ULTIPLE LINEAR REGRESSION VIEWPOINTS
olume 11, Number 1, September, 1981

MISSING CELLS IN A TWO-WAY CLASSIFICATION

JOHN D. WILLIAMS
THE UNIVERSITY OF NORTH DAKOTA

\bstract - A data set with one missing cell is investigated with a number of plausible hypotheses regarding the means. It is shown that the set of hypotheses likely to be of interest correspond to a result computationally identical to the unadjusted main effects solution.

The two-way fixed effects analysis of variance with disproportionate cell frequencies has been considered by many different researchers. The "full rank model" solution, described by Timm and Carlson (1975), has been purported to be a "best" solution. Overall, Spiegel and Cohen (1975) have also opted for this solution, though Overall and Spiegel (1969) earlier had shown a preference for the fitting constants solution. Cohen (1968) described a hierarchical model that has the advantage of being an additive solution. Jennings (1967) and Williams (1972) describe a solution that address probable hypotheses of interest. Jennings approached the problem in a classical regression formulation, whereas Williams showed that the same results could be computationally found in a simpler manner. Perhaps unfortunately, Williams termed the solution the unadjusted main effect solution.

Other researchers have used a combination of approaches rather than use exclusively a single solution. Among such researchers are Searle (1971) and Applebaum and Cramer (1974).

Focusing on the hypothesis tested has been the direction of Speed and Hocking (1976) and Searle, Speed and Henderson (1981). In the latter article, they show that, with missing cells, the usual hypotheses for rows and columns lose their meaning and that it is much more beneficial to concentrate on cells. This approach would seem to be in keeping with Jenning's (1967) earlier article.

Comparisons of the hypotheses tested in the full rank model solution of Timm and Carlson (1975) and the unadjusted main effects solution was shown in Williams (1977). In a companion to the present paper Williams (this issue) showed that when the data are proportional but not equal in cell frequencies that the hypotheses tested could vary from those a researcher wishes to test for the Timm and Carlson full rank model solution. The direction of the present paper is to examine the hypotheses when missing cells occur.

The data are taken from Williams (1974, p. 77) except that the three data points in the last cell (engineering females) are omitted.

ACT Scores

Sex	Arts and Sciences	Education	Engineering
Male	20	21	21
	18	17	22
	18	19	16
	16	14	18
	21	12	23
	22	26	
	24	28	
	28	21	
	29	14	
	16	15	
	18		
	13		
	15		
	18		
	17		
Female	19	23	
	17	29	
	17	21	
	16	17	
	18	15	
	27	13	
	14		
	15		
	16		

Note that if the three data points had been included for the engineering females, the data would be proportional, and the analysis is given in Williams (this issue).

To analyse the data, an analysis using contrast coding is used to effect what might be termed a "quasi-analysis of variance solution" using the full rank model approach of Timm and Carlson and the unadjusted main effect solution of Williams. In addition to the Y (criterion) variable, four other variables are defined:

$X_1^* = 1$ if male, -1 if female;

$X_2^* = 1$ if in the College of Arts and Sciences, 0 if in the College of Education, -1 if in the College of Engineering;

$X_3^* = 0$ if in the College of Arts and Sciences, 1 if in the College of Education, -1 if in the College of Engineering; and

$X_4^* = X_1^* \cdot X_2^*$.

Six models can be defined:

$Y = b_0 + b_1X_1^* + e_1,$ (1)

$Y = b_0 + b_2X_2^* + b_3X_3^* + e_2,$ (2)

$Y = b_0 + b_1X_1^* + b_2X_2^* + b_3X_3^* + e_3,$ (3)

$Y = b_0 + b_1X_1^* + b_2X_2^* + b_3X_3^* + b_4X_4^* + e_4,$ (4)

$Y = b_0 + b_1X_1^* + b_4X_4^* + e_5,$ (5) and

$Y = b_0 + b_2X_2^* + b_3X_3^* + b_4X_4^* + e_6.$ (6)

In equations 1 through 6, the b's are regression coefficients specific to an equation (b_0 will likely be different for the different equations; so also the b's are specific to an equation); the e's are error terms associated with each equation.

Table 1 shows the sums of squares generated by these models.

While the results for each main effect are different depending on whether the measurement is made in the presence of the other main effect or the main effect and the interaction, this outcome would be expected from our knowledge of the disproportionate case.

It is instructive to set up binary coded predictors and then state and test likely hypotheses of interest. Five cell variables can be defined:

Table 1

Two-Way Solution for the Missing Cell Data

Source of Variation	df	SS	MS	F
Sex	1	SS_1= 7.51	7.57	.35
Sex (Independent of **College**)	1	SS_3-SS_2=10.68-5.64=5.64=5.14	5.14	.24
Sex (Independent of **College and Interaction**)	1	SS_4-SS_5=28.74-25.24=3.50	3.50	.16
College	2	SS_2=5.64	2.84	.13
College (Independent of Sex)	2	SS_3-SS_1=10.68-7.51=3.17	1.59	.07
College (**Independent of Sex and Interaction**)	2	SS_4-SS_6=28.74-8.71=20.03	10.02	.46
Interaction	1	SS_4-SS_5=28.74-10.68=18.06	18.06	.83
Within	40	SS_{DEV_4}=873.16	21.83	

X_1 = 1 if in cell 1 (Male, Arts & Science), 0 otherwise;

X_2 = 1 if in cell 2 (Male, Education), 0 otherwise;

X_3 = 1 if in cell 3 (Male, Engineering), 0 otherwise;

X_4 = 1 if in cell 4 (Female, Arts and Science), 0 otherwise; and

X_5 = 1 if in cell 5 (Female, Education), 0 otherwise.

Then a full model can be defined:

$$Y = b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + b_5X_5 + e_7. \quad (7)$$

An alternative full model utilizing the unit vector is

$$Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + e_7. \quad (8)$$

Hypotheses for Rows

Now, several different hypotheses that might be of interest for the row effect can be investigated. Four such hypotheses will be treated:

H_1: $b_1 + b_2 = b_4 + b_5$, a hypothesis for regression coefficients that corresponds to $\frac{\bar{Y}_1 + \bar{Y}_2}{2} = \frac{\bar{Y}_4 + \bar{Y}_5}{2}$. Note that H_1 fails to address altogether membership in cell 3; it also tests a hypothesis among the means that does not take into account the varying cell frequencies.

H_2: $\frac{b_1 + b_2 + b_3}{3} = \frac{b_4 + b_5}{2}$. While H_2 takes cell 3 into account, it does not address the varying cell frequencies.

H_3: $\frac{15b_1 + 10b_2}{25} = \frac{9b_4 + 6b_5}{15}$. H_3 takes into account the unequal cell frequencies, it does not take into account cell 3.

H_4: $\frac{15b_1 + 10b_2 + 5b_3}{30} = \frac{9b_4 + 6b_5}{15}$. H_4 takes into account both the unequal sized groups and cell 3.

t H_1, the restriction shown in H_1 is imposed on the full model;

$_1$: $b_1 + b_2 = b_4 + b_5$

r $b_1 = b_4 + b_5 - b_2$.

$= (b_4 + b_5 - b_2) X_1 + b_2 X_2 + b_3 X_3 + b_4 X_4 + b_5 X_5 + e_8$

r

$= b_2(X_2 - X_1) + b_4(X_4 + X_1) + b_5(X_5 + X_1) + b_3 X_3 + e_8$

$= X_2 - X_1$;

$= X_4 + X_1$; and

$= X_5 + X_1$,

$= b_2 D_1 + b_4 D_2 + b_5 D_3 + b_3 X_3 + e_8$. (9)

the model shown for equation 9,

$_9 = 26.92$; $SS_4 - SS_9 = 28.74 - 26.92 = 1.82$;

$\frac{1.82}{21.83} = .08$, a value that does not correspond to any given in Table 1.

interest is in using R^2s rather than SS, the equation

$\frac{(R^2_{FM} - R^2_{RM})/df_1}{(1 - R^2_{FM})/df_2}$ where R^2_{FM} refers to the R^2 term for the full model

$_{RM}$ refers to the R^2 term for the restricted model.

$= \frac{(.03187 - .02985)/1}{(1 - .03187)/40} = .08$, as before.

H_2: $\frac{b_1 + b_2 + b_3}{3} = \frac{b_4 + b_5}{2}$;

$b_1 = 3/2b_4 + 3/2b_5 - b_2 - b_3$. Imposing this restriction of the full

yields

$(3/2b_4 + 3/2b_5 - b_2 - b_3)X_1 + b_2X_2 + b_3X_3 + b_4X_4 + b_5X_5 + e_9$.

$b_2(X_2 - X_1) + b_3(X_3 - X_1) + b_4(X_4 + 3/2X_1) + b_5(X_5 + 3/2X_1) + e_9$.

Let $D_1 = X_2 - X_1$;

$D_4 = X_3 - X_1$;

$D_5 = X_4 + 3/2\ X_1$; and

$D_6 = X_5 + 3/2\ X_1$.

Then $Y = b_2D_2 + b_3D_4 + b_4D_5 + b_5D_6 + e_9$. (10)

Using the model shown for equation 10, $SS_{10} = 23.72$; $SS_4 - SS_{10} = 28.74 - 23.72 =$ [illegible]

$F = \frac{5.02}{21.83} = .23$. Note that H_2 does not yield any solution for sex shown in Table 1.

Using H_3: $\frac{15b_1 + 10b_2}{25} \quad \frac{9b_4 + 6b_5}{15}$, or $b_1 = b_4 + 2/3b_5 - 2/3\ b_2$, an impositio[illegible] is made on the full model:

$Y = (b_4 + 2/3\ b_5 - 2/3b_2)X_1 + b_2X_2 + b_3X_3 + b_4X_4 + b_5X_5 + e_{10}$.

$Y = b_2(X_2 - 2/3X_1) + b_3X_3 + b_4(X_4 + X_1) + b_5(X_5 + 2/3X_1) + e_{10}$.

Let $D_7 = X_2 - 2/3X_1$;

$D_2 = X_4 + X_1$; and

$D_8 = X_5 + 2/3X_1$.

Then $Y = b_2D_7 + b_3X_3 + b_4D_2 + b_5D_8 + e_{10}$. (11)

Using the model shown for equation 11,

$SS_{11} = 23.70$, $SS_4 - SS_{11} = 28.74 - 23.70 = 5.04$.

$F = \frac{5.04}{21.83} = .23$, a value that does not correspond to any outcome for the sex effect shown in Table 1.

Consider H_4: $\frac{15b_1 + 10b_2 + 5b_3}{30} = \frac{9b_4 + 6b_5}{15}$, or

$$b_1 = 6/5\ b_4 + 4/5b_5 - 2/3b_2 - 1/3b_3.$$

Imposing this restriction on the full model yields

$Y = (6/5b_4 + 4/5b_5 - 2/3b_2 - 1/3b_3)X_1 + b_2X_2 + b_3X_3 + b_4X_4 + b_5X_5 + e_{11}.$

$Y = b_2(X_2 - 2/3X_1) + b_3(X_3 - 1/3X_1) + b_4(X_4 + 6/5X_1) + b_5(X_5 + 4/5X_1) + e_{11}.$

Let $D_9 = X_2 - 2/3X_1$;

$D_{10} = X_3 - 1/3X_1$;

$D_{11} = X_4 + 6/5X_1$; and

$D_{12} = X_5 + 4/5X_1$.

Then $Y = b_2D_9 + b_3D_{10} + b_4D_{11} + b_5D_{12} + e_{11}$. (12)

Using equation 12,

$SS_{12} = 21.23$; $SS_4 - SS_{12} = 28.74 - 21.23 = 7.51$;

$F = \frac{7.51}{21.83} = .35$. It can be noted that the result for H_4 is identical with the use of equation 1, which is the unadjusted sex effect.

Hypotheses for Columns

Four different hypotheses can be given for the column effect also:

H_5: $b_1 + b_4 = b_2 + b_5$. Note that H_5, like H_1, disregards cell 3 and does not take into account the unequal sized cell frequencies.

H_6: $\frac{b_1 + b_4}{2} = \frac{b_2 + b_5}{2} = b_3$. H_6, like H_2, does not take into account the unequal sized cell frequencies.

H_7: $\frac{15b_1 + 9b_4}{24} = \frac{10b_2 + 6b_5}{16}$. H_7, like H_3, takes into account the unequal sized cell frequencies, but disregards cell 3.

H_8: $\frac{15b_1 + 9b_4}{24} = \frac{10b_2 + 6b_5}{16} = b_3$. H_8, like H_4, takes into account the unequal sized cell frequencies and cell 3.

To test H_5: $b_1 + b_4 = b_2 + b_5$, or $b_1 = b_2 + b_5 - b_4$.

Imposing this restriction on the full model yields

$Y = (b_2 + b_5 - b_4)X_1 + b_2X_2 + b_3X_3 + b_4X_4 + b_5X_5 + e_{12}.$

$Y = b_2(X_2 + X_1) + b_3X_3 + b_4(X_4 - X_1) + b_5(X_5 + X_1) + e_{12}.$

Let $D_{13} = X_2 + X_1$;

$D_{14} = X_4 - X_1$; and

$D_3 = X_5 + x_1$.

Then $Y = b_2D_{13} + b_3X_3 + b_4D_{14} + b_5D_3 + e_{12}$. (13)

Using equation 13, $SS_{13} = 25.68$; $SS_4 - SS_{13} = 28.74 - 25.68 = 3.06$.

$F = \frac{3.06}{21.83} = .14$, a value that does not correspond to any of the outcomes for the college effect in Table 1.

Regarding H_6: $\frac{b_1 + b_4}{2} = \frac{b_2 + b_5}{2} = b_3$, two different restrictions are implied.

Solving for b_1 and b_3, $b_1 = b_2 + b_5 - b_4$; $b_3 = \frac{b_2 + b_5}{2}$.

Imposing these restrictions on the full model,

$Y = (b_2 + b_5 - b_4)X_1 + b_2X_2 + \frac{(b_2 + b_5)X_3}{2} + b_4X_4 + b_5X_5 + e_{13}.$

$Y = b_2(X_2 + X_1 + 1/2X_3) + b_4(X_4 - X_1) + b_5(X_5 + X_1 + 1/2X_3) + e_{13}.$

Let $D_{15} = X_2 + x_1 + 1/2X_3$;

$D_{14} = X_4 - X_1$; and

$D_{16} = X_5 + X_1 + 1/2X_3$.

Then $Y = b_2D_{15} + b_4D_{14} + b_5D_{16} + e_{13}$. (14)

Using equation 14, $SS_{14} = 19.68$; $SS_4 - SS_{14} = 28.74 - 9.06 = 19.68$.

$F = \frac{9.06/2}{21.83} = .21$, a value that does not correspond to any outcome for the college effect in Table 1.

H_7 is given as $\frac{15b_1 + 9b_4}{24} = \frac{10b_2 + 6b_5}{16}$.

Solving for b_1, $b_1 = b_2 + 3/5b_5 - 3/5b_4$.

posing this restriction on the full model yields,

$Y = (b_2 + 3/5b_5 - 3/5b_4)X_1 + b_2X_2 + b_3X_3 + b_4X_4 + b_5X_5 + e_{14}$.

$Y = b_2(X_2 + X_1) + b_3X_3 + b_4(X_4 - 3/5X_1) + b_5(X_5 + 3/5X_1) + e_{14}$.

Let $D_{13} = X_2 + X_1$;

$D_{17} = X_4 - 3/5X_1$; and

$D_{18} = X_5 + 3/5X_1$.

ien $Y = b_2D_{13} + b_3X_3 + b_4D_{17} + b_5D_{18} + e_{14}$. (15)

ing equation 15, $SS_{15} = 28.24$; $SS_4 - SS_{15} = 28.74 - 28.24 = .50$.

$F = \frac{.50}{21.83} = .02$, a value that does not correspond to any outcome shown

or the college effect in Table 1.

Finally, regarding H_8: $\frac{15b_1 + 9b_4}{24} = \frac{10b_2 + 6b_5}{16} = b_3$, two restrictions

hown in terms of b_1 and b_3) are implied:

$b_1 = b_2 + 3/5b_5 - 3/5b_4$ and

$b_3 = 5/8b_2 + 3/8b_5$.

mposing these restrictions on the full model,

$Y = (b_2 + 3/5b_5 - 3/5b_4)X_1 + b_2X_2 + (5/8b_2 + 3/8b_5)X_3 + b_4X_4 + b_5X_5 + e_{15}$.

$Y = b_2(X_2 + X_1 + 5/8X_3) + b_4(X_4 - 3/5X_1) + b_5(X_5 + 3/5X_1 + 3/8X_3) + e_{15}$.

Let $D_{19} = X_2 + X_1 + 5/8X_3$;

$D_{17} = X_4 - 3/5X_1$; and

$D_{20} = X_5 + 3/5X_1 + 3/8X_3$.

Then $Y = b_2D_{19} + b_4D_{17} + b_5D_{20} + e_{15}$. (16)

Using equation 16, $SS_{16} = 23.10$; $SS_4 - SS_{16} = 28.74 - 23.10 = 5.64$.

$F = \frac{5.64/2}{21.83} = .13$, the same result, in a computational sense, of the unadjusted main effect for colleges.

Hypothesis for Interaction

In testing the hypothesis for interaction, it can be noted that cell 3 does not enter into the interaction. Thus, the likely hypothesis of interest in terms of the means is $\overline{Y}_1 - \overline{Y}_4 = \overline{Y}_2 - \overline{Y}_5$.

In terms of the regression coefficients, the null hypothesis would be tested by $b_1 - b_4 = b_2 - b_5$ (this hypothesis will be called H_9).

Then, in terms of b_1, $b_1 = b_2 - b_5 + b_4$. Imposing this restriction on the full model yields,

$$Y = (b_2 - b_5 + b_4)X_1 + b_2X_2 + b_3X_3 + b_4X_4 + b_5X_5 + e_{16}.$$

$$Y = b_2(X_2 + X_1) + b_3X_3 + b_4(X_4 + X_1) + b_5(X_5 - X_1) + e_{16}.$$

Let $D_{13} = X_2 + X_1$;

$D_2 = X_4 + X_1$; and

$D_{21} = X_5 - X_1$.

Then $Y = b_2D_{13} + b_3X_3 + b_4D_2 + b_5D_{21} + e_{16}$. (17)

Using equation 17, $SS_{17} = 10.68$; $SS_4 - SS_{17} = 28.74 - 10.68 = 18.06$, thus

$F = \frac{18.06}{21.83} = .83$, the result given for interaction in Table 1. The results of using H_1 thru H_9 with both SS and R^2s are shown in Table 2.

Discussion

A considerable amount of effort has been expended by many different researchers in investigating the two-way analysis of variance with disproportionate cell frequencies. In regard to any two-way layout of data, four situations regarding the cell frequencies can be put in order of their stress on the analysis: 1) equal numbers in each cell; 2) unequal but proportional

Table 2

Hypotheses for Two-Way Analysis of Variance with a Missing Cell

Hypothesis	SS_R	$SS_F - SS_R$	R^2	$R_F^2 - R_R^2$
H_1: $b_1 + b_2 = b_4 + b_5$	26.92	1.82	.02985	.00202
H_2: $\frac{b_1 + b_2 + b_3}{3} = \frac{b_4 + b_5}{2}$	23.72	5.02	.02629	.00558
H_3: $\frac{15b_1 + 10b_2}{25} = \frac{9b_4 + 6b_5}{15}$	23.70	5.04	.02628	.00559
H_4: $\frac{15b_1 + 10b_2 + 5b_3}{30} = \frac{9b_4 + 6b_5}{15}$	21.23	7.51	.02354	.00833
H_5: $b_1 + b_4 = b_2 + b_5$	25.68	3.06	.02848	.00339
H_6: $\frac{b_1 + b_4}{2} = \frac{b_2 + b_5}{2} = b_3$	19.68	9.06	.02182	.01005
H_7: $\frac{15b_1 + 9b_4}{24} = \frac{10b_2 + 6b_5}{16}$	28.24	.50	.03131	.00056
H_8: $\frac{15b_1 + 9b_4}{24} = \frac{10b_2 + 6b_5}{16} = b_3$	23.10	5.64	.02562	.00625
H_9: $b_1 - b_4 = b_2 - b_5$	10.68	18.06	.01184	.02003

numbers in each cell; 3) disproportionate numbers in each cell; and 4) at least one missing cell.

Addressing the four situations, the solution described by Jennings (1967) and shown to be computationally equivalent to the unadjusted main effects solution by Williams (1972) is robust in that it addresses likely hypotheses regarding the cell means in all four instances. The full rank model, described by Timm and Carlson (1975) can be criticized as addressing likely hypotheses of interest only for the equal cell frequency situation; the hypotheses tested in the proportionate case may very well deviate from those a researcher is likely to be most interested in. The hypotheses that are tested in the missing cells case do not appear to have any likely contrasts among the cells that address usual analysis of variance questions. It is of course possible that the hypotheses tested by the full rank model are of interest to the researcher. However, as a general data-analytic tool, the full rank model as described by Timm and Carlson would seem to lack the robustness needed to suggest itself to the statistically unsophisticated user.

REFERENCES

elbaum, M. I. and Cramer, E. M. Some problems in the non-orthogonal analysis of variance. Psychological Bulletin, 1974, 81, 335-343.

en, J. Multiple regression as a general data - analytic system. Psychological Bulletin, 1968, 70, 426-443.

nings, E. Fixed effects analysis of variance by regression analysis. Multivariate Behavioral Research, 1967, 2, 95-108.

rall, J. E. and Spiegel, D. H. Concerning least squares analysis of experimental data. Psychological Bulletin, 1969, 72, 311-322.

rall, J. E., Spiegel, D. H. and Cohen, J. Equivalence of orthogonal and nonorthogonal analysis of variance. Psychological Bulletin, 1975, 82, 182-186.

rle, S. R. Linear models. New York: Wiley, 1971.

rle, S. R., Speed, F. M. and Henderson, H. V. Some computational and model equivalences in analyses of variance of unequal-subclass-numbers data. The American Statistician, 1981, 35, 16-33.

ed, F. A. and Hocking, R. R. The use of R () - Notation with unbalanced data. The American Statistician, 1976, 30, 30-33.

m, N. H. and Carlson, J. E. Analysis of variance through full rank models. Multivariate Behavioral Research: Monograph, 1975, 75-1.

llams, J. D. Two way fixed effects analysis of variance with disproportionate cell frequencies. Multivariate Behavioral Research, 1972, 7, 67-83.

llams, J. D. Regression analysis in education research. New York: MSS Publishing Corp., 1974.

llams, J. D. Full rank and non-full rank models with contrast and binary coding systems for two way disproportionate cell frequencies analysis. Multiple Linear Regression Viewpoints, 1977, 8, No. 1, 1-18.

llams, J. D. A note on proprotional cell frequencies in a two-way classification. Multiple Linear Regression Viewpoints (this issue).