Global Journal of Science Frontier Research: F
Mathematics and Decision Sciences
Volume 14 Issue 6 Version 1.0 Year 2014
Type : Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 2249-4626 & Print ISSN: 0975-5896

Two Factor Analysis of Variance and Dummy Variable Multiple Regression Models

By Oyeka ICA & Okeh UM
Nnamdi Azikiwe University, Nigeria

Abstract- This paper proposes and presents a method that would enable the use of dummy variable multiple regression techniques for the analysis of sample data appropriate for analysis with the traditional two factor analysis of variance techniques with one, equal and unequal replications per treatment combination and with interaction.

The proposed method, applying the extra sum of squares principle develops F ratio-test statistics for testing the significance of factor and interaction effects in analysis of variance models. The method also shows how using the extra sum of squares principle to build more parsimonious explanatory models for dependent or criterion variables of interest.

In addition, unlike the traditional approach with analysis of variance models the proposed method easily enables the simultaneous estimation of total or absolute and the so-called direct and indirect effects of independent or explanatory variables on given criterion variables. The proposed methods are illustrated with some sample data and shown to yield essentially the same results as would the two factor analysis of variance techniques when the later methods are equally applicable.

Keywords: dummy variable regression, Analysis of variance, degrees of freedom, treatment, regression coefficient..

GJSFR-F Classification : MSC 2010: 62J05

Notes

Two Factor Analysis of Variance and Dummy Variable Multiple Regression Models

Oyeka ICA [α] & Okeh UM [σ]

Abstract- This paper proposes and presents a method that would enable the use of dummy variable multiple regression techniques for the analysis of sample data appropriate for analysis with the traditional two factor analysis of variance techniques with one, equal and unequal replications per treatment combination and with interaction.

The proposed method, applying the extra sum of squares principle develops F ratio-test statistics for testing the significance of factor and interaction effects in analysis of variance models. The method also shows how using the extra sum of squares principle to build more parsimonious explanatory models for dependent or criterion variables of interest.

In addition, unlike the traditional approach with analysis of variance models the proposed method easily enables the simultaneous estimation of total or absolute and the so-called direct and indirect effects of independent or explanatory variables on given criterion variables. The proposed methods are illustrated with some sample data and shown to yield essentially the same results as would the two factor analysis of variance techniques when the later methods are equally applicable.

Keywords: dummy variable regression, Analysis of variance, degrees of freedom, treatment, regression coefficient.

I. Introduction

Analysis of variance and regression analysis whether single-factor or multi-factor, sometimes both in theory and applications have often been treated and presented as rather different concepts by various authors. In fact only limited attempts seem to have been made to present analysis of variance as a regression problem (Draper and Smith, 1966; Neter and Wasserman, 1974).

Nonetheless analysis of variance and regression analysis are actually similar concepts, especially when analysis of variance is presented from the perspective of dummy variable regression models. This is the focus of the present paper, which attempts to develop a method to use dummy variable multiple regression models and apply the "extra sum of squares principle" in the analysis of two-factor analysis of variance models with unequal replications per treatment combination as a multiple regression problem.

II. The Proposed Method

Regression techniques can be used for the analysis of data appropriate for two factor or two –way analysis of variance with replications and possible interactions. This approach is a more efficient method than the method of unweighted means discussed in Oyeka et al (2012).

Author α : Department of Statistics, Nnamdi Azikiwe University Awka, Nigeria. Abakaliki, Nigeria. e-mail: uzomaokey@ymail.com
Author σ : Department of Industrial Mathematics and Applied Statistics, Ebonyi State University Abakaliki, Nigeria.

In a two factor analysis of variance involving factors A and B with interactions between these two factors, as discussed in Oyeka (2013), the resulting model is

$$y_{ilj} = \mu + \alpha_l + \beta_j + \lambda_{lj} + e_{ilj} \tag{1}$$

Where y_{ilj} is the i^{th} observation or response at the l^{th} level of factor A and j^{th} level of factor B; μ is the grand or overall mean, α_l is the effect of the l^{th} level of factor A, β_j is the effect of the j^{th} level of factor B; λ_{lj} is the interaction effect between the l^{th} level of factor A and j^{th} level of factor B; e_{ilj} are independent and normally distributed error terms with constant variance, for $i = 1,2..nij, \imath = 1,2\ldots a,$ the 'a' levels of factor A; $j = 1,2..b,$ the 'b' levels of factor B, subject to the constraints

$$\sum_{l=1}^{a} \alpha_i = \sum_{j=1}^{b} \beta_j = \sum_{l=1}^{a} \lambda_{lj} = \sum_{j=1}^{b} \lambda_{lj} = 0 \tag{2}$$

Let $n = \sum_{l=1}^{a} \sum_{j=1}^{b} n_{ij}$ be the total sample size or observations for use in the analysis.

To obtain a dummy variable regression model of 1s and 0s equivalent to equation 1 and also subject to the constraints imposed on the parameters by equation 2, we would as usual use for each factor one dummy variable of 1s and 0s less than the number of levels, classes, or categories that factor has (Boyle 1974). Similarly the interaction effects will be factored in by taking the cross-products of the set of dummy variables representing one of the factors with the set of dummy variables representing the other factor. Thus factor A with 'a' levels will be represented by a-1 dummy variables of 1s and 0s, factor B with 'b' levels will be represented by b-1 dummy variables of 1s and 0s and the factors A by B interaction effects will be represented by (a-1)(b-1)dummy variables of 1s and 0s.Specifically to obtain the required dummy variables for factors A and B. we may define

$$x_{il;A} = \begin{cases} 1, if\ \ the\ i^{th} observation\ or\ response,\ y_{ilj}\ \ is\ at\ the\ l^{th}\ \ level\ of\ \ factor\ A \\ 0,\ otherwise \end{cases} \tag{3}$$

for $i = 1,2 \ldots n_{lj}\ \ l = 1,2 \ldots a - 1, for\ all\ j = 1,2 \ldots b$

For factor B define

$$x_{ij;B} = \begin{cases} 1, if\ \ the\ i^{th} observation\ or\ response,\ y_{ilj}\ \ is\ at\ the\ J^{th}\ \ level\ of\ \ factor\ B \\ 0,\ otherwise \end{cases} \tag{4}$$

for $i = 1,2 \ldots n_{lj}, j = 1,2 \ldots b - 1, for\ all\ l = 1,2 \ldots a$

Using these specifications we have that the dummy variable multiple regression model equivalent to the two factor analysis of variance model of equation 1 is

$$y_{ilj} = \beta_0 + \beta_{1;A}.x_{i1;A} + \beta_{2;A.}x_{i2;A} + \ldots + \beta_{a-1;A}.x_{ia-1;A} + \beta_{1;B}.x_{i1;B} + \beta_2;_B.x_{i2;B} + \ldots + \beta_{b-1};_B.x_{ib-1};_B + \beta_1;I.x_{il};_I\ I$$
$$+ \beta_2;I.x_{i2};I + \ldots + \beta_{(a-1)(b-1)} : I.x_{i(a-1)(b-1)};I + e_{ilj}$$

OR when more compactly expressed

$$y_{ilj} = \beta_0 + \sum_{l=1}^{a-1} \beta_{l;A}.x_{il;A} + \sum_{j=1}^{b-1} \beta_{j;B}.x_{ij;B} + \sum_{k=1}^{(a-1)(b-1)} \beta_{k;I}.x_{ik};I + e_{ilj} \tag{5}$$

Where the β_s are partial regression coefficients and e_{ij} are independent and normally distributed error terms with constant variance with $E(e_{ilj}) = 0.$

The expected value of y_{ilj} of Equation (5) is

$$E(y_{ilj}) = \beta_0 + \sum_{l=1}^{a-1} \beta_{l;A}.x_{il;A} + \sum_{j=1}^{b-1} \beta_{j;B}.x_{ij;B} + \sum_{k=1}^{(a-1)(b-1)} \beta_{k;}I.x_{ik};I \tag{6}$$

Note that the interaction terms may be more completely represented as

$$x_{ik;}I = x_{il:A}.x_{ij:B} = x_{il}.x_{ij}; and\ \beta_{k};I = \beta_{ij}:\ AB = \beta_{lj}$$

For $l = 1,2 \ldots a-1 : j = 1,2, \ldots b-1$

Hence Equation 6 may alternatively be expressed as

$$E(y_{ilj}) = \beta_0 + \sum_{l=1}^{a-1} \beta_{l;A}.x_{il;A} + \sum_{j=1}^{b-1} \beta_{j;B}.x_{ij;B} + \sum_{l=1}^{(a-1)}\sum_{j=1}^{(b-1)} \beta_{lj}.x_{il}.x_{ij} \tag{7}$$

Now the mean value or mean response in the language of analysis of variance at the l^{th} level factor A and j^{th} level of factor B is obtained by setting $x_{il;A} = x_{ij:B} = 1$the l^{th} $and\ x_{ig} = 0$ for all 'g'not equal to $l.j$ in Equation (7) to obtain

$$E(y_{ilj}) = \mu_{lj} = \beta_0 + \beta_{l;A} + \beta_{j:B} + B_{lj} \tag{8}$$

For $l = 1,2, \ldots a-1; j = 1,2, \ldots b-1$

Similarly the mean response or mean of the criterion variable at the l^{th} level of factor A is obtained by setting x_{il};A=1 and all other $x_{igs} = 0$ $(g \neq l)$ while the mean response at the j^{th} level of factor B is obtained by setting $x_{ij}:B = 1$ and all other $x_{igs} = 0$ $(g \neq j)$ in Equati on (6) giving

$$\mu_j = \beta_0 + \beta_l:A; and\ \mu_j = \beta_0 + \beta_j:B \tag{9}$$

For $l = 1,2, \ldots a-1; j = 1,2, \ldots b-1$

These are the same results that are obtained using conventional two factor analysis of variance methods. The partial regression parameter β_l: A is as usual interpreted as the change in the dependent variable 'Y' percent change in the l^{th} level of factor A compared with all its other levels holding the levels of all other independent variables in the model constant; β_j : B is similarly interpreted. The interaction effect β_{lj} is interpreted as the dependent variable Y per unit change at the l^{th} level of factor A cj^{th} level of the change at the l^{th} level of factor B confounded by or in the presence of the effect of the j^{th} level of factor B (l^{th} level of factor A).

Now Equation 5 can be more compactly expressed in matrix form as

$$\underline{y} = X\underline{\beta} + \underline{e} \tag{10}$$

Where $\underline{y}$ is an $nx1$ column vector of response outcomes or values of the criterion or dependent variable; X is an nxr design matrix of 1s and 0s; $\underline{\beta}$ is an $rx1$ column vector of partial regression parameters and $\underline{e}$ is an $nx1$ column vector of normally distributed error terms with constant variance with $E(\underline{e}) = \underline{0}$,where $r=a.b-1$ representing the number of dummy variables of 1s and 0s in the model.

The corresponding expected value of the criterion variable equivalent to Equation (6) is

$$E(\underline{y}) = X.\underline{\beta} \qquad (11)$$

Note that use of Equations 3-5 or 10 makes it unnecessary, at least for the fixed effects model of primary interest here, to treat one observation per cell, equal and unequal observations per cell in two factor analysis of variance problems differently. The same dummy variable regression models can be used in all these cases except that in the case of one observation per cell where it is not possible to calculate the error sum of squares and hence the corresponding error mean square, the interaction mean square is instead used in all tests.

Use of the usual least squares methods with either Equations (5) or (10) yields unbiased estimates of the partial regression parameters which again expressed in matrix form is

$$\underline{\hat{\beta}} = \underline{b} = (X'X)^{-1}.X'\underline{y} \qquad (12)$$

Where $(X'X)^{-1}$ is the matrix inverse of the non singular variance-covariance matrix $X'X$.

The resulting estimated or fitted value of the response or dependent variable is

$$\underline{\hat{y}} = X.\underline{b} \qquad (13)$$

In the conventional two factor analysis of variance a null hypothesis that is usually of interest first is that treatment means are equal for all treatment combinations. In the dummy variable regression approach an equivalent null hypothesis would be that the specified model that is either Equations (5) or (10) fits. This null hypothesis when expressed in terms of the regression parameters would be

$$H_o : \underline{\beta} = \underline{0} \; versus \; H_1 : \underline{\beta} \neq \underline{0} \qquad (14)$$

This null hypothesis is tested using the usual F_{test} presented in the familiar analysis of variance table where the required sums of squares are obtained as follows:-

The total sum of squares is as usual calculated as

$$SS_{Total} = SS_{Tot} = \underline{y}'\underline{y} - n\bar{y}^2 \qquad (15)$$

Which has the chi-square distribution with $n-1$ degrees of freedom where $\bar{y}$ is the mean of the criterion or dependent variable. The sum of squares regression or the so-called treatment sum of squares in analysis of variance parlance is

$$SSR = SST = \underline{b}'.X'\underline{y} - n\bar{y}^2 \qquad (16)$$

Which has the chi-square distribution with $r = a.b-1$ degrees of freedom. Similarly the error sum of squares is

Notes

$$SSE = SS_{Total} - SSR = \underline{y}'\,\underline{y} - \underline{b}'.X'\,\underline{y} \tag{17}$$

With $(n-1)$-($a.b-1) = n-a.b$ degrees of freedom.
These results may be summarized in an analysis of variance Table (Table 1)

Table 1: Analysis of variance table for regression model of Equation (10)

Source of Variation	Sum of Squares (SS)	Degrees of Freedom (DF)	Mean sum of Squares (MS)	F-Ratio
Regression (treatment)	$SSR = SST = \underline{b}'X'\underline{y} - n.\bar{y}^2$	$a.b-1$	MSR=$\frac{SSR}{a.b-1}$	F=$\frac{MSR}{MSE}$
Error	$SSE = \underline{y}'\,\underline{y} - \underline{b}'X'\underline{y}$	$n-a.b$	MSE=$\frac{SSE}{n-a.b}$	
Total	$SS_{Toat} = \underline{y}'\,\underline{y} - n.\bar{y}^2$	$n-1$		

The null hypothesis of Equation 4 is rejected at the α level of significance if the calculated $F-ratio$ of Table 1 is such that

$$F \ge F_{1-\alpha;a.b-1}, n-a.b \tag{18}$$

Otherwise the null hypothesis is accepted.

If the model fits, that is if the null hypothesis of Equation (14) is rejected, in which case not all the regression parameters are equal to zero, then one can proceed to test other null hypothesis concerning factors A and B level effects as well as factors A by B interaction effects. Thus additional null hypothesis that may be tested are that factor A has no effects on the criterion variable; factor B has no effects on the criterion variable; and that there are no factors A by B interaction effects. Stated notation ally the null hypotheses are

$$H_{0A}: \underline{\beta}_A = \underline{0} \text{ Versus } H_{1A}: \underline{\beta}_A \ne \underline{0} \tag{19}$$

$$H_{0B}: \underline{\beta}_B = \underline{0} \text{ Versus } H_{1B}: \underline{\beta}_B \ne \underline{0} \tag{20}$$

$$H_{0AB}: \underline{\beta}_{AB} = \underline{0} \text{ Versus } H_{1AB}: \underline{\beta}_{AB} \ne \underline{0} \tag{21}$$

To test these hypotheses one needs to calculate the contribution of each factor separately to the treatment or regression sum of squares. The treatment or regression sum of squares SST in analysis of variance parlance which is the regression sum of squares SSR in regression models distributed as chi-square with $a.b-1$ degrees of freedom is made up of three sums of squares each having the chi-square distribution, namely the sum of squares due to row or factor A, SSA with $a-1$ degrees of freedom, the sum of squares due to column or factor B, SSB with $b-1$ degrees of freedom, and the row by column of factors A by B interaction sum of squares, SSAB with ($a-1$)($b-1$) degrees of freedom. Thus notationally we have that

$$SST = SSR = SSA + SSB + SSAB \tag{22}$$

To obtain these sums of squares we note that the design matrix X of Equation (10) with $ab-1$ dummy variables 0s and 1s because of 0s and 1s of dummy variables of 1s and 0s can be partitioned into three sub matrices namely an $n \times (a-1)$ matrix X_A of $a-1$ dummy variables representing the $(a-1)$. included levels of factor A, the

$n\times(b-1)$matrix X_B comprising $b-1$ dummy variables of 1s and 0s representing the $b-1$included levels of factor B; and the $n\times(a-1)(b-1)$matrix $X_{AB}$of $(a-1)(b-1)$ dummy variable of 1s and 0s representing interaction between factors A and B .

The $(ab-1)\times 1$column vector of estimated partial regression coefficients b can be similarly partitioned into the corresponding $(a-1)\times 1$column vector b_A of effects due to factor A; the $(b-1)\times 1$column vector of b_B of effects due to factor B and the $(a-1)(b-1)\times 1$column vector $\underline{b}_{AB}$ of effects due to factors A by B interaction. Now the sum of squares $\underline{b}'.X'.\underline{y}$ of Equation (16) may hence equivalently be expressed as

$$\underline{b}'X'\underline{y} = (X.\underline{b})'.\underline{y} = \left(\begin{pmatrix} X_A & X_B & X_{AB} \end{pmatrix}\begin{bmatrix} b_A \\ b_B \\ b_{AB} \end{bmatrix}\right)'\underline{y}$$

OR

$$\underline{b}'X'\underline{y} = \underline{b}'_A.X'_A.\underline{y} + \underline{b}'_B.X'_B.\underline{y} + b'_{AB}.X'_{AB}.\underline{y} \qquad (23)$$

The sum of squares regression or the treatment sum of squares, $SSR = SST$ of the full model of Equation 10 is

$$\begin{aligned} SSR &= \underline{b}'.X'.\underline{y} - n.\bar{y}^2 = \left(\underline{b}'_A.X'_A.\underline{y} - n.\bar{y}^2\right) + \left(\underline{b}'_B.X'_B.\underline{y} - n.\bar{y}^2\right) + \left(\underline{b}'_{AB}.X'_{AB}.\underline{y} - n.\bar{y}^2\right) + 2.n.\bar{y}^2 \qquad (24) \\ SSR(SST) &= SSA \qquad + SSB \qquad + SSAB \qquad + mean \\ & \qquad\qquad\qquad\qquad\qquad\qquad (adjustment\ \ factor) \end{aligned}$$

Now to find the required sums of squares after fitting the full regression model of Equation (10) one then proceeds to fit, that is regress the dependent variable $\underline{y}$ separately as reduced models on $X_A \quad X_B \ and \ X_{AB}$ to obtain using the usual least square methods, the three terms of Equation (20) or (24). Now the sums of squares and the corresponding estimated regression coefficients on the right hand side of Equation (24) are obtained by fitting reduced regression models separately of $X_A, X_B \ and \ X_{AB}$ as reduced design matrices. That is the dependent variable $\underline{y}$ is separately fitted, that is regressed on each of the reduced design matrices $X_A, X_B \ and \ X_{AB}$.

These regression models would yield estimates of the corresponding reduced partial regression parameters, $\underline{\beta}_A, \underline{\beta}_B \ and \ \underline{\beta}_{AB}$ as respectively

$$\hat{\underline{\beta}}_A = \underline{b}_A = (X'_A.X_A)^{-1}.X'_A.\underline{y}; \hat{\underline{\beta}}_B = \underline{b}_B = (X'_B.X_B)^{-1}.X'_B.\underline{y}; \hat{\underline{\beta}}_{AB} = \underline{b}_{AB} = (X'_{AB}.X_{AB})^{-1}.X'_{AB}.\underline{y} \qquad (25)$$

If the full model of Equation (10) fits, that is if the null hypothesis of Equation (14) is rejected, then the additional null hypothesis of Equations 19-21 may be tested using the extra sum of squares principle (Drapa and Smith, 1966). If we denote the sum of squares due to the full model of Equation (10) and the reduced models due to the fitting of the criterion variable $\underline{y}$ to any of the reduced design matrices by $SS(F) \ and \ SS(R)$, respectively then following the extra sum of squares principle (Draper and Smith, 1966;Neter and Wasserman 1974), the extra sum of squares due to a given factor is calculated as $ESS = SS(F) - SS(R)$ Equation (26)with degrees of freedom obtained as the difference between the degrees of freedom of $SS(F) \ and \ SS(R)$. That is as

$$Edf = df(F) - df(R) \tag{27}$$

Thus the extra sums of squares for factors A, B and A by B interaction are obtained as respectively

$$ESSA = SSR - SSA; ESSB = SSR - SSB; ESSAB = SSR - SSAB \tag{28}$$

With degrees of freedom of respectively

$$(ab-1)-(a-1) = a(b-1); (ab-1)-(b-1) = b(a-1); (ab-1)-(a-1)(b-1) = a+b-2 \tag{29}$$

Note that since each of the reduced models and the full model have the same total sum of squares, SS_{Tot} ,the extra sum of squares may alternatively be obtained as the difference between the error sum of squares of each reduced model and the error sum of squares of the full model. In other words the extra sum of squares is equivalently calculated as

$$ESS = SS(F) - SS(R) = (SS_{Tot} - SS(F)) - (SS_{Tot} - SS(R)) = SSE(R) - SSE(F) \tag{30}$$

With the degrees of freedom similarly obtained as

$$Edf = df\, SSE(R) - df\, SSE(F) \tag{31}$$

Thus the extra sums of squares due to factors A, B and A by B interaction are alternatively obtained as

$$ESSA = SSEA - SSE; ESSB = SSEA - SSE; ESSAB = SSEAB - SSE \tag{32}$$

Where SSR and SSE are respectively the regression sum of squares and the error sum of squares for the full model. The null hypothesis of Equations (19) - (21) are tested using the F ratios as summarized in Table 2 which for completeness also includes the values of Table 1 for the full model.

Table 2 : Two factor Analysis of Variance Table showing Extra Sums of Squares

Source of Variation	Sum of Squares (SS)	Degrees of Freedom (DF)	Mean sum of Squares (MS)	F Ratio	Extra Sum of Squares (SS(F)-SS(R))	Degrees of Freedom (DF)	Extra Mean Sum of Squares (EMSR)	F Ratio
Regression	$SSR=\underline{b}'.X'.\underline{y}-n\bar{y}^2$	$ab-1$	$MSR=\frac{SSR}{ab-1}$	$F=\frac{MSR}{MSE}$	$SSR=\underline{b}'X'.\underline{y}-n\bar{y}^2$	$ab-1$	$MSR=\frac{SSR}{ab-1}$	$F=\frac{MSR}{MSE}$
Full Model								
Error	$SSE=\underline{y}'.\underline{y}-\underline{b}'.X'.\underline{y}$	$n-a.b$	$MSE=\frac{SSE}{n-a.b}$		$SSE=\underline{y}'\underline{y}-\underline{b}'X'.\underline{y}$	$n-a.b$	$MSE=\frac{SSE}{n-a.b}$	
Factor A								
Regression	$SSA=\underline{b}'_A.X'_A.\underline{y}-n\bar{y}^2$	$a-1$	$MSA=\frac{SSA}{a-1}$	$F=\frac{MSA}{MSEA}$	$ESSA=SSR-SSA$	$a(b-1)$	$EMSA=\frac{ESSA}{a(b-1)}$	$F=\frac{EMSA}{MSE}$
Error	$SSEA=\underline{y}'.\underline{y}-\underline{b}'_A.X'_A.\underline{y}$	$n-a$	$MSEA=\frac{SSEA}{n-a}$		$ESSEA=SSEA-SSE$ $=ESSA$	$a(b-1)$	$EMSSA=\frac{ESSEA}{a(b-1)}$	
Factor B								
Regression	$SSB=\underline{b}'_B.X'_B.\underline{y}-n\bar{y}^2$	$b-1$	$MSB=\frac{SSB}{b-1}$	$F=\frac{MSB}{MSEB}$	$ESSB=SSR-SSB$	$b(a-1)$	$EMSB=\frac{ESSB}{b(a-1)}$	$F=\frac{EMSB}{MSE}$
Error	$SSEB=\underline{y}'.\underline{y}-\underline{b}'_B.X'_B.\underline{y}$	$n-b$	$MSEB=\frac{SSEB}{n-a}$		$ESSEB=SSEB-SSE$ $=ESSB$	$b(a-1)$	$EMSEB=\frac{ESSEB}{b(a-1)}$	
Factor A by B interaction								
Regression	$SSAB=\underline{b}'_{AB}.X'_{AB}.\underline{y}-n\bar{y}^2$	$(a-1)$ $(b-1)$	$MSAB=\frac{SSAB}{(a-1)(b-1)}$	$F=\frac{MSAB}{MSEAB}$	$ESSAB=SSR-SSAB$	$a+b$ -2	$EMSAB=\frac{ESAB}{a+b-2}$	$F=\frac{EMSAB}{MSE}$
Error	$SSEB=\underline{y}'.\underline{y}-\underline{b}'_{AB}.X'_{AB}.\underline{y}$	$n-$ $(a-1)$ $(b-1)$	$MSEAB=\frac{SSEAB}{n-(a-1)(b-1)}$		$SSEAB-SSE=ESSAB$	$a+b$ -2	$EMSEAB=\frac{ESSEAB}{a+b-2}$	
Total	$SS_{Tot}=\underline{y}'.\underline{y}-n.\bar{y}^2$	$n-1$			$SS_{Tot}=y'.y-n.\bar{y}^2$	$n-1$		

Notes

Note that the F ratios of Table 2 are each the ratio of the extra mean sum of squares of the corresponding reduced model to the mean sum of squares of the full model. Where SSR is the regression sum of Squares with $a.b-1$ degrees of freedom, SSE is the error sum of Squares with $n-a.b$ degrees of freedom and MSE is the mean error sum of Squares, all for the full model of Equation (10). The results of Table 2 are the same as would be obtained using the conventional two factor or two way analysis of variance with replications and interactions. Usually, the hypothesis of no interaction is tested first using the corresponding F ratio of Table 2. If the hypothesis of no interaction is accepted, then one may proceed to test the null hypotheses about factors A and B effects again using the corresponding F ratios of Table 2. If however the null hypothesis of no interaction is rejected, then one may use any of the familiar and appropriate methods of treating interactions and proceed with further analysis.

Thus if the model of Equation (10) fits, that is if the null hypothesis of Equation (14) is rejected then the null hypotheses of Equations 19-21 of no factors A, B and A by B interaction are respectively tested using the corresponding test statistics (see Table 2), namely

$$F_A = \frac{EMSA}{MSE}; F_B = \frac{EMSB}{MSE}; F_{AB} = \frac{EMSAB}{MSE} \tag{33}$$

With numerator degrees of freedom of $a(b-1), b(a-1),$ and $a+b-2$ respectively and common denominator degrees of freedom of $n-a.b$ for use to obtain the necessary critical $F-ratios$ for comparative purposes for rejection or acceptance of the corresponding null hypothesis.

Note that in general whether or not the independent or explanatory variables used in a regression model are dummy variables or numeric measurements, the extra sum of squares principle is most useful in determining the contribution of an independent variable or a subset of the independent variables among all the independent variables in the model in explaining the variation of a specified dependent on criterion variable. This would inform the inclusion or exclusion of the independent variable or the subset of the independent variables in the hypothesized model depending on the significance of the contribution.

Thus the extra sum of squares principle enables one select important variables and formulate a more parsimonious statistical model of explanatory variables for a dependent variable of interest. To do this, for example, for one independent variable $X_{j,}$ included in a regression model with say a total of $'r'$independent variables, over fits the full model with all the independent variables and reduced model with only one independent variable X_{jA} .Suppose as discussed earlier that the regression sums of squares for the full model and the reduced model are respectively $SS(F)\ and\ SS(R)$ which have degrees of freedom of $'r'$and 1 respectively. Then from Equation (28) the extra sum of squares regression with respect to X_j is

$$ESS(X_j) = SS(F) - SS(R) \tag{34}$$

With r-1 degrees of freedom. The corresponding extra mean sum of squares is

$$EMS(X_j) = \frac{ESS(X_j)}{r-1} \tag{35}$$

The significance of β_j, the partial regression coefficient or effect X_j on the criterion variable y is determined using the test statistic.

$$FX_j = \frac{EMS(X_j)}{MSE} \tag{36}$$

Which has $r-1$ and $n-r$ degrees of freedom for $j=1,2,...r;$ where MSE is the error mean square for the full model and 'n' is the total sample size.

An advantage of using dummy variable regression models in two factor and multi factor analysis of variance is that the method enables the estimation of other effects separately of several factors on a specified dependent or criterion variable. For example it enables the estimation of the total or absolute effect, the partial regression coefficient or the so-called direct effect of a given independent variable on the dependent variable through the effects of its representative dummy variables as well as the indirect effect of that parent independent variable through the meditation of other parent independent variables in the model (Wright, 1934).

The total or absolute effect of a parent independent variable on a dependent variable is estimated as the simple regression coefficient of that independent variable represented by codes assigned to its various categories, when regressed on the dependent variable. The direct effect of a parent independent variable on a dependent variable is the weighted sum of the partial regression coefficients or effects of the dummy variables representing that parent independent variable on the dependent variable, where the weights are the simple regression coefficients of each representative dummy variable regressing on the specified parent independent variable represented by codes. The indirect effect of a given parent independent variable on a dependent variable is then simply the difference between its total and direct effects (Wright 1973). Now the direct effect or partial regression coefficient of a given parent independent variable A say on a dependent variable Y is obtained by taking the partial derivative of the expected value of the corresponding regression model with respect to that parent independent variable. Thus the direct effect of the parent independent variable A on the dependent variable Y is obtained from Equation 7 as

$$\beta_A dir = \frac{dE(y_{ilj})}{d_A} = \sum_{l=1}^{a-1} \beta_l ; A.\frac{dE(x_{il};A)}{d_A} + \sum_g \beta_g ; Z\frac{dE(x_{ig};Z)}{dA} \quad OR$$

$$\beta_A dir = \sum_{l=1}^{a-1} \beta_l ; A\frac{dE(x_{il;A})}{dA} \tag{37}$$

Since $\sum_g \beta_g ; Z\frac{dE(x_{ig}:Z)}{dA} = 0,$ for all other independent variables Z in the model different from A.

The weight, $\alpha_l ; A = \frac{dE(x_{il;A})}{dA}$ is estimated by fitting a simple regression line of the dummy variable $x_{il;A}$ regressing on its parent independent variable, A represented by codes and taking the derivative of its expected value with respect to A, Thus if the expected value of the dummy variable $x_{il;A}$ regressing on its parent independent variable A is expressed as

Notes

Then the derivative of $E(x_{il;A})$ with respect to A is

$$\frac{dE(x_{il;A})}{dA} = \alpha_{j;A} \tag{39}$$

Hence using Equation 39 in Equation 37 gives the direct effect of the parent independent variable A on the dependent variable Y as

$$\beta_{Adir} = \sum_{l=1}^{a-1} \alpha_{l;A_l} . \beta_{l;A} \tag{40}$$

Whose sample estimate is from Equation 12

$$\hat{\beta}_{Adir} = b_{Adir} = \sum_{l=1}^{a-1} \alpha_{l;A} . b_{l;A} \tag{41}$$

The total or absolute effect of A on Y is estimated as the simple regression coefficient or effect of the parent independent variable A represented by codes on the dependent variable Y as

$$\hat{\beta}_{A} = b_{A} \tag{42}$$

Where b_A is the estimated simple regression coefficient or effect of A on Y. The indirect effect of A on Y is estimated as the difference between b_A and $b_{Adir,}$ that is as

$$\hat{\beta}_{Aind} = b_{Aind} = b_A - b_{Adir} \tag{43}$$

The total, direct and indirect effects of factor B are similarly estimated. These results clearly give additional useful information on the effects of given factors on a specified dependent or criterion variable than would the traditional two factor analysis of variance model.

III. Illustrative Example

In a study of Encephalitic and Meningitic brain damage each of a random sample of 36 patients is given a battery of tests on mental acuity recording a composite score for each patient. Low scores on this composite measure indicate some degree of brain damage. The patients are divided into 3 groups according to the predisposing factor of initial infection and into 3crossed groups according to time to observed physical recovery from the illness. A control group of other mental patients are similarly studied with the following results. (Table 3)

Table 3 : Mental acuity of sample data of patients with diagnosed metal illness by factor and time to recovery.

	Time to Recovery (B)		
Predisposing factor (A)	**1 – 2 years (1)**	**3 – 5 years (2)**	**7 – 10 years (3)**
Encephalitis (1)	76 73 75 62	69 53 72	59, 43 41 57, 55
Meningitis (2)	81 89 83	82 70 91, 74 75	68 50 75 47
Other (Control (3)	75 79 84 65 63	85 76 87	98 100 82 79

Do there seem to be significant differences in performance among the encephalitic, meningitic and other (control) patients? Among patients according to time to recovery? Is there any interaction between predisposing factor of illness and time to recovery of patients?

To answer these questions using dummy variable multiple regression analysis or model, we represent the predisposing factor here called factor A which has three levels with two dummy variables of Is and Os namely $x_{i1.A}$ for (1) Encephalitis and $x_{12;A}$ for (2) Meningitis. Time to recovery here called factor B also with three levels is represented by two dummy variables of Is and 0s namely $x_{i1;B}$ for (1) 1 – 2 years and $x_{i2;B}$ for (2) 3 – 5 years. The interaction terms are represented by the cross products of these dummy variables namely $x_{i3} = x_{i1;A}.x_{i2;B}; x_{i4}; = x_{i1;A}.x_{i2;B}; x_{i5} = x_{i2;B}.x_{i1;B}$ *and* $x_{i6} = x_{i2;A.}.x_{i2;B}$ *for* $i = 1, 2 \ldots 36$ yielding the design matrix of Table 4.

Table 4 : Design Matrix X for the Data of Table 3

S/ N	y_{ilj}	x_{i0}	$x_{i1;A}$	$x_{i2;A}$	$x_{i1;B}$	$x_{i2;B}$	x_{i3} $(x_{i1;A}.x_{i1;B})$	x_{i4} $(x_{i1;A}.x_{i2;B})$	x_{i5} $(x_{i2;A}.x_{i1;B})$	x_{i6} $(x_{i2;A}.x_{i2;B})$
1.	76	1	1	0	1	0	1	0	0	0
2.	73	1	1	0	1	0	1	0	0	0
3.	75	1	1	0	1	0	1	0	0	0
4.	62	1	1	0	1	0	1	0	0	0
5.	69	1	1	0	0	1	0	1	0	0
6.	53	1	1	0	0	1	0	1	0	0
7.	72	1	1	0	0	1	0	1	0	0
8.	59	1	1	0	0	0	0	0	0	0
9.	43	1	1	0	0	0	0	0	0	0
10.	41	1	1	0	0	0	0	0	0	0
11.	57	1	1	0	0	0	0	0	0	0
12.	55	1	1	0	0	0	0	0	0	0
13.	81	1	0	1	1	0	0	0	1	0
14.	89	1	0	1	1	0	0	0	1	0
15.	83	1	0	1	1	0	0	0	1	0
16.	82	1	0	1	0	1	0	0	0	1
17.	70	1	0	1	0	1	0	0	0	1
18.	91	1	0	1	0	1	0	0	0	1
19.	74	1	0	1	0	1	0	0	0	1
20.	75	1	0	1	0	1	0	0	0	1
21.	68	1	0	1	0	0	0	0	0	0
22.	50	1	0	1	0	0	0	0	0	0
23.	75	1	0	1	0	0	0	0	0	0
24.	47	1	0	1	0	0	0	0	0	0
25.	75	1	0	0	1	0	0	0	0	0
26.	79	1	0	0	0	0	0	0	0	0
27.	84	1	0	0	0	0	0	0	0	0
28.	65	1	0	0	0	0	0	0	0	0
29.	63	1	0	0	0	0	0	0	0	0
30.	85	1	0	0	1	1	0	0	0	0
31.	76	1	0	0	1	1	0	0	0	0
32.	87	1	0	0	1	1	0	0	0	0
33.	98	1	0	0	1	0	0	0	0	0
34.	100	1	0	0	1	0	0	0	0	0
35.	82	1	0	0	1	0	0	0	0	0
36.	79	1	0	0	1	0	0	0	0	0

Fitting the full model of Eqn 10 using the design matrix X of table 4, we obtain the fitted regression equation

$$\hat{y}_{ilj} = 83.642 - 32.642x_{i1;A} - 18.726x_{i2A} - 10.442x_{i1;B} + 7.168x_{i2;B} - 30.942x_{i3} + 6.499x_{i4} + 29.860x_{i5} + 2.383x_{i6} \quad (Pvalue = 0.0000) \tag{44}$$

A P-value of 0.0000 clearly shows that the model fits.

The expected score by patients on the mental acuity test by predisposing factor (factor A), is obtained by setting $x_{i1;A} = x_{i2;A} = 1$, and all other $x_{ijs} = 0$ in equation (44) giving

$$\hat{y}_{ilj} = 83.642 - 32.642 - 18.726 = 32.274$$

The estimated response or score on the mental acuity test by length of time to observed physical recovery is similarly estimated by setting $x_{i1;B} = x_{i2;B} = 1$ and all other $x_{ils} = 0$ in Equation (44) yielding

$$\hat{y}_{ilj} = 83.642 - 10.442 + 7.168 = 80.368$$

The corresponding analysis of variance table for the full model is presented in Table 5.

Table 5 : Anova Table for the Full Model of Equation (44)

Source of Variation	**Sum of Squares (SS)**	**Degrees of freedom (Df)**	**Mean Sum of Squares (MS)**	**F-Ratio**	**P-Value**
Regression (treatment)	4597.321	8	574.665	5.468	0.0000
Error	2837.652	27	105.098		
Total	7434.972	35			

Having fitted the full model which is here seen to fit, we now proceed to fit the dependent variably $\underline{y}$ separately on each of the sub matrices X_A and X_B each with two dummy variables of Is and 0s and X_{AB} with four dummy variables of Is and 0s as reduced models to obtain the corresponding sum of squares due to each of these factors. The sums of squares due to factor A, B and A by B interaction are calculated following Equation (24). The results are summarized in a two factor analysis of variance Table with extra sums of squares (Table 6)

Table 6 : Two factor Analysis of Variance Table with Extra Sums of Squares for the Sample Data of Table 3

Source of Variation	Sum of Squares (SS)	Degrees of freedom (Df)	Mean sum of squares MS	F-Ratio	Extra Sum of Squares (ESS)	Degrees of freedom (Df)	Extra mean sum of squares (EMS)	F-Ratio	Critical F value P-value
Full Model									
Regression	4597.321	8	574.665	5.468	4597.321	8	574.665	5.468	3.030
Error	2837.652	27	105.098		2837.652	27	105.098		
Factor A									
Regression	2413.556	2	1206.778	7.931	2183.765	6	363.963	3.463	2.46
Error	5021.417	33	152.164		2837.652	6	472.942		
Factor B									
Regression	817.650	2	408.825	2.039	3779.671	6	629.945	1.096	2.46
Error	6617.322	33	200.525		-3779.67	6	-629.95		
Factor A by B Interaction									
Regression	624.201	4	156.050	0.710	3973.12	4	993.28	1.728	2.73
Error	6810.771	31	219.702		-3973.12	4	-993.23		
Total	7434.972	35			7434.972	35			

Note: * indicates statistical significance at the 5 percent level

These analyses indicate that the hypothesized model fits, that is that not all the factor level effects are zero. Furthermore, there does not seem to exist any significant interaction between predisposing factor of illness A and time to observed physical recovery B. However only the predisposing factor of illness A is seen to have significant effect on the criterion variable Y namely patient composite score on mental acuity.

Finally to estimate the direct effect or partial regression coefficient of A, say, represented by the dummy variables $x_{i1;A}$ and $x_{i2;A}$, we first estimate the simple regression coefficient resulting when theses dummy variables are each regressed on A using Equation 39, yielding.

$$\alpha_{1;A} = -\frac{1}{2} = -0.5; \alpha_{2;A} = 0$$

Using these results with Equation (44) in (41), we obtain an estimate of the direct effect of A on 'y' as

$$b_A dir = (-0.5)(-32.642) + (0)(-10.442) = 16.321$$

The estimated simple regression coefficient or effect of A on y is $b_A = 9.917$
Hence the estimated indirect effect of A on'y' is from Equation (43)

$$b_A ind = 9.917 - 16.321 = -6.404$$

The absolute, direct and indirect effects of B on 'y' are similarly estimated.

IV. Summary and Conclusion

We have in this paper proposed and developed a method that enabled the use of dummy variable multiple regression techniques for the analysis of data appropriate for use with two factor analysis of variance models with unequal observations per treatment combination and with interactions. The proposed model and method employed the extra sum of squares principle to develop appropriate test statistics of F ratios to test for the significance of factor and interaction effects.

The method which was illustrated with some sample data was shown to yield essentially the same results as would the traditional two factor analysis of variance model with unequal observations per cell and interaction. However the proposed method is more generalized in its use than the traditional method since it can easily be used in the analysis of two-factor models with one observation, equal, and unequal observations per cell as a rather unified analysis of variance problem.

Furthermore unlike the traditional analysis of variance models the proposed method is able to enable one using the extra sum of squares principle, to determine the relative contributions of independent variables or some combinations of these variables in explaining variations in a given dependent variable and hence build a more parsimonious explanatory model for any variable of interest. In addition, the method enables the simultaneous estimation of the total or absolute, direct and indirect effects of a given independent variable on a dependent variable, which provide additional useful information.

Notes

References Références Referencias

1. Boyle, Richard P (1974) Path Analysis and Ordinal Data. In Blalock, H M (ed) Causal Model in the Social Sciences Aldine Publishing Company Chicago 1974.
2. Draper, N.R and Smith, H. (1966). Applied Regression Analysis: John Wiley & sons, Inc., New York.
3. Neter, J.and Wasserman, W. (1974). Applied Linear Statistical Models. Richard D. Irwin Inc, ISBN 0256014981, ISSN- 101-423-199. New York.
4. Oyeka, I.C.A, Afuecheta E.O, Ebuh G.U and Nnanatu C.C (2012): Partitioning the total chi-square for matched Dichotomous Data. International Journal of Mathematics and Computations (IJMC), ISBN 0974-570X (online), ISSN-0974-5718 (print) vol 16; issue no 3, pp 41-50.
5. Oyeka, I.C.A, Uzuke C.U, Obiora-ilouno H.O and Mmaduakor C (2013): Ties Adjusted Two way Analysis of Variance tests with unequal observations per cell. Science Journal of Mathematics & Statistics (SJMS), ISSN:2276-6324:.
6. Wright, Sewall (1934): The Methods of Path Coefficients. Annals of Mathematical Statistics: Vol 5

Notes

Notes

This page is intentionally left blank