$ which are presented in the following analysis of variance table.

Table 6. Analysis of variance table

Source	d.f.	S.S.
$R(\mu^*)$	1	2160.00
$R(\alpha^*/\mu^*)$	1	40.00
$R(\beta^*/\mu^*, \alpha^*)$	2	19.14
$R(\gamma^*/\mu^*, \alpha^*, \beta^*)$	2	0.86
Error	9	52.00
S.S.Total(crude)	15	2272.00

While comparing the analysis of variance table 5 given in method 1 and table 6 in method 2, we observe that

$R(\mu) = R(\mu^*) = 2160.00$

$R(\alpha/\mu) = R(\alpha^*/\mu^*) = 40.00$

$R(\beta/\mu,\alpha) = R(\beta^*/\mu^*,\alpha^*) = 19.14$

$R(\gamma/\mu,\alpha,\beta) = R(\gamma^*/\mu^*,\alpha^*,\beta^*) = 0.86$

Also the sum of square due to error is same for both the methods and is 52.00.

4.5 The R( ) and R(*) Notation

When we use β-model, the reduction in sum of squares in ANOVA are often described in terms of symbol like $R(\underline{\beta}_2/\underline{\beta}_1)$ which is defined as

$$R(\underline{\beta}_2/\underline{\beta}_1) = R(\underline{\beta}_1,\underline{\beta}_2) - R(\underline{\beta}_1)$$

$$= R(\underline{\beta}) - R(\underline{\beta}_1)$$

Here $R(\underline{\beta}_1,\underline{\beta}_2)$ or $R(\underline{\beta})$ is the regression sum of squares for fitting the model with $\underline{\beta}_1$ and $\underline{\beta}_2$ parameters and $R(\underline{\beta}_1)$ is analogous quantity using only $\underline{\beta}_1$. For example, some of the common sums of squares are computed from the model shown as below:

$$
\begin{aligned}
\mu_{ij} &= \mu \\
&= \mu + \alpha_i && R(\alpha/u) \\
&= \mu + \alpha_i + \beta_j && R(\beta/\mu, \alpha) \\
&= \mu + \alpha_i + \beta_j + \gamma_{ij} && R(\gamma/\mu, \alpha, \beta)
\end{aligned}
$$

Some confusion arises when the R( ) notation is applied to the model after it has been converted to full rank model by imposing certain non-estimable conditions on the parameters. Two essential features to be noted are:

1. Sum of squares such as $R(\beta/\mu^*, \alpha^*, \gamma^*)$ are generally nonzero and correspond to specific hypothesis.

2. The R(*) notation is not unique since it depends on the nonestimable conditions such as $R(\beta^*/\mu^*, \alpha^*, \gamma^*)$ depends on the choice of nonestimable conditions.

CHAPTER - V

LINEAR HYPOTHESES

The main purpose of this chapter is to describe the hypotheses commonly tested in linear models with unbalanced data including the case of zero cell frequency. The sum of the squares for test statistics had been developed either on heuristic principles or because of computational convenience. Precise statements of the corresponding hypothesis associated to the R( ) notation are rarely available in the literature and in those cases where the hypotheses are stated, these are usually described in terms of the parameters of the nonfull rank model which may be difficult to interpret statistically. The hypotheses associated with the reduction in sum of squares or the R( ) notation, are described in terms of means of the observed population in case of two way classification model with interaction. Some examples of the hypotheses tested with missing cells are also described.

5.1 Linear Hypotheses.

The linear hypotheses are discussed when β -model and μ-model are taken separately.

5.1.1 Linear hypothesis in case of β -model

Let us discuss first those hypotheses in case of

classified fixed effects β-model for designed experiments in case of unbalanced data. The linear model is

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}$$

$$i = 1, 2, \ldots, a$$

$$j = 1, 2, \ldots, b$$

$$k = 0, 1, 2, \ldots, n_{ij}$$

Various hypotheses are shown in the following table which are generally tested in terms of population parameters.

Table 7. Showing the hypotheses to be tested!

Effect	Hypothesis Number	Hypothesis in terms of the parameter
		Rows
1. Rows	H_1 :	$\alpha_i + \sum_j \gamma_{ij} = \alpha_{i'} + \sum_j \gamma_{i'j}$
2. Main row	H_2 :	$\alpha_i + \sum_j \frac{n_{ij}\beta_j}{n_{i\cdot}} + \sum_j \frac{n_{ij}\gamma_{ij}}{n_{i\cdot}} = \alpha_{i'} + \sum_j \frac{n_{i'j}\beta_j}{n_{i'\cdot}} + \sum_j \frac{n_{i'j}\gamma_{i'j}}{n_{i'\cdot}}$
3. Partial Row(adj)	H_3 :	$(n_{i\cdot} - \sum_j \frac{n_{ij}^2}{n_{\cdot j}})\alpha_i + \sum_j (n_{ij} - \frac{n_{ij}^2}{n_{\cdot j}})\gamma_{ij} = \sum_{i \neq i'} (\sum_j \frac{n_{ij}n_{i'j}}{n_{\cdot j}})\alpha_{i'} + \sum_{i \neq i'} (\sum_j \frac{n_{ij}n_{i'j}}{n_{\cdot j}})\gamma_{i'j}$
4. First row	H_4 :	$\alpha_1 - \alpha_i + \gamma_{11} - \gamma_{i1} = 0$
		Columns
5. Columns	H_5 :	$\beta_j + \sum_i \gamma_{ij} = \beta_{j'} + \sum_i \gamma_{ij'}$
6. Main columns	H_6 :	$\beta_j + \sum_i \frac{n_{ij}\alpha_i}{n_{\cdot j}} + \sum_i \frac{n_{ij}\gamma_{ij}}{n_{\cdot j}} = \beta_{j'} + \sum_i \frac{n_{ij'}\alpha_i}{n_{\cdot j'}} + \sum_i \frac{n_{ij'}\gamma_{ij'}}{n_{\cdot j'}}$
7. Partial column (adj)	H_7 :	$(n_{\cdot j} - \sum_i n_{ij}^2/n_{i\cdot})\beta_j + \sum_i (n_{ij} - n_{ij}^2/n_{i\cdot})\gamma_{ij} = \sum_{j \neq j'} (\sum_i n_{ij}n_{ij'}/n_{i\cdot})\beta_{j'} + \sum_{j \neq j'} \sum_i (n_{ij}n_{ij'}/n_{i\cdot})\gamma_{ij'}$
8. First column	H_8 :	$\beta_1 - \beta_j + \gamma_{11} - \gamma_{1j} = 0$
		Interaction
9. Interaction	H_9 :	$\gamma_{ij'} - \gamma_{ij} - \gamma_{i'j'} + \gamma_{i'j} = 0$

5.1.2 Linear hypothesis in case of μ-model

The hypotheses are described when μ-model in two way classification with interaction is considered. Following are the various hypotheses in terms of cell means as parameter.

Table 8. Linear hypotheses to be tested in terms of μ.

Effect	Hypothesis No.	Hypothesis in terms of μ
		Rows
1. Rows	H_1	$\bar{\mu}_{i\cdot} = \bar{\mu}_{i'\cdot}$
2. Main row	H_2	$\sum_j \frac{n_{ij}\cdot \mu_{ij}}{n_{i\cdot}} = \sum_j \frac{n_{i'j}\cdot \mu_{i'j}}{n_{i'\cdot}}$
3. Partial row (adj)	H_3	$\sum_j n_{ij}\cdot \mu_{ij} = \sum_{i'} \sum_j \frac{n_{ij}\cdot n_{i'j}\cdot \mu_{i'j}}{n_{\cdot j}}$
4. First row	H_4	$\mu_{i1} = \mu_{i'1}$
		Columns
5. Columns	H_5	$\bar{\mu}_{\cdot j} = \bar{\mu}_{\cdot j'}$
6. Main column	H_6	$\sum_i n_{ij}\cdot \mu_{ij}/n_{\cdot j} = \sum_i n_{ij'}\cdot \mu_{ij'}/n_{\cdot j'}$
7. Partial row	H_7	$\sum_i n_{ij}\mu_{ij} = \sum_{j'} \sum_i \frac{n_{ij}\cdot n_{ij'}\mu_{ij'}}{n_{i\cdot}}$
8. First column	H_8	$\mu_{1j}\cdot = \mu_{j'}$
		Interaction
9. Interaction		$\mu_{ij} - \mu_{ij'} - \mu_{i'j} + \mu_{i'j'} = 0$

All the hypotheses statements hold for all i and j and i' and j' s.t. $i \neq i'$ and $j \neq j'$. The restriction in the hypotheses H_1, H_4, H_5, H_8 and H_9 is that the number of observations in n_{ij} should be greater than zero. If we examine the hypotheses H_1, H_2, H_3 and H_4, it is observed that all are different hypothesis.

5.2 Examples

5.2.1 Two way crossed classification

Consider a model of two way classification without interaction with unequal observation in each cell for the data page-57

$$Y_{ijk} = \mu + \alpha_i + \beta_j + e_{ijk}$$

$$i = 1, 2, \ldots, 3$$

$$j = 1, 2, \ldots, 3$$

$$k = 0, 1$$

Where

Y_{ijk} is the yield of the i^{th} wheat variety under j^{th} fertilizer of k^{th} plot.

α_i is the effect of i^{th} wheat variety.

β_j is the effect of j^{th} fertilizer type and e_{ijk} is an error terms.

Suppose we want to know the relative performance of three varieties of wheat and three types of fertilizers which are usually formulated in terms of linear hypothesis

about the cell means. Following are the different analysis of variance table showing the different reduction in sum of squares.

Table 9. Analysis of variance for two way classification without interaction.

Source of variation	D.F.	S.S.	M.S.	'F'
(a) For fitting μ, and α and β after μ				
Mean	1	$R(\mu)$	-	F(M)
Fitting α and β after μ	a+b-2=4	$R(\alpha, \beta/\mu)$	-	F(Rm)
Residual error	15	S.S.E.	-	
Total	20	S.S.T.		
(b) For fitting μ, α after μ and β after μ and α				
Mean	1	$R(\mu)$	-	F(M)
Fitting α after μ	a-1=2	$R(\alpha/\mu)$	-	$F(\alpha/\mu)H_1$
Fitting β after μ and α	b-1=2	$R(\beta/\mu,\alpha)$	-	$F(\beta/\mu,\alpha)H_3$
Residual	15	S.S.E.	-	
Total	20	S.S.T.		
(c) For fitting μ, β after μ and α after μ and β				
Mean	1	$R(\mu)$	-	F(M)
Fitting β after μ	b-1=2	$R(\beta/\mu)$	-	$F(\beta/\mu)H_5$
Fitting α after μ and β	a-1=2	$R(\alpha/\mu,\beta)$	-	$F(\alpha/\mu,\beta)H_7$
Residual error	15	S.S.E.	-	
Total	20	S.S.T.		

From the Table 9, it is clear that $R(\alpha/\mu)$ and $R(\alpha/\mu, \beta)$ are not used for the same purpose. These R( ) notation are associated with the different hypotheses. Similarly $R(\beta/\mu)$ and $R(\beta/\mu,\alpha)$ test the different hypotheses. Distinguishing between these two $F(\alpha/\mu)$ and $F(\alpha/\mu,\beta)$ is of paramount importance because it is a distinction that occurs repeatedly in fitting other models. These two F's may be identical in case of balanced data but it differs in case of unbalanced data. That is the test based on $F(\alpha/\mu)$ is testing the effectiveness of adding α effects to the model over and above μ whereas $F(\alpha/\mu, \beta)$ tests the effectiveness of adding α -effects to the model over and above having μ and β -effects. These tests must be described more completely.

i) Now suppose we want to compare three fertilizers. The hypothesis that the mean response of three fertilizers when averaged over all varieties is same that is

$$HF : \bar{\mu}.1 = \bar{\mu}.2 = \bar{\mu}.3$$

Similarly for three varieties, the hypothesis is that the mean response of the three varieties when averaged over all fertilizers is same.

$$H_V : \bar{\mu}_{1.} = \bar{\mu}_{2.} = \bar{\mu}_{3.}$$

These hypotheses correspond to the hypotheses to H_1 and H_5

respectively and their sum of squares are $R(\beta/\mu)$ and $R(\alpha/\mu)$ respectively.

Here H_5 is related with average performance of the three fertilizers. But if any farmer is interested to grow all the three varieties of wheat and wants to purchase only one type of fertilizer then this hypothesis HF would be appropriate. These hypothesis correspond to H_4. Alternatively if he is willing to purchase different fertilizers for different varieties, then the hypothesis may be of the form $= H_F^* : \mu_{ij} - \mu_{ij'} = 0$ which is the differential response of fertilizers j and j' on a specific variety i, is zero. This hypothesis H_8 is synonymus to hypothesis H_4.

It would be nice if the farmer can decide that a particular fertilizer is best for all the three varieties rather than just best for a particular variety as in H_F or best on the average as in H_F^*. Suppose we assume

$$\mu_{ij} - \mu_{ij}' - \mu_{ij'} + \mu_{ij}'' = 0$$

for all i, i', j and j' that is the differential effect of the fertilizer j and j' is same for varieties i and i', which is synonymous to H_9. If this assumption is valid, this hypothesis H_F is equivalent to the much stronger

hypothesis

$$H_F^{**} : \mu_{ij} - \mu_{ij'} = 0 \text{ for all } i, j \text{ and } j'$$

Example 2. Two fold nested model

Consider a two-fold nested design.

$$Y_{ijk} = \mu_{ij} + e_{ijk}$$

$$i = 1,2, \ldots ,a$$

$$j = 1,2, \ldots ,b_i$$

$$k = 1,2, \ldots ,n_{ij}$$

Suppose a Govt. Officer wants to study the fish capture in various districts in some state where there are different number of ponds in each districts. Under each ponds different number of fisherman catch the fish. For example, he collects the data on a = 3 districts and b_i = 2,2,3 ponds nested in the i^{th} districts. He interviews n_{ij} fishman at the j^{th} pond in the i^{th} district as shown below:

Table 10. Data on fish capture in different districts.

		Districts						
		D_1		D_2		D_3		
Ponds		P_1	P_2	P_1	P_2	P_1	P_2	P_3
Fisherman	1	6	5	4	2	3	2	4
	2	3	7	3	5	3	3	5
	3	5	6	-	3	5	4	-
	4	-	8	-	4	-	6	-
	5	-	-	-	1	-	5	-
	n_{ij}	3	4	2	5	3	5	2

Here μ_{ij} is the mean response (i.e., average number of fish caught per day) at the j^{th} pond in the i^{th} district. The same can be expressed in the form of β-model

$$Y_{ijk} = \mu + \alpha_i + \gamma_{ij} + e_{ijk}$$

Where

μ is the overall mean

α_i is the effect of i^{th} district and

γ_{ij} is effect of j^{th} pond in the i^{th} district.

The analysis of the β-model is summarized in the following ANOVA table i...

Table II.

Source	d.f.	S.S.	Hypothesis
Bet.Districts	2	$R(\alpha/\mu)$	HA(H_2 and H_6)
Bet.Ponds within districts	4	$R(\gamma/\mu;\alpha)$	HB/A (H_4 and H_8)
Error	17	SSE.	

The sum of squares $R(\gamma/\mu;\alpha)$ is the difference in the residual sum of squares obtained by fitting the model

$$Y_{ijk} = \mu + \alpha_i + e_{ijk}$$

and SSE_{Ω} the full rank model residual sum of squares.

But the analysis of the μ-model forces us to specify the hypotheses of our interest as opposed to β-model. The question, of which hypothesis to test, depends on the interest of the research worker but two hypothesis can be considered.

$$H_1 : \bar{\mu}_{i.} = \bar{\mu}_{i'.} \text{ for } i \text{ and } i'$$

and

$$H_2 : \mu_{ij} = \mu_{ij'} \text{ for all } i, j, j'$$

Here H_1 demonstrates that the average catch per pond is the same in each of the district i.e.

$$\bar{\mu}_{1:} = \bar{\mu}_{2.} = \bar{\mu}_{3.} \text{ and}$$

H_2 says that the average catch in each pond is the same in a given district i.e.

$$\mu_{11} = \mu_{12}$$

$$\mu_{21} = \mu_{22}$$

$$\mu_{31} = \mu_{32} = \mu_{33}$$

It seems that H_2 hypothesis is identical to $H_{B/A}$ but H_1 is different from H_A, unless n_{ij} is the same for each pond. In the μ-model H_A correspond to the hypothesis H_2 which is equivalent to here

$$\frac{3\mu_{11} + 4\mu_{12}}{7} = \frac{2\mu_{21} + 5\mu_{22}}{7} = \frac{3\mu_{31} + 5\mu_{32} + 2\mu_{33}}{10}$$

and H_1 is equal to

$$\frac{\mu_{11} + \mu_{12}}{2} = \frac{\mu_{21} + \mu_{22}}{2} = \frac{\mu_{31} + \mu_{32} + \mu_{33}}{3}$$

It seems that we cannot conclude which hypothesis is more appropriate, but it is clear that H_1 and H_2 hypothesis are different. The advantage of the cell mean model is that the user can clearly evaluate which of the hypothesis is of interest to him. The μ-model for this nested situation is identical if $b_i = b$ to that in two way classification (nested)

5.4 Missing Cells Case

Earlier we have assumed that all $n_{ij} > 0$ which is not essential for all methods. For example, the sums of squares generated in the method of fitting constants requires only that design be connected. The hypothesis indicated are valid with $n_{ij} = 0$ inserted where no observations are made on the $(ij)^{th}$ cell. However, the hypothesis associated with the method of weighted squares of means are valid only when $n_{ij} > 0$. Several computer programmes employ the modification of the basic R( ) procedure which sets $\gamma_{ij} = 0$ if $n_{ij} = 0$ which enables the generation of sums of squares but no indication is given which hypothesis being tested.

5.4.1 Example

Let a = b = 3 and suppose $n_{ii} = 0$; i = 1,2,3. The remaining cell frequencies are as below:

	Column		
	1	2	3
Rows			
1	0	1	2
2	1	0	1
3	2	1	0

Here the method of fitting constants will give the sum of squares for main effects and interaction. The following table gives the number of hypothesis:

Table 12. Hypothesis for missing cells analysis

Hypothesis	Rows Effects
H_1 :	$\mu_{12} + \mu_{13} = \mu_{23} + \mu_{32}$
	$\mu_{21} + \mu_{23} = \mu_{13} + \mu_{31}$
H_2 :	$\frac{\mu_{12} + 2\mu_{13}}{3} = \frac{\mu_{21} + \mu_{23}}{2} = \frac{2\mu_{31} + \mu_{32}}{3}$
H_3 :	$\frac{\mu_{12} - \mu_{32}}{2} = \frac{2(\mu_{23} - \mu_{13})}{3}, \mu_{21} + \mu_{23} = \mu_{13} + \mu_{31}$
H_4 :	$\mu_{13} = \mu_{23}$
	$\mu_{21} = \mu_{31}$

Columns

H_5 : $\mu_{21} + \mu_{31} = \mu_{23} + \mu_{32}$

$\mu_{12} + \mu_{32} = \mu_{31} + \mu_{13}$

H_6 : $\frac{\mu_{21} + 2\mu_{31}}{3} = \frac{\mu_{12} + \mu_{32}}{2} = \frac{2\mu_{13} + \mu_{23}}{3}$

H_7 : $\frac{\mu_{21} - \mu_{23}}{2} = \frac{2(\mu_{32} - \mu_{31})}{3}$

$\mu_{12} + \mu_{32} = \mu_{31} + \mu_{13}$

H_8 : $\mu_{21} = \mu_{23}$

$\mu_{12} = \mu_{13}$

Interaction

H_9 : $\mu_{12} - \mu_{13} - \mu_{21} = \mu_{32} - \mu_{23} - \mu_{31}$

CHAPTER - VI

METHODS OF ANALYSIS OF LINEAR MODELS

In this chapter various methods of analysis of experimental design models with unbalanced data are described which are most commonly found in the literature. These methods are differentiated by the hypotheses associated with the sum of squares or R( ) notation. The method should be chosen based on the appropriateness of the hypotheses rather than on computational convenience. These sums of squares are calculated from the full rank cell means μ-model which will remove many sources of prevailing confusion, since the hypothesis can be stated in terms of the population parameter. The following are the methods for the analysis of linear models.

6.1 The Methods of Unweighted Means

The method of unweighted means was described by Yates (1934) as an approximate method which is computationally very simple method. Later on Anderson and Bancroft (1952), Senedecor and Cochran (1967), Searle (1971) and Winer (1971) have discussed it. The procedure was applied in two way classification model without interaction with x_{ij} denoting the cell means. The error sum of squares (SSE) is just the usual residual sum of squares.

Table 13. ANOVA Table of Method of Unweighted Means.

Source	S.S.	Hypothesis which can be tested
A factor	$\bar{n}_h \sum_i \sum_j (\bar{x}_{i\cdot} - \bar{x}_{\cdot\cdot})^2$	None
B factor	$\bar{n}_h \sum_i \sum_j (\bar{x}_{\cdot j} - \bar{x}_{\cdot\cdot})^2$	None
Interaction A x B	$\bar{n}_h \sum_i \sum_j (x_{ij} - \bar{x}_{i\cdot} - \bar{x}_{\cdot j} + \bar{x}_{\cdot\cdot})^2$	None
Error	S.S.E.	

Where $\bar{n}_h = ab \Big/ \sum\sum \frac{1}{n_{ij}}$ $\qquad x_{ij} = \bar{Y}_{ij}.$

This method is approximate since the sum of square due to A, B and AB are not distributed as χ^2 and hence do not correspond to the numerator sum of squares (due to A, B, AB) for testing linear hypothesis about the cell means.

6.2 The Method of Weighted Squares of Means

This method was also described by Yates (1934) and afterwards by Anderson and Bancroft (1952), Elston and Bush (1964), Searle (1971) and Neter and Wasserman (1974) in two way classification with unequal number of observations in the cells but atleast one observation per cell. Amend the normal equations by requiring the

solutions to satisfy.

$$\sum_i \alpha_i = \sum_j \beta_j = \sum_i \gamma_{ij} = \sum_j \gamma_{ij} = 0$$

for all i, and j

The same method can be used for solving the equations in case of balanced data and same has been extended for unbalanced data. Here the sum of squares for α-effects in the weighted squares of means is calculated as

$$SSA_W = \sum_{i=1}^{a} \frac{\left[\sum_j \bar{\gamma}_{ij.}\right]^2}{\sum_j \frac{1}{n_{ij}}} - \left[\sum_i \frac{\sum_j \bar{\gamma}_{ij.}}{\sum_j \frac{1}{n_{ij}}}\right]^2 \Big/ \sum_{i=1}^{a} \frac{1}{\sum_j \frac{1}{n_{ij}}}$$

then $SSA_W = R\ (\alpha^*/\mu^*, \beta^*, \gamma^*)$

Table 14. ANOVA Table

Source	S.S.	Hypothesis
Rows A	$R(\alpha^*/\mu^*, \beta^*, \gamma^*) = SSAW = \sum_i \omega_i (\bar{x}_{i.} - \bar{x}_{w})^2$	H_1
Column B	$R(\beta^*/\mu^*, \alpha^*, \gamma^*) = SSBW = \sum_j v_j (\bar{x}_{.j} - \bar{x}_{v})^2$	H_5
Interaction AB	$R(\gamma^*/\mu^*, \alpha^*, \beta^*) = SSAB = \sum_{ij} (x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x}_{..})^2$	H_9

Where

$$\bar{x}_{(1)} = \frac{\sum_i \omega_i \bar{x}_{i.}}{\sum \omega_i} \qquad \bar{x}_{(2)} = \frac{\sum_j v_j \bar{x}_{.j}}{\sum_j v_j}$$

$$w_i = b^2 \Big/ \sum_j \frac{1}{n_{ij}} \qquad v_j = a^2 \Big/ \sum_i \frac{1}{n_{ij}}$$

$$x_{ij} = \bar{Y}_{ij.}$$

In this analysis, the hypothesis associated with Rows A and Columns B of weighted means are unweighted hypothesis H_1 and H_5. The interaction SS AB exists only when there are at least two levels of each factor. In this case $R(\gamma^*/\mu^*, \alpha^*, \beta^*)$ is independent of the procedure of the nonestimable conditions imposed. The sum of squares SSA can be obtained from $R(\alpha^*/\mu^*, \beta^*, \gamma^*)$ if we use nonestimable restrictions.

$$\alpha_{i\cdot} = \beta_{\cdot j} = \gamma_{i\cdot} = \gamma_{\cdot j} = 0 \text{ for all i and j}$$

Searle (1971) had linked this discussion of the hypothesis being tested by S.S.A. to these nonestimable conditions. But this is misleading as H_1 hypothesis is regardless of how the computations are performed. The R( ) notation procedure is just a computational mechanism.

6.3 The Method of Fitting Constants

This method of fitting constants was also discussed by Yates (1934). Here the sums of squares are most easily described by using the R( ) notation in two way classification model with interaction. Here $R(\alpha/\mu,\beta)$ is computed by applying the R( ) procedure to the two way classification model without interaction and $R(\alpha/\mu)$ is computed from one way classification reduced model

$$Y_{ijk} = \mu + \alpha_i + e_{ijk}.$$

Table 15. ANOVA Table for the Method of Fitting Constants

Source	S.S.	Hypothesis
A (unadj)	$R(\alpha/\mu)$	H_2
A (adj)	$R(\alpha/\mu,\beta)$	H_3
B (unadj)	$R(\beta/\mu)$	H_6
B (adj)	$R(\beta/\mu,\alpha)$	H_7
AB	$R(\gamma/\mu, \alpha, \beta)$	H_9

In the above table, it is observed that there are two "maineffects" sum of squares for both A and B. This type of description of sums of squares is not informative with regard to identification of the hypothesis being tested. $R(\alpha/\mu)$ is referred as the sum of squares for "Testing α effects ignoring β and γ" by some authors. Precise statements of hypotheses being tested in terms of the cell means are provided by H_2 and H_3.

6.4 The Methods of Overall and Spiegel (1969)

The authors presented three different methods for analysing the β-model and μ-model which are appropriate under different circumstances.

6.4.1 Complete least squares :

The method is precisely like the method of weighted squares of means with the same sum of squares and corresponding hypotheses H_1, H_5 and H_9.

6.4.2 Experimental design method :

This method is also just like the method of fitting constants in which the main effects adjusted sum of square $R(\alpha/\mu,\beta)$ and $R(\beta/\mu,\alpha)$ are used corresponding to the hypothesis H_3 and H_7 being tested.

6.4.3 A prior ordering :

This is also a special case of method of fitting constants in which we use $R(\alpha/\mu)$ and $R(\beta/\mu,\alpha)$ as the main effects sum of squares for testing the hypothesis H_2 and H_7 or $R(\beta/\mu)$ and $R(\alpha/\mu,\beta)$ are used for testing hypothesis H_6 and H_3. The authors further suggested that second method is the proper generalization of the analysis for balanced data. Since all the three methods are identical with balanced data, Method 1 complete least squares is the natural generalization. The hypothesis H_1 and H_5 are easy to interpret. They also suggested that Method 3 'A prior ordering' is better appropriation if a logical ordering exists for effects being examined.

6.5 Henderson's Methods

Henderson (1953) proposed three methods for generalizing sums of squares for the estimation of variance covariance components. Since the sum of squares have been used for testing hypothesis in fixed effects model, it is of interest to relate them with present development. Only two methods 1 and 3 are considered here.

6.5.1 Method 1 :

Here unadjusted main effects (unadj) sum of squares $R(\alpha/\mu)$ and $R(\beta/\mu)$ are considered as in the method of fitting constants which test the hypotheses H_2 and H_6. Searle (1971) observed that interaction sum of squares obtained by substraction of both main effects is not a sum of squares and hence it does not correspond to any linear hypothesis or it can not be expressed as the linear function of population parameter.

6.5.2 Method 3 :

Here this method can be used with any ANOVA table, it is generally associated with the sum of squares suggested by Overall and Spiegel's (1969)'s method,"a prior ordering" which tests the hypothesis H_2 and H_7 or alternatively H_6 and H_3.

6.6 Hemmele's Iterative Method

This method proposed by Hemmerle (1974) is a procedure for computing $R(\beta_2/\beta_1)$ of which computations are performed iteratively. Since the sum of squares are calculated iteratively, the quantities calculated are approximations to the actual sum of squares. The specific hypotheses being tested depend on the partition of the parameter $\underset{\sim}{\beta}$ into $\underset{\sim}{\beta}_1$ and $\underset{\sim}{\beta}_2$ as well as on the choice of the non-estimable conditions. The hypotheses being tested with this method given by Hemmerle are H_1, H_5 and H_9. The nonestimable conditions are same as in 6.2 method.

6.7 The Methods used in SAS-76

Four methods have been incorporated into SAS-76 system which form part of the General Linear Model (GLM) procedure given by Barr *et al* (1976). These methods are summarized as follows.

6.7.1 SAS-76 Type 1 method

This method resembles with the method of fitting constants and hence H_2 and H_7 main effects hypotheses are tested. Thus it is noted that it is"a prior ordering method" of Overall and Spiegel (1969).

6.7.2 SAS-76 Type II method

This is also most common procedure based on the method of fitting constants in which the adjusted sum of squares viz. $R(\alpha/\mu, \beta)$ and $R(\beta/\mu, \alpha)$ are used to test the main effects hypotheses. This is the experimental design method of Overall and Spiegel (1969) in which H_3 and H_7 hypotheses are being tested.

6.7.3 SAS-76 Type III and IV method

These two methods are equivalent to the method of weighted squares of means or complete least squares method of Overall and Spiegel (1969) where main effects hypotheses H_1 and H_5 can be tested.

The following table provides a summary of the hypotheses tested by the various methods.

Table 16. Hypothesis tested by different methodswith asterisk.

Methods	Row effects				Column effects				Int.
	H_1	H_2	H_3	H_4	H_5	H_6	H_7	H_8	H_9
1.Unweighted means	No hypothesis being tested.								
2.Weighted squares of means	*	-	-	-	*	-	-	-	*
3.Fitting constants	-	*	*	-	-	*	*	-	*
4.Overall & Spiegel's									
Method 1	*	-	-	-	*	-	-	-	*
Method II	-	-	*	-	-	-	*	-	*
Method III	-	*	-	-	-	-	*	-	*
5.Henderson's									
Method I	-	*	-	-	-	*	-	-	-
Method III	-	-	*	-	-	-	*	-	*
6.Hammerle's iterative method	*	-	-	-	*	-	-	-	*
7.SAS-76									
Type I	-	*	-	-	-	-	*	-	*
Type II	-	-	*	-	-	-	*	-	*
Type III & IV	*	-	-	-	*	-	-	-	*

CHAPTER - VII

REGRESSION MODEL WITH DUMMY VARIABLES

In this chapter we shall discuss the regression model with dummy variables. The regression model with dummy variables is an alternative for analysis of variance technique for balanced sets, though it is not usually recommended. The concept of regression on dummy variables is introduced for some factors and their levels. For unbalanced data, having more than one factor in a cross classification situation, the usual analysis of variance technique has several limitations and the regression model is the most viable tool for analysing the two or higher way cross classified data with unequal cell frequencies. The preferred analysis of unbalanced data is the regression analysis with dummy variables.

7.1 Dummy Variables

There are many ways of quantitatively identifying the classes of a qualitative variable. A dummy variable is any variable in a regression equation that takes on a finite number of values for identifying different categories of qualitative variable. The term "dummy" variable describes no meaningful measurement level of the variable but rather

act only to indicate the categories of the nominal variable. For example a variable X_1 takes the value 1 if some level of a factor is used and zero otherwise or if alternatively a variable X_2 takes the values in the following way

$$X_2 = \begin{cases} 1 & \text{If an individual is a male} \\ -1 & \text{If an individual is female} \\ 0 & \text{Otherwise.} \end{cases}$$

then the variables so defined act as indicator variables in the sense usually associated with them and are called "dummy variables".

If the nominal independent variable has k categories, we define (k-1) number of dummy variables to index the categories, provided, the regression model contains a constant term (an intercept). If the regression model does not contain a constant term, then k dummy variables are defined to identify the k categories.

Due to the use of dummy variables in the regression analysis, the regression analysis has more range of application in the statistical field. In particular we can get the same information by employing the regression analysis with dummy variables what we obtain by use of distinct analytical procedures such as analysis of variance, analysis of covariance and discriminant analysis etc. Also we can

compare many regression equation with the use of dummy variables in a single multiple regression model.

7.2 Linear Model

Consider a two way classification model with interaction

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk} \quad \ldots \quad (7.2.1)$$

Here in equation (7.2.1) none of the parameter is estimable. The hypothesis such as HO : $\alpha_i = 0$ for all i are also not testable. If nothing more is stated about these parameters, thus they are not uniquely defined but it is nevertheless possible to estimate certain linear functions of them. Suppose we put the additional restrictions such as

$$\sum_{i=1}^{a} \nu_i \alpha_i = \sum_{j=1}^{b} \omega_j \beta_j = 0 \quad \ldots \quad (7.2.2)$$

and

$$\sum_{i=1}^{a} \nu_i \gamma_{ij} = 0 \text{ for all } j, \quad \sum_{j=1}^{b} \omega_j \gamma_{ij} = 0 \text{ for all } i \quad \ldots \quad (7.2.3)$$

where ν_i and ω_j are (a+b) given positive quantities and at least one ν_i and ω_j are non zero. If we consider (7.2.1) along with (7.2.2) and (7.2.3) thus all the parameters in (7.2.1) are uniquely defined in terms of cell means and also the parameters became estimable as well. It is important to note that none of the above restrictions

in (7.2.2) and (7.2.3) are estimable. We shall consider three particular cases of the above restrictions which are generally of interest.

1. $\nu_i = \omega_j = 1$ for all i, j

i.e. $\sum_{i=1}^{a} \alpha_i = 0 = \sum_{j=1}^{b} \beta_j$ . . . (7.2.4)

and

$\sum_{i} \gamma_{ij} = 0$ for all j

$\sum_{j} \gamma_{ij} = 0$ for all i

2. $\nu_1 = \omega_1 = 1$ and all other ν_i's and ω_j's are zero.

i.e. $\alpha_1 = \beta_1 = \gamma_{1j} = \gamma_{i1} = 0$ for all i & j . . . (7.2.5)

3. $\nu_i = \sum_{j=1}^{b} n_{ij} = n_{i.}$ for all i

$\omega_j = \sum_{i=1}^{a} n_{ij} = n_{.j}$ for all j

i.e. $\sum_{i=1}^{a} n_{i.} \alpha_i = \sum_{j=1}^{b} n_{.j} \beta_j = 0$ . . . (7.2.6)

$\sum_{i}^{a} n_{i.} \gamma_{ij} = 0$ for all j $\sum_{j=1}^{b} n_{.j} \gamma_{ij} = 0$ for all i.

Definitions of the parameters in the model (7.2.1) depend on the restrictions considered along with the model. If the restrictions given in (7.2.4) are considered part of the model (7.2.1), then μ is a general mean, α_i and β_j are main effects and γ_{ij} is an interaction effect as are usually defined for a balanced ANOVA model. If $n_{ij} = m$

for all i and j, it becomes the balanced case. All other n_{ij} configuration are collectively called unbalanced data. A special case of unbalanced data is when the cell frequencies are proportional.

7.3 Regression Model With Dummy Variables

Most linear models can be alternatively considered in a regression model setting. This can be done by defining dummy variables in a regression model appropriately. A regression formulation often is desirable, if not mandatory when dealing with unbalanced data involving two or more factors. The general regression formulation with dummy variables corresponding the model (7.2.1) is as follows:

$$Y = \mu + \sum_{i=1}^{a-1} \alpha_i x_i + \sum_{j=1}^{b-1} \beta_j z_j + \sum_{i=1}^{a-1} \sum_{j=1}^{b-1} \gamma_{ij} x_i z_j + e \quad \ldots.(7.3.1)$$

Where μ, α_i, β_j and γ_{ij} are regression coefficients and x_i and z_j are appropriately defined dummy variables. In section (7.2) we have considered three particular cases of the restrictions under (7.2.4), (7.2.5) and (7.2.6) to be considered along with the model (7.2.1). In regression model, we will define three coding schemes for the dummy variables in the general regression model (7.3.1) which corresponds to the three particular cases of restrictions in (7.2.2)and (7.2.3). The following are the three schemes:

Scheme A: Define

$$x_i = \begin{cases} 1 \text{ for level } i \text{ of the row factor} \\ -1 \text{ for level } a \text{ of the row factor} \\ 0 \text{ otherwise} \end{cases}$$

$$i = 1, 2, \ldots, a-1 \qquad (7.3.2)$$

and

$$z_j = \begin{cases} 1 \text{ for level } j \text{ of the column factor} \\ -1 \text{ for level } b \text{ of the column factor} \\ 0 \text{ otherwise} \end{cases}$$

$$j = 1, 2, \ldots, b-1$$

This coding scheme corresponds to the restriction in (7.2.4).

Scheme B: Define

$$x_i = \begin{cases} 1 \quad \text{For level } (i+1) \text{ of the row factor} \\ 0 \quad \text{otherwise} \end{cases}$$

$$i = 1, 2, \ldots, a-1 \qquad (7.3.3)$$

and

$$z_j = \begin{cases} 1 \text{ for level } (j+1) \text{ of the column factor} \\ 0 \text{ otherwise} \end{cases}$$

and $j = 1, 2, \ldots, b-1$

This coding scheme corresponds to the restrictions given in (7.2.5).

Scheme C: Define

$$x_i = \begin{cases} n_{a:} & \text{for level i of the row factor} \\ -n_{i:} & \text{for level a of the row factor} \\ 0 & \text{otherwise} \end{cases}$$

$$i = 1, 2, \ldots, a-1 \qquad (7.3.4)$$

and

$$z_j = \begin{cases} n_{.b} & \text{for level j of the column factor} \\ -n_{.j} & \text{for level b of the column factor} \\ 0 & \text{otherwise} \end{cases}$$

$$j = 1, 2, \ldots, b-1$$

This coding scheme corresponds to the restrictions in (7.2.6).

7.4 Common ANOVA Hypothesis.

The general regression model (7.3.1) is applicable to both equal and unequal cell frequencies cases. We will express the equivalance of hypothesis tested by these approaches in terms of the cell means and will make specific recommendations in regard to their uses to present a unified approach. Also the reasons for their disagreement for certain types of hypothesis denoted by R( ) will be presented

to clarify the prevailing confusion in the analysis of linear models with unbalanced data. Table 17 gives the various common hypotheses which are related to different coding scheme.

It is seen from Table 17 that $R^*(\alpha/\mu, \beta, \gamma)$, under coding Scheme A, tests the hypothesis H_1 and likewise $R^*(\beta/\mu, \alpha, \gamma)$ tests the hypothesis H_5 in two way classification with interaction. By inspecting these hypothesis it is revealed that the sense in which row or column affect measured is different for the four hypothesis. It is better ~~to prefer~~ to call the hypothesis H_3 as hypothesis for testing the partial row effect in two way classification model. Since it is consistent with the definition of the partial regression coefficient because we are using here general regression model as given in (7.3.1) where the columns of the design matrix for the two effects are not mutually orthogonal in case of unbalanced data. Here the regression coefficients are measuring the partial effects only. The hypothesis H_1, H_2 and H_3 agree i.e. they are equivalent only when $n_{ij} = n$ for all i and j and not otherwise. The hypothesis H_4 can only be tested when coding scheme B $B(1, 0)$ is adopted. The hypothesis H_1 is simplest for interpretation which is that there is no difference in the levels of row factor A when averaged overall levels of

Table 17. Various hypotheses in regression model with dummy variables.

Effect	Regression model	Hypothesis	Test procedure coding scheme		
			Scheme A	Scheme B	Scheme C
Row	$Y = \mu + \sum_i \alpha_i x_i + \sum \beta_j z_j + \sum \gamma_{ij} x_i z_j + e$	$H_1: \bar{\mu}_i = \bar{\mu}_{i'}$	$R^*(\alpha/\mu,\beta,\gamma)$	-	-
Main row	$Y = \mu + \sum \alpha_i x_i + e$	$H_2: \frac{\sum n_{ij}\mu_{ij}}{n_{i.}} = \frac{\sum n_{i'j}\mu_{i'j}}{n_{i'.}}$	$R^*(\alpha/\mu)$	$R^*(\alpha/\mu)$	$R^*(\alpha/\mu)$
Partial row (Adj)	$Y = \mu + \sum \alpha_i x_i + \sum \beta_j z_j + e$	$H_3: \sum_j n_{ij}\mu_{ij} = \sum_{i'}\sum_j \frac{n_{ij}n_{i'j}\mu_{i'j}}{n_{.j}}$	$R^*(\alpha/\mu,\beta)$	$R^*(\alpha/\mu,\beta)$	$R^*(\alpha/\mu,\beta)$
First row	$Y = \mu + \sum \alpha_i x_i + \sum \beta_j z_j + \sum \gamma_{ij} x_i z_j + e$	$H_4: \mu_{i1} = \mu_{i'1}$	-	$\overset{*}{R}(\alpha/\mu,\beta,\gamma)$	-
Column	$Y = \mu + \sum \alpha_i x_i + \sum \beta_j z_j + \sum \gamma_{ij} x_i z_j + e$	$H_5: \bar{\mu}_{.j} = \bar{\mu}_{.j'}$	$R^*(\beta/\mu,\alpha,\gamma)$	-	-
Main column	$Y = \mu + \sum \beta_j z_j + e$	$H_6: \sum_i \frac{n_{ij}u_{ij}}{n_{.j}} = \sum_i \frac{n_{ij'}\mu_{ij'}}{n_{.j'}}$	$R^*(\beta/\mu)$	$\overset{*}{R}(\beta/\mu)$	$\overset{*}{R}(\beta/\mu)$
Partial column (Adj)	$Y = \mu + \sum \alpha_i x_i + \sum \beta_j z_j + e$	$H_7: \sum_i n_{ij}\mu_{ij} = \sum_{j'}\sum_i \frac{n_{ij}n_{ij'}\mu_{ij'}}{n_{i.}}$	$R^*(\beta/\mu,\alpha)$	$R^*(\beta/\mu,\alpha)$	$R^*(\beta/\mu,\alpha)$
First column	$Y = \mu + \sum \alpha_i x_i + \sum \beta_j z_j + \sum \gamma_{ij} x_i z_j + e$	$H_8: \mu_{1j} = \mu_{1j'}$		$R^*(\beta/\mu,\alpha,\gamma)$	-
Interaction		$H_9: \mu_{ij} - \mu_{ij'} - \mu_{i'j} + \mu_{i'j'} = 0$	$\overset{*}{R}(\gamma/\mu,\alpha,\beta)$	$\overset{*}{R}(\gamma/\mu,\alpha,\beta)$	$\overset{*}{R}(\gamma/\mu,\alpha,\beta)$

All statements hold good for i, i', j, j' s.t, $i \neq i'$ and $j \neq j'$.

column factor B. Similarly H_5 hypothesis tests that there is no difference in the levels of column factor B when averaged over all the levels of row factor A. $R^*(\alpha/\mu, \beta, \gamma)$ and $R^*(\beta/\mu, \alpha, \gamma)$ are the corresponding sums of squares for α effects and β effects in the method of analysis of weighted squares of means suggested by Yates (1934). In addition, it is noted that there is another method to obtain these sums of squares. The hypothesis H_4 which test the cell means between rows under specified column, is distinctly different from others and easy to test and interpret. There are specific situations in applications when hypothesis H_4 is of special importance. The hypothesis H_5, H_6, H_7 and H_8 are the counterparts of H_1, H_2, H_3 and H_4 respectively. If H_1, H_2, H_3, H_4 are used for rows then H_5, H_6, H_7 and H_8 are the hypothesis for columns respectively. The last hypothesis H_9, which is the hypothesis of no interaction between the levels of two factor A and B, is common to all three coding schemes under the assumption that $n_{ij} > 0$.

7.5 Examples

In order to illustrate the procedure for obtaining the various sum of squares required for the different entries in the analysis of variance using coding scheme A (1, -1, 0) and coding scheme B (1, 0), we consider the

same experimental data discussed earlier under method 1 in chapter IV.

The appropriate regression model for this experimental data is

$$Y = \mu + \alpha_1 x_1 + \beta_1 z_1 + \beta_2 z_2 + \gamma_{11} x_1 z_1 + \gamma_{12} x_1 z_2 + e \quad \ldots(7.5.1)$$

where μ, α_1, β_1, β_2, γ_{11} and γ_{12} are regression coefficients and x_1, z_1 and z_2 are dummy variables defined as in section (7.3).

7.5.1 Coding Scheme A (1, -1, 0)

x_1 = 1 for level 1 of the row factor

and

$$z_j = \begin{cases} 1 & \text{for the level } j \text{ of the column factor} \\ 0 & \text{otherwise} \end{cases}$$

$j = 1, 2.$

The design matrix $\underset{a}{X}^{*\prime}$ for the regression model (7.5.1), turns out to be

$$\underset{a}{X}^{*\prime} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 \\ 1 & 1 & 1 & 0 & -1 & -1 & 1 & 1 & 1 & 0 & 0 & -1 & -1 & -1 & -1 \\ 0 & 0 & 0 & 1 & -1 & -1 & 0 & 0 & 0 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & 1 & 1 & 0 & -1 & -1 & -1 & -1 & -1 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & -1 & -1 & 0 & 0 & 0 & -1 & -1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

and $\underset{(1\times15)}{Y'} = (8,13,9,12,7,11,11,14,17,14,16,10,11,14,13)$

And the normal equations $\underline{X}_a^{*\prime}\underline{X}_a^{*}\ \hat{\underline{\beta}}_a^{*} = \underline{X}_a^{*\prime}\underline{Y}$ are

$$\begin{bmatrix} 15 & -3 & 0 & -3 & 2 & 1 \\ -3 & 15 & 2 & 1 & 0 & -3 \\ 0 & 2 & 12 & 6 & -2 & -2 \\ -3 & 1 & 6 & 9 & -2 & -3 \\ 2 & 0 & -2 & -2 & 12 & 6 \\ 1 & -3 & -2 & -3 & 6 & 9 \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{\alpha}_1 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\gamma}_{11} \\ \hat{\gamma}_{12} \end{bmatrix} = \begin{bmatrix} 180 \\ -60 \\ 6 \\ -24 \\ 18 \\ 12 \end{bmatrix}$$

Following the usual analysis procedure, we obtain

$R^*(\mu, \alpha, \beta, \gamma) = 2220.00.$

The sum of squares for reduced models can be obtained by using the appropriate portion of $\underline{X}_a^{*\prime}\underline{X}_a^{*}$ and $\underline{X}_a^{\prime}\underline{Y}$ matrices for the complete model. This is achieved by setting each of those parameters equal to zero which are not present in the reduced model. For example to evaluate $R(\mu)$, we set

$$\alpha_1 = \beta_1 = \beta_2 = \gamma_{11} = \gamma_{12} = 0$$

and the reduced normal equations for the model having parameter μ is $15\hat{\mu} = 180$ which gives $R^*(\mu) = \frac{180^2}{15} = 2160.00$

And to evaluate $R^*(\mu, \beta, \gamma)$, we set $\alpha_1 = 0$ in the model and corresponding the reduced normal equation are

$$\begin{bmatrix} 15 & 0 & -3 & 2 & 1 \\ 0 & 12 & 6 & -2 & -2 \\ -3 & 6 & 9 & -2 & -3 \\ 2 & -2 & -2 & 12 & 6 \\ 1 & -2 & -3 & 6 & 9 \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\gamma}_{11} \\ \hat{\gamma}_{12} \end{bmatrix} = \begin{bmatrix} 180 \\ 6 \\ -24 \\ 18 \\ 12 \end{bmatrix}$$

and $R^*(\mu, \beta, \gamma)$ comes out to be 2185.71. The reduction in sum of squares due to fitting various sub models is presented in the following table.

Table 18. ANOVA Table.

Source	d.f.	S.S.
$R^*(\mu)$	1	2160.00
$R^*(\alpha/\mu)$	1	40.00
$R^*(\beta/\mu)$	2	18.00
$R^*(\alpha, \beta/\mu)$	3	59.14
$R^*(\alpha, \gamma/\mu)$	3	43.14
$R^*(\beta, \gamma/\mu)$	4	25.71
$R^*(\alpha, \beta, \gamma/\mu)$	5	60.00
$R^*(\beta/\mu, \alpha)$	2	19.14
$R^*(\alpha/\mu, \beta)$	1	41.14
$R^*(\gamma/\mu, \alpha, \beta)$	2	0.86
$R^*(\mu, \alpha)$	1	2200.00
$R^*(\mu, \beta)$	2	2178.00
$R^*(\mu, \alpha, \beta)$	3	2219.14
$R^*(\mu, \alpha, \gamma)$	3	2203.14
$R^*(\mu, \beta, \gamma)$	4	2185.71
$R^*(\mu, \alpha, \beta, \gamma)$	5	2220.00

cont.

Error S.S.	9	52.00
S.S.Total	15	2272.00

7.5.2 Coding Scheme B(1, 0)

Here we define

x_1 = 1 for the level 2 of the row factor

$$z_j = \begin{cases} 1 & \text{for level } (j+1) \text{ of the column factor} \\ 0 & \text{otherwise} \end{cases}$$

$j = 1, 2$

The design matrix $\underset{\sim}{X}_b^{*\prime}$ for the regression model (7.5.1) becomes

$$\underset{\sim}{X}_b^{*\prime} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix}$$

and

$$\underset{\sim}{Y}_{1\times 15}^{\prime} = (8,\ 13,\ 9, 12, 7, 11, 11, 14, 17, 14, 16, 10, 11, 14, 13.)$$

Consequently the normal equations $\underline{X}_b^{*\prime}\,\underline{X}_b^{*}\,\hat{\underline{\beta}}_b^{*} = \underline{X}_b^{*\prime}\underline{Y}$ are:

$$\begin{bmatrix} 15 & 9 & 3 & 6 & 2 & 4 \\ 9 & 9 & 2 & 4 & 2 & 4 \\ 3 & 2 & 3 & 0 & 2 & 0 \\ 6 & 4 & 0 & 6 & 0 & 4 \\ 2 & 2 & 2 & 0 & 2 & 0 \\ 4 & 4 & 0 & 4 & 0 & 4 \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{\alpha}_2 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \\ \hat{\gamma}_{22} \\ \hat{\gamma}_{23} \end{bmatrix} = \begin{bmatrix} 180 \\ 120 \\ 42 \\ 66 \\ 30 \\ 48 \end{bmatrix}$$

Following the usual procedure of analysis we get

$R^*(\mu, \alpha, \beta, \gamma) = 2220.00.$

The other sum of squares for various reduced models can be obtained by using the appropriate portion of the $\underline{X}_b^{*\prime}\,\underline{X}_{\beta}^{*}$ and $X_b^{*\prime}\,\underline{Y}$ matrices for the complete model. The different reduction in sumsof squares obtained are presented in the table given below:

Table 19. ANOVA table.

Source	d.f.	S.S.
$R^*(\mu)$	1	2160.00
$R^*(\alpha/\mu)$	1	40.00
$R^*(\beta/\mu)$	2	37.60
$R^*(\alpha,\beta/\mu)$	3	59.14

continued...

$R^*(\alpha, \gamma/\mu)$	3	54.00
$R^*(\beta, \gamma/\mu)$	4	35.00
$R^*(\alpha, \beta, \gamma/\mu)$	5	60.00
$R^*(\beta/\mu, \alpha)$	2	19.14
$R^*(\alpha/\mu, \beta)$	1	21.54
$R^*(\gamma/\mu, \alpha, \beta)$	2	0.86
$R^*(\mu, \alpha)$	1	2200.00
$R^*(\mu, \beta)$	2	2197.60
$R^*(\mu, \alpha, \beta)$	3	2219.14
$R^*(\mu, \alpha, \gamma)$	3	2214.00
$R^*(\mu, \beta, \gamma)$	4	2196.00
$R^*(\mu, \alpha, \beta, \gamma)$	5	2220.00
Error S.S.	9	52.00
Total S.S.	15	2272.00

Coding Scheme 'C'

The analyses of unbalanced data using Coding Scheme C poses no new problems. It is not given in details here in because of space. Moreover, the hypotheses tested under this scheme are not of much practical interest as they depend upon the number of observations going into each cell. If one chooses to test these hypotheses the design matrix for this scheme can be constructed on the same lines as done for Coding Schemes A and B. Then the rest of the steps are the same in the analysis procedure.

CHAPTER - VIII

DISCUSSION

The chapter deals with the discussion on three coding schemes which are used in general regression with dummy variables. We shall compare and contrast all these three schemes and examine the suitability that under which coding scheme, a particular hypothesis is tested. The index of non-orthogonality will also be discussed.

While reviewing the analysis of regression model on dummy variables, several authors have discussed the regression on dummy variables by adopting only one single coding scheme B (1, 0) which is quite inadequate. Searle (1971) and Speed, Hocking and Hackney (1978) also considered and discussed the regression method with dummy variables defined by coding scheme B (1, 0). Kleinbaum and Kuper (1978) made a passing remark that three coding schemeswill yield the identical results in case of balanced data which is entirely incorrect and misleading. They did not compare the three methods in case of unbalanced data. To the best of our knowledge, the detailed comparisons of different regression methods with dummy variables has not been done so far. The regression methods with dummy variables can be defined in many ways but here only three methods in over parametrized model discussed in the Chapter VII will be compared which

~~are~~ correspondi~~ngly~~ to the three side conditions viz.

$$(1)\quad \sum_i \alpha_i = \sum_j \beta_j = \sum_i \gamma_{ij} = \sum_j \gamma_{ij} = 0$$

$$(2)\quad \alpha_1 = \beta_1 = \gamma_{1j} = \gamma_{i1} = 0$$

$$(3)\quad \sum_i n_{i.}\alpha_{i.} = \sum_j n_{.j}\,\beta_{.j} = \sum_i n_{ij}(\alpha_i + \gamma_{ij}) = \sum_j n_{ij}(\beta_j + \gamma_{ij}) = 0$$

8.1 Comparison Of Coding Scheme (A (1, -1, 0) and Coding Scheme B(1, 0)

Consider the two coding schemes A(1, -1, 0) and scheme B(1, 0). If $n_{ij} = n$ for all i and j or in case of balanced data, then the reduction in sum of squares

$$R^*(\alpha/\mu) = R^*(\alpha/\mu,\beta) = R^*(\alpha/\mu,\beta,\gamma) \quad . \quad . \quad . \quad (8.1.1)$$

is obtained by the application of coding scheme A. But if we apply coding scheme B (1,0) with $n_{ij} = n$ it can be easily verified that $R^*(\alpha/\mu) = R^*(\alpha/\mu,\beta) \neq R^*(\alpha/\mu,\beta,\gamma)$..(8.1.2).

Searle (1971) states that unbalanced data have their own analysis of variance technique and those for balanced data, are merely special cases to technique for unbalanced data, which is incorrect. It is evident that if we adopt coding scheme B (1, 0) for the regression method with dummy variables, the result obtained using

usual analysis procedure for balanced case.

$$R^*(\alpha/\mu) = R^*(\alpha/\mu,\beta) = R^*(\alpha/\mu,\beta,\gamma)$$

does not result in as a special case of the technique for unbalanced data due to (8.1.2). On the contrary if the dummy variables are defined by the coding scheme A(1, -1, 0) the balanced case is a special case of the technique of unbalanced data yielding

$$R^*(\alpha/\mu) = R^*(\alpha/\mu,\beta) = R^*(\alpha/\mu,\beta,\gamma) \quad \ldots \quad (8.1.3)$$

~~Here~~ Searle (1971) idea is considered a very strong one and this should be the guiding principle in deciding on the analysis of variance technique for unbalanced data. Using coding scheme A(1, -1, 0) we can test the hypothesis $H_1 : \bar{\mu}_{i.} = \bar{\mu}_{i'.}$ or mean responses of the i and i'^{th} rows averaged over the columns. This hypothesis is associated with the reduction in sum of squares due to α, adjusted for β and γ, $R^*(\alpha/\mu,\beta,\gamma)$ in case of two way classification with interaction. When the regression model is without interaction, the H_3 hypothesis tests the weighted mean of two rows i and i', partial row effect (adj.)

$\sum_{j} n_{ij}\mu_{ij} = \sum_{i'}\sum_{j} \frac{n_{ij}\, n'_{ij}\, \mu'_{ij}}{n_{.j}}$ and it is associated with the reduction in sum of squares $R^*(\alpha/\mu,\beta)$ i.e. the sum of squares due to α after adjustment for β and neglecting

γ and H_3 hypothesis can be tested with any of the scheme. The hypothesis H_2 also tests similarly the two rows means difference i and i' when only single factor is considered in the regression model. The row means $\sum_{j} \frac{n_{ij}\mu_{ij}}{n_{i.}} = \sum_{j} \frac{n'_{ij}\mu'_{ij}}{n_{i'.}}$ are also weighted means and hypothesis H_2 is associated with the S.S. $R^*(\alpha/\mu)$. The hypothesis H_2 and H_3 can be tested by any coding scheme in case of unbalanced data, where H_1 is tested only by coding scheme A(1, -1, 0). If we are interested in comparing the two row's means under a particular column, then these means will be tested by hypothesis H_4 using coding scheme B(1, 0) which is associated with $R^*(\alpha/\mu, \beta, \gamma)$. The hypothesis H_9: $\mu_{ij} - \mu_{ij'} - \mu_{ij'} + \mu_{ij'} = 0$ of no interaction can be tested with any scheme and is associated with the reduction in sum of squares $R^*(\gamma/\mu, \alpha, \beta)$ which is S.S. due to γ after adjusting α and β effects together in the model. The hypotheses H_5, H_6, H_7 and H_8 for columns are synonymous to H_1, H_2, H_3 and H_4 hypothesis and test the mean differences among column means.

Consider the ANOVA table as given below:

Table 20. ANOVA table for coding scheme A and scheme B.

Source	d.f.	Coding scheme A S.S.	Coding scheme B S.S.
$R^*(\mu)$	1	2160.00	2160.00
$R^*(\alpha/\mu)$	1	40.00	40.00
$R^*(\beta/\mu)$	2	18.00	37.60

$R^*(\alpha, \beta/\mu)$	3	59.14	59.14
$R^*(\alpha, \gamma/\mu)$	3	43.14	54.00
$R^*(\beta, \gamma/\mu)$	4	25.71	36.00
$R^*(\alpha, \beta, \gamma/\mu)$	5	60.00	60.00
$R^*(\beta/\mu, \alpha)$	2	19.14	19.14
$R^*(\alpha/\mu, \beta)$	1	41.14	21.54
$R^*(\gamma/\mu, \alpha, \beta)$	2	0.86	0.86
$R^*(\mu, \alpha)$	1	2200.00	2200.00
$R^*(\mu, \beta)$	2	2178.00	2197.00
$R^*(\mu, \alpha, \beta)$	3	2219.14	2219.14
$R^*(\mu, \alpha, \gamma)$	3	2203.14	2214.00
$R^*(\mu, \beta, \gamma)$	4	2185.91	2196.00
$R^*(\mu, \alpha, \beta, \gamma)$	5	2220.00	2220.00
S.S. Error	9	52.00	52.00
S.S. Total	15	2272.00	2272.00

Examining the above ANOVA table for coding scheme A and coding scheme B, it is observed that the reduction in sum of squares due to the effects α, β, and γ, $R^*(\alpha, \beta, \gamma/\mu)$ was found 60.00 using the coding scheme A and scheme B. By examining the S.S. due to α, β, γ, , $R^*(\alpha, \beta, \gamma/\mu)$ and $R(\alpha^*, \beta^*, \gamma^*/\mu^*)$, obtained with the application of Method 1 and Method 2 in Chapter IV, it is established that $R(\alpha, \beta, \gamma/\mu)$ remains same by the application of any scheme or any method. This total sum

of squares due to α, β, and γ, $R^*(\alpha, \beta, \gamma/\mu)$ effect will be unaffected and unchanged whether the model is non full rank or full rank reparametrized model.

Consider the main effect sum of squares due to α, and β. Take one S.S. unadjusted for factor A and another S.S. adjusted for factor B. It is found that total S.S. for unadjusted and adjusted effect's is same using the coding scheme A and coding scheme B viz. $R^*(\alpha/\mu) + R^*(\beta/\mu,\alpha)$ or $R^*(\beta/\mu) + R^*(\alpha/\mu,\beta)$ is 59.14 for both the schemes. Then we can say that the reduction in sum of squares $R^*(\alpha, \beta/\mu)$ will also be equal using the coding scheme A or coding scheme B.

In fact $R(\alpha, \beta/\mu)$ will be same when we apply method 1 and method 2 adopting non full rank and reparametrized full rank model. $R(\alpha/\mu) + R(\beta/\mu,\alpha)$ or $R(\alpha^*/\mu^*) + R(\beta^*/\mu^*, \alpha^*)$ is also same using method 1 and method 2 in case of non full rank model and reparametrized full rank model. Although we are getting the different values of $R^*(\alpha/\mu)$ and $R^*(\beta/\mu)$, $R^*(\alpha/\mu,\beta)$ and $R^*(\beta/\mu,\alpha)$ adopting two coding schemes but the above result for the total will always hold good. In fact $R(\alpha, \beta/\mu)$ will be same whether we apply any method adopting non full rank or reparametrized full rank model. The significance of these reduction in sum of squares will also different. In one case it may be

significantly different and in another it may be nonsignificant. For example $R^*(\alpha/\mu, \beta)$ was found significantly different using coding scheme A, but it was not significant in case of coding scheme B. $R^*(\alpha/\mu)$ was found significant for both the schemes but $R^*(\beta/\mu)$ was not significant for any coding schemes.

In case of balanced data for method 1 and method 2 $R(\alpha/\mu, \beta, \gamma)$ was zero but here $R^*(\alpha/\mu, \beta, \gamma)$ and $R^*(\beta/\mu, \alpha, \gamma)$ were calculated non zero equal to 16.86 and 34.29 for the coding scheme A (1, -1, 0) and 16.86 and 24.29 for coding scheme B. All the values were found different.

8.2 Coding Scheme C

The coding scheme C is not as common as coding scheme A or coding scheme B. If we examine the Table17, . in chapter VII it does not test any new hypothesis. The hypothesis H_2 and H_3 associated with the sum of squares $R^*(\alpha/\mu)$ and $R^*(\alpha/\mu, \beta)$ are tested using the scheme C. The hypothesis H_9 of no interaction can also be tested with this scheme. In the light of the above arguments and discussions held if we examine all three coding schemes, it is advised and recommended that use of scheme A defined as (1, -1, 0), the regression method of analysis is most appropriate to test all hypotheses.

8.3 Statistical packages

It is interesting to note that the analysis of variance tables using BMDP-2V (Dixon, 1979) for balanced data are widely different for the two regression methods defining dummy variables by coding scheme A(1, -1, 0) and coding scheme B(1, 0). The reason being is that it first fits the regression model containing all the dummy variables for the grouping variables and their interaction and then the model containing all dummy variables but those of the main effects or interaction being tested. SPSS (1975) option 9 uses the general linear regression approach defining the dummy variables by coding scheme B (1, 0) and partitions individual effects by adjusting for all other effects and this option cannot be used for balanced case since this will result entirely in a different ANOVA table and not the standard ANOVA table for balanced case, and the hypotheses H_1 and H_5 cannot be tested by this package. On the other hand, with BMDP package, the hypotheses H_1 and H_5 can easily be tested, we do not recommend the use of the regression method with dummy variables defined by scheme B (1, 0).

This discussion does not mean that the regression approach with dummy variables defined by the coding scheme B (1, 0) does not serve any useful purpose. It does serve useful purpose in some important applications. Suppose

the cell means are presented in the table 2 1 below in a two way classification model in terms of μ-model and the general linear regression model with dummy variables defined by coding scheme B (1, 0).

Table 21. Cell means

$\mu_{11} = \mu$	$\mu_{12} = \mu + \beta_1, \; - - - \; , \mu_{1b} = \mu + \beta_{b-1}$	
$\mu_{21} = \mu + \alpha_1$	$\mu_{22} = \mu + \alpha_1 + \beta_1 + \gamma_{22}, \; - - - \; , \mu_{2b} = \mu + \alpha_1 + \beta_{b-1} + \gamma_{2b}$	
$\mu_{31} = \mu + \alpha_2$	$\mu_{32} = \mu + \alpha_2 + \beta_1 + \gamma_{32}, \; - - - \; , \mu_{3b} = \mu + \alpha_2 + \beta_{b-1} + \gamma_{3b}$	
$\vdots$	$\vdots$	$\vdots$
$\mu_{a1} = \mu + \alpha_{a-1}$	$\mu_{a2} = \mu + \alpha_{a-1} + \beta_1 + \gamma_{a2}, \; - - - \; , \mu_{ab} = \mu + \alpha_{a-1} + \beta_{b-1} + \gamma_{ab}$	

Obviously, it is clear from the above table that in the presence of interaction row effects are estimated and tested from the information in the first column, similarly, the column effects in the presence of interaction, are estimated and tested from the information in the first row only. In other words $R^*(\alpha/\mu, \beta, \gamma)$ is the sum of squares for testing the hypothesis $H_4: \mu_{i1} = \mu_{i'1}$ or testing the two row cell means under first column. And (a-1) (b-1) components of interaction are derived from the last (a-1) rows and (b-1) columns only. In general, it does not seem a good proposition.

Next if the row effects are assumed alike γ_{ij} is given by $\mu_{ij} - \mu_{1j}$ which is not the difference of the difference, a general concept associated with interaction. It is, obviously, parallel to the row effect. Similarly arguments hold good if the column effects are assumed constant. Therefore, it is inferred from the above discussion that in deciding on a test procedure, it is very important to define the parameters in the model in a purposeful way. It is very desirable also to define the parameters uniquely for the proper interpretation of the results of ANOVA table. So long as the interaction component is not present in the model, the row and column effects are defined with the common meaning irrespective of the coding schemes. The presence of the interaction component in the model and the way it is defined in the model, both are very important and which affects the definitions of the row and column effects. This explains the cause for the different hypothesis tested with the same R( ) notation. We recommend that the objectives and the type of the treatments in the experiment would govern the decision for making a choice of coding scheme.

Next we consider a situation where the regression method with dummy variables defined by the coding scheme B (1, 0) is the only appropriate method. Suppose we want

to conduct a multinutrient factorial experiment when each nutrient is tried at more than one level and level one of each nutrient is zero which is called control. The hypothesis H_9 of no two factor interaction between nutrient type and level is given by H_9:

$$\mu_{ij} - \mu_{i'j} - \mu_{ij'} + \mu_{i'j'} = 0$$

for all i, i', j and j' except j, j' = 1

This hypothesis H_9 of no two factor interaction is illustrated graphically in Figure 1.

The hypothesis states that the line segments after the first interval are parallel. If the coding scheme B (1, 0) is not used for regression method with dummy variables, then the nonparallelism in the first interval could give a false indication of the presence of interaction. And in the presence of interaction, the main effects are still purposeful and these main effects should be based on the first row and the first column means which is only done by the coding scheme B (1, 0).

Another important situation where hypotheses of the type H_4 and H_8 are considered of interest, is the varietal and fertilizer trial when zero level of fertilizer is included in the experiment and interaction is found statistically significant. Such type of experiment can be best

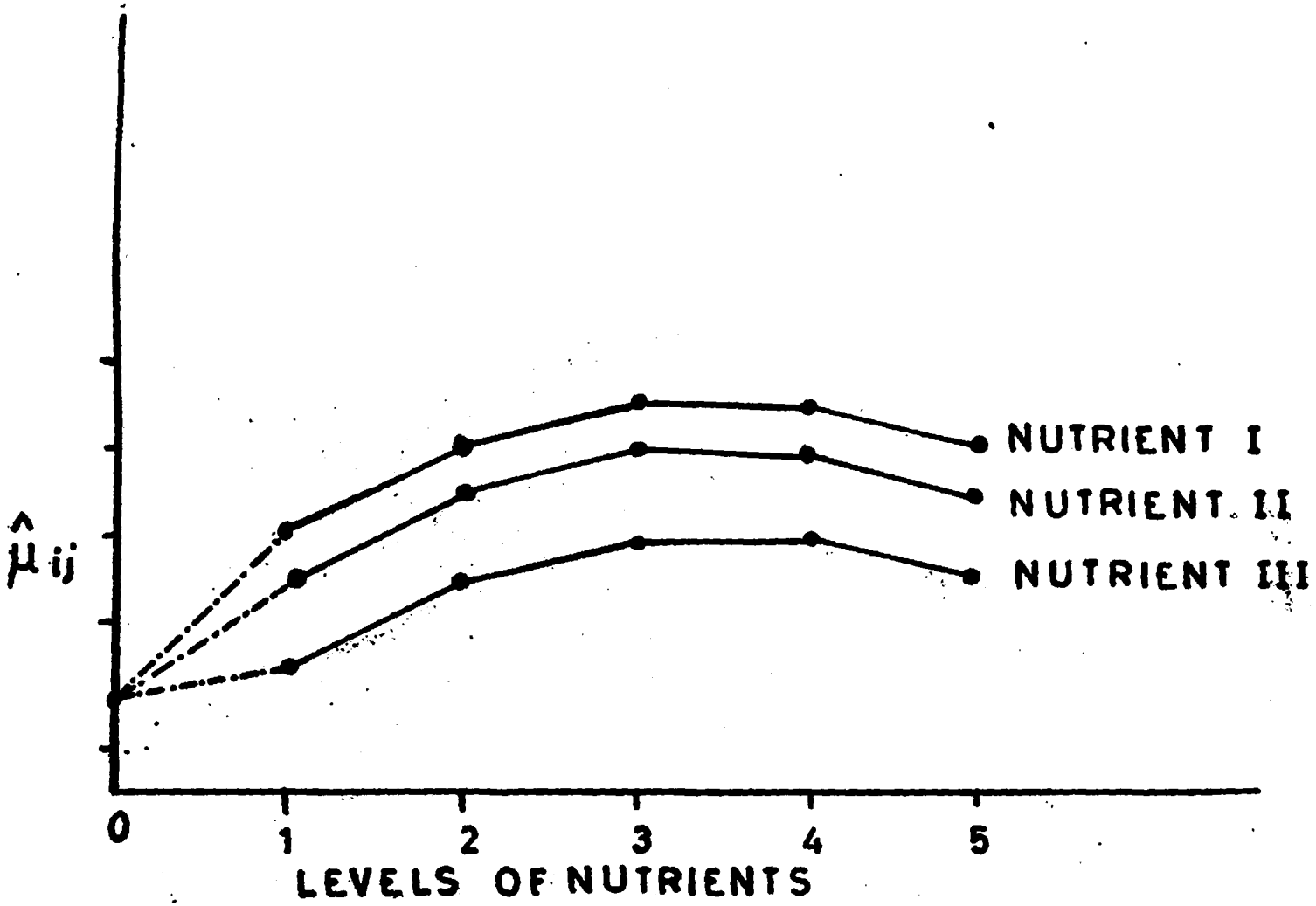

Fig-1- PLOT OF CELL MEANS

analyzed by regression method by adopting the coding scheme B (1, 0). Several authors have highlighted the hypothesis of the type H_1 and H_5 which simply test the difference in the two rows means or two columns means respectively. Another important observation we observed with several sets of unbalanced data sets, is that $R^*(\alpha/\mu, \beta, \gamma)$ for coding scheme B (1, 0) is always smaller considerably than $R^*(\alpha/\mu, \beta, \gamma)$ calculated for any other coding scheme discussed here. In our example, the S.S. for $R^*(\alpha/\mu, \beta, \gamma)$ are 35.00 and 24.00 for coding scheme A(1, -1, 0) and coding scheme B(1, 0) respectively. Similarly, it is true for $R^*(\beta/\mu, \alpha, \gamma)$ which is 16.86 and 6.00 for coding scheme A(1, -1, 0) and coding scheme B(1, 0).

8.5 Index of Non-orthogonality

The index of non-orthogonality is measured through the concept of expected average distance between the estimated parameter vector and the parameter vector. Suppose $\underline{X}$ is the design matrix of p parameter vector and (r_{ij}) is the correlation matrix of p parameter for the design matrix $\underline{X}$ of a linear regression model. If $\lambda_1, \lambda_2, \ldots, \lambda_p$ are the eigen values of the correlation matrix (r_{ij}), then

$$\sum_{i=1}^{p} \lambda_i = \sum_{i=1}^{p} \frac{1}{\lambda_i} = p \text{ for balanced data}$$

and

$$\sum_{i=1}^{p} \frac{1}{\lambda_i} > p \text{ for unbalanced data}$$

Defining $\theta = \sum_{i=1}^{p} \frac{1}{\lambda_i} / p$ then θ is called index of non-orthogonality. We find that coding scheme B(1, 0) always yields higher value of the measure of non-orthogonality, over the other schemes. Hence, if the estimation of the regression coefficients in the regression model is one of the primary purpose of the study then it is not desirable to use the coding scheme B(1, 0).

8.6 Zero Cell Frequency

It has been seen that in many situations scientists are not aware of testing of specific hypothesis when the data are analyzed and particularly in case of unbalanced data, there may be some misunderstanding among statisticians as to what hypothesis is being tested. In this section that problem is discussed where some of the n_{ij}'s observation in any cell are to be zero which may happen either by loss of some animals in case of animal experiments, or by damage or infestation in the field due to pest insects.

Consider two way classification without interaction. Suppose $n_{11} = 0$ in the Ist row of the Ist column. The model is $\underline{Y} = \underline{W}\underline{\mu} + \underline{e}$

The model is of (ab-1) length consisting of μ_{ij} for (i, j) $\neq$ (1, 1). The (a-1) (b-1)-1 constants describing the absence of interaction are obtained from

$$\mu_{ij} - \mu_{ij'} - \mu_{i'j} + \mu_{i'j'} = 0$$ by eliminating μ_{11} using the relation

$$\mu_{11} = \mu_{12} + \mu_{21} - \mu_{22}$$

Now suppose we want to test the row effects. If $n_{11} = 1$, the hypothesis $\mu_{ij} = \mu_{ij'}$ for all i, i' and j. If $n_{11} = 0$ the corresponding hypothesis is derived by eliminating μ_{11}. Thus the hypothesis is $\mu_{ij} = \mu_{ij'}$ for all i, i' and j. If the model is two way classification with interaction with some empty cells, then the situation is little complicated. A standard computer programme is there to set up a full rank model by imposing

$$\sum_{i=1}^{b} \alpha_i = \sum_{j} \beta_j = \sum_{i} \gamma_{ij} = \sum_{j} \gamma_{ij} = 0$$

and $\gamma_{.j} = 0$ if $n_{ij} = 0$. This can be analyzed by developing the normal equations and the respective sum of squares may be obtained.

CHAPTER - IX

SUMMARY

In the analysis of linear models for designed experiments with balanced data, there is a general agreement on the appropriate analysis of variance table, but there is a disagreement on the proper analysis of unbalanced data. Here, the analysis of unbalanced data has been discussed using several techniques.

The analysis of non full rank model is done by two different approaches. In one approach, the side conditions are incorporated into the normal equations whereas in the second approach the model is reparameterized by using the side conditions thus making the model full rank. Illustrations of these two methods (approaches) are given by using actual data. Various reductions in sums of squares are calculated and tested for corresponding hypotheses. All hypotheses are listed which are related to the respective reduction in sum of squares in case of one way, two way classification with and without interaction.

Different methods of analysis of linear models have been described in brief. Their sums of squares are related to the corresponding hypothesis being tested. A unified regression

approach using dummy variables for unbalanced data has been introduced and discussed in depth of which analysis of variance techniques are special cases for unbalanced data, which has removed many prevailing confusions in the analysis of linear models with unbalanced data. The regression model clearly defines the parameters and hypothesis statements.

Regression methods with dummy variables in linear models are critically examined in case of unbalanced data using three coding schemes. Various reductions in sum of squares have been calculated using coding scheme A(1, -1, 0) and coding scheme B (1, 0), their sum of squares were related to the corresponding hypothesis to be tested. It has been observed that the reduction in sum of squares fitting the model μ, α, β and γ is the same using any method or any coding schemes, SS error also turns out equal under both the schemes. The total of main effects for the factor A and factor B is same when we consider SS (adj) for one factor A and SS (unadj) for another factor B and vice versa. It is found that $R^*(\alpha/\mu, \beta, \gamma)$ and $R^*(\beta/\mu, \alpha, \gamma)$ using coding scheme B(1, 0) are considerably smaller than using ~~any~~ other scheme.

The zero cell frequency case or missing cell case has been discussed in case of unbalanced data giving various hypothesis. Measure of extent of the non-orthogonality

through the concept of eign values of the correlation matrix (r_{ij}) for the design matrix $\underline{X}$ of parameter of linear regression model has been proposed for unbalanced case. The proposed index of non-orthogonality is a thumb rule only. The non-orthogonality index is generally higher if coding scheme B (1, 0) is adopted for the analysis of unbalanced data. Therefore, it is recommended that one should use regression method with dummy variables using coding scheme A(1, -1, 0) for appropriate analysis of linear models for unbalanced data.

BIBLIOGRAPHY

Anderson,R.L. and Bancroft,T.A. 1952. Statistical Theory in Research. McGraw Hill Book Co. N.Y.

Ballas,J.A. and Webester,J.T. 1966. On dependent tests from a non-orthogonal design. J. Am. Stat. Assoc. 29, 51-56.

Barr,A.J., Goodnight,J.H., Sall,J.P. and Helwig,J.T. 1976. A users guide to SAS-76, Raleigh, N.C. SAS Institute.

Carlson,J.E. and Timm,N.H. 1974. Analysis of non-orthogonal fixed effects design. Psychological Bulletin. 81, 563-570.

Crump,S.I. 1946. The estimation of variance components in analysis of variance. Biom. 2, 7-11.

Draper,N.R. and Smith,H. 1966. Applied regression analysis. John Wiley and Sons, N.Y.

Elston,R.C. and Bush,N. 1964. The hypothesis that can be tested when there are interaction in the analysis of variance models. Biom. 20, 681-690.

Federer,W.T. 1963. Relationship between a three way classification. Disproportionate number analysis of variance and several two way classification and nested analysis. Biom. 19, 629.

Federer,W.T. and Paik,U.B. 1974. Analysis of non-orthogonal n-way classification. Annals. Stat. 2, 1000-1021.

Federer,W.T. and Zelon,M. 1966. Analysis of multifactor classification with unequal numbers of observation. Biom. 22, 525-562.

Francis,I. 1973. A comparison of several analysis of variance programmes. J. Am. Stat. Assoc. 68, 860-865.

Freeman,G.H. and Jeffers,J.N.R. 1962. Estimation of means and standard errors in the analysis of non-orthogonal experiments by electronic computer. J. Royal Stat. Soc. Sr. B.: 24, 425-435.

Gosslac,D.G. and Lucas,H.L. 1965. Analysis of variance of disproportionate data when interactions are present. Biom. 21, 115-133.

Gabriel,K.R. 1963. Analysis of variance of proportions with unequal frequencies. J. Am. Stat. Assoc. 58, 1133.

Graybill,F.A. 1961. An introduction to linear statistical models. vol.1. McGraw Hill Book Co. N.York.

Golhar,M. and Skillings,J. 1976. A comparison of several analysis of variance. Programmes with unequal cell size. Communication in Statistics Simulations and Computation B.S., 43-54.

Hammerle,W.J. 1974. Non-orthogonal analysis of variance using iterative improvement and balanced residuals. J. Am. Stat. Assoc. 69, 722-779.

Hartwell,T.D. and Gaylor,D.W. 1973. Estimating variance components for two way disproportionate data with missing cells by method of unweighted means. J. Am. Stat. Assoc. 68, 379.

Harvey,W.R. 1968. Instruction for use of least squares and maximum likelihood, general purpose programme. Ohio State University, Columbus, Ohio.

Harvey,W.R. 1968. Least square analysis of data with unequal subclass numbers. A.R.S. USAID July 20, 8.

Henderson,C.R. 1953. Estimation of variance covariance components. Biom. 9, 226-262.

Hocking,R.R., Hackney,O.P. and Speed,F.M. 1978. The analysis of linear models with unbalanced data. J. Am. Stat. Assoc. 73, 133-151.

Hocking,R.R. and Speed,F.M. 1975. A full rank analysis of some linear model problems. J. Am. Stat. Assoc. 70, 706-712.

Hocking,R.R. and Speed,F.M. 1979. The analysis of linear models with unbalanced data: Concepts and methods. 35th Annual Conference of Applied Statistics. Villanova Univ. December 6.

Kleinbaum,D.G. and Kuper,L.L. 1978. Applied regression analysis and multivariate methods. Duxbury Press, Massachusetts.

Kutner,M.H. 1974. Hypothesis testing in linear models. Am. Stat. 28, 98-100.

Mielke,P.W. and Mchugh,R.B. 1965. Two way analysis of variance for mixed model with disproportionate subclass frequencies. Biom. 21, 308.

Nair,K.R. 1941. A note on the fitting constants for analysis of non-orthogonal data arranged in double classification. Sankhya 5, 317-328.

Nelder,J.A. 1974. Letter to the editor. J. Am. Stat. Assoc. Series C. 23, 232.

Neter,J. and Wasserman,W. 1974. Applied linear statistical models. Illinois, Richard, D. Irwin.

Overall, J.E. and Klett,J.C. 1972. Applied multivariate analysis. McGraw Hill Book Co. N.York.

Overall,J.E. and Spiegel,D.K. 1969. Concerning least squares analysis of experimental data. Psychological Bull. 7, 311-322.

Pearce,S.C. 1963. The use and classification of nonorthogonal designs. J. Royal Stat. Soc. Series A. 126,353.

Rao,C.R. 1946. On the linear combination of observations and the general theory of least squares. Sankhya. 7, 237-256.

Rao,C.R. 1955. Analysis of dispersion for multiply classified data with unequal numbers in cells. Sankhya 15, 253-280.

Rao,C.R. 1965. Linear statistical inference and its application. John Wiley and Sons, New York.

Raut,K.C. 1960. Generalized nonorthogonal design and its analysis with recovery of inter block information. J. Ind. Soc. Ag. Stat. 12, 190-199.

Read,R.R. 1961. On quadratic estimates of inter class variance for unbalanced data. Jr. Royal Stat. Soc. Series B. 23, 493.

Rees,D.H. 1969. The analysis of variance of some nonorthogonal designs with split plots. Biometrika. 56, 43.

Scheffe,H. 1959. Analysis of variance. John Wiley and Sons, New York.

Searle,S.R. 1970. Large sample variances of maximum liklihood estimates of variance components using unbalanced data. Biom. 26, 505-524.

Searle,S.R. 1971. Linear models. John Wiley and Sons, New York.

Searle,S.R. 1972. Using the R( ) notation for reduction in sums of squares when fitting linear models. Presented at Spring Regional meeting of ENAR. Ames, Iowa.

Searle,S.R. 1976. Comments on ANOVA calculations for messy data. Paper presented at the Ist International SAS users meeting. Kissimee, Florida, USA. Jan.26-28.

Searle,S.R. 1979. Annoted computer output for analysis of variance of unequal subclass numbers data. Am. Stat. 33, No.4.

Smith,H.F. 1951. Analysis of variance with unequal but proportionate numbers of observation in the subclass of two way classification. Biom. 7:70-74.

Snedecor,W.G. and Cochran,W.G. 1967. Statistical methods. 6th Edition, Ames,Iowa. The Iowa Univ. Press.

SPSS . . 1975. Statistical Package for the Social Sciences, New York. McGraw Hill Book Co.

Speed,F.M. and Hocking,R.R. 1976. The use of R( ) notation with unbalanced data. Am. Stat. 30, 30-34.

Speed,F.M., Hocking,R.R. and Hackney,O.P. 1978. Method of analysis of linear models with unbalanced data. J. Am. Stat. Assoc. 73, 105-112.

Steele, R.G.D. and Torrie, J.H. 1960. Principles and procedures of statistics. McGraw Hill Book Co. N.York.

Winer, B.J. 1971. Statistical principles in experimental design. McGraw Hill Book Co. N.York.

Yates, F. 1934. The analysis of multiple classification with unequal numbers in different classes. J. Am. Stat. Assoc. 29, 52-66.