MULTIPLE LINEAR REGRESSION VIEWPOINTS
Volume 11, Number 1, September, 1981

A NOTE ON PROPORTIONAL CELL FREQUENCIES IN A TWO-WAY CLASSIFICATION

JOHN D. WILLIAMS
THE UNIVERSITY OF NORTH DAKOTA

Abstract - A proportional, but non-equal two-way data set is analyzed, comparing the full rank model solution to the fitting constants, hierarchical model and unadjusted main effects solutions. The latter three models yield identical results; the full rank model, yielding different results, is shown to be testing different main effect hypotheses.

Several writers have explored different approaches to the analysis of disproportionate cell frequencies data in a two-way (or higher) layout. One such solution, the "full rank model" solution, as described by Timm and Carlson (1975) has been purported to be the "best" solution to the traditional two-way design; Overall, Spiegel and Cohen (1975) appear to concur in this position. One rather interesting circumstance is that, for proportional, but non-equal cell entries, the full rank model solution fails to yield an additive solution. While this problem has been pointed out before (see Overall and Spiegel, 1969; also, Williams 1977), a simple example together with the sums of squares should be helpful.

Consider the following data (taken from Williams, 1974, p. 77):

ACT Scores

Sex	College		
	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		

Sex	College		
	Arts and Sciences	Education	Engineering
Female	19	23	27
	17	29	24
	17	21	22
	16	17	
	18	15	
	27	13	
	14		
	15		
	16		

Several different procedures could be effected to code the data or obtain suitable solutions. Because contrast coding is an effective means to obtain a solution for the full rank model approach of Timm and Carlson, contrast coding is used for the other solutions as well. In addition to the Y (criterion) variable, five other variables can be 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;

$X_4 = X_1 \cdot X_2$; and

$X_5 = X_1 \cdot X_3$.

Six models can be defined:

$$Y = b_0 + b_1X_1 + e_1, \quad (1)$$

$$Y = b_0 + b_2X_2 + b_3X_3 + e_2, \quad (2)$$

$$Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + e_3, \quad (3)$$

$$Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + b_5X_5 + e_4, \quad (4)$$

$$Y = b_0 + b_1X_1 + b_4X_4 + b_5X_5 + e_5, \quad (5) \text{ and}$$

$$Y = b_0 + b_2X_2 + b_3X_3 + b_4X_4 + b_5X_5 + e_6. \quad (6)$$

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 other b's are specific to an equation); the e's are error terms associated with each equation.

Table 1 shows the sums of squares for the various approaches to analyzing the data on sex and college.

It can be noticed in Table 1 that the main effect for sex and sex independent of college are equal, but are unequal to sex independent of college and interaction; a similar result occurs for the college effect. If either an unadjusted main effects solution, a fitting constants solution or a hierarchical model are completed, an additive solution is found. See Table 2. (The terminology for type of solution is the same as in Williams, 1972).

However, if a full rank model solution (as suggested by Timm and Carlson) is executed a non-additive model results. See Table 3.

The difference in the solutions shown in Tables 2 and 3 are that different hypotheses are being tested. It can be shown (see Williams, 1977) that the solution in Table 2 corresponds to the one proposed by Jennings (1967); the hypothesis for sex differences is given by (in terms of sample means)

$$\frac{n_1\bar{Y}_1 + n_2\bar{Y}_2 + n_3\bar{Y}_3}{n_1 + n_2 + n_3} = \frac{n_4\bar{Y}_4 + n_5\bar{Y}_5 + n_6\bar{Y}_6}{n_4 + n_5 + n_6} \qquad (7)$$

where the n's and $\bar{Y}$'s correspond to the cells in the two way layout. $\bar{Y}_1$ is the mean of males in arts and science, $\bar{Y}_2$ is the mean of males in education, and $\bar{Y}_3$ is the mean of males in engineering; means for females ($\bar{Y}_4$, $\bar{Y}_5$, $\bar{Y}_6$) are similarly defined.

Since proportionality holds, the numerator and denominator of the left side of equation 7 can be multiplied by $\frac{n_4}{n_1}$ (or by $\frac{n_5}{n_2}$ or $\frac{n_6}{n_3}$ or any combination thereof, since the proportion is the same):

$$\frac{\frac{n_4}{n_1}n_1\bar{Y}_1 + \frac{n_4}{n_1}n_2\bar{Y}_2 + \frac{n_4}{n_1}n_3\bar{Y}_3}{\frac{n_4}{n_1}(n_1 + n_2 + n_3)} = \frac{n_4\bar{Y}_4 + n_5\bar{Y}_5 + n_6\bar{Y}_6}{n_4 + n_5 + n_6} \qquad (8)$$

Since $\frac{n_4}{n_1} = \frac{n_5}{n_2} = \frac{n_6}{n_3}$, equation 8 can be simplified:

Table 1

Two-Way Solution for Proportionate Cell Frequency

Source of Variation	df	SS	MS	F
Sex	1	$SS_1 = .14$	.14	.01
Sex (Independent of College)	1	$SS_3 - SS_2 = 49.24 - 49.10 = .14$	.14	.01
Sex (Independent of College and Interaction)	1	$SS_4 - SS_5 = 107.42 - 95.36 = 12.06$	12.06	.57
College	2	$SS_2 = 49.10$	24.55	1.16
College (Independent of Sex)	2	$SS_3 - SS_1 = 49.24 - .14 = 49.10$	24.55	1.16
College (Independent of Sex and Interaction)	2	$SS_4 - SS_6 = 107.42 - 34.91 = 72.51$	36.26	1.72
Interaction	2	$SS_4 - SS_3 = 107.42 - 49.24 = 58.18$	29.09	1.38
Within	42	$SS_{DEV_4} = 885.83$	21.09	
Total	47	$SS_T = 993.25$		

Table 2

Summary Table for the Unadjusted Main Effects Solution, Fifting Constants Solution and Hierarchical Model With Proportional Data

Source of Variation	df	SS	MS	F
Sex	1	.14	.14	.01
College	2	49.10	24.55	1.16
Interaction	2	58.18	29.09	1.38
Within	42	885.83	21.09	
Total	47	993.25		

Table 3

Summary Table For Full Rank Model Solution With Proportional Data

Source of Variation	df	SS	MS	F
Sex (Independent of College & Interaction)	1	12.06	12.06	.57
College (Independent of Sex & Interaction)	2	72.51	36.26	1.72
Interaction	2	58.18	29.09	1.38
Within	42	885.83	21.09	
Total	47	1028.58 ≠	993.25	

$$n_4\bar{Y}_1 + n_5\bar{Y}_2 + n_6\bar{Y}_3 = n_4\bar{Y}_4 + n_5\bar{Y}_5 + n_6\bar{Y}_6. \quad (9)$$

While equation 9 could be expressed in several alternative forms, it is clear that the number of members in a cell are incorporated into the hypothesis. The full rank model solution addresses a different hypothesis. For the sex effect, the hypothesis tested is (in terms of sample means)

$$\frac{\bar{Y}_1 + \bar{Y}_2 + \bar{Y}_3}{3} = \frac{\bar{Y}_4 + \bar{Y}_5 + \bar{Y}_6}{3}. \quad (10)$$

Note that equation 10 tests a hypothesis regarding means that suggest all groups have the same number of members, even if they do not. The actual mean for males is 19.33 and for females is 19.44. The cell means are $\bar{Y}_1 = 19.53$, $\bar{Y}_2 = 18.70$, $\bar{Y}_3 = 20$, $\bar{Y}_4 = 17.67$, $\bar{Y}_5 = 19.67$ and $\bar{Y}_6 = 24.33$. Thus $\frac{\bar{Y}_1 + \bar{Y}_2 + \bar{Y}_3}{3} = 19.41$ and $\frac{\bar{Y}_4 + \bar{Y}_5 + \bar{Y}_6}{3} = 20.55$. It is this writer's opinion that the additive solution is more likely to be of interest than the solution found through the full rank model solution suggested by Timm and Carlson.

REFERENCES

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

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

Overall, J.E., Spiegel, D.H., and Cohen, J. Equivalence of orthogonal and nonorthogonal analysis of variance. Psychological Bulletin, 1975, 82, 185-186.

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

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

Williams, J.D. Regression analysis in educational research. New York: MSS Publishing Corp., 1974.

Williams, 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.

MULTIPLE LINEAR REGRESSION VIEWPOINTS
Volume 11, Number 1, September, 1981

UNDERSTANDING PARTIAL REGRESSION COEFFICIENTS IN THE PRESENCE OF CORRELATED REGRESSORS

JEFFREY K. SMITH AND LINDA F. LEARY
RUTGERS UNIVERSITY

ABSTRACT

The interpretation of partial regression coefficients in the presence correlated regressors causes difficulty for students in the social sciences. nce correlation among regressors is the typical case in the social sciences, is presents a considerable instructional problem. This article presents an planation of the partial regression coefficient in the presence of correlated gressors that is a simple and direct extension of the case where regressors e mutually orthogonal. The interpretation presented emphasizes the rela- nship between the partial regression coefficient and the simple regression efficient. An example using SAS computer package is provided.

troduction

The extension of the principles and techniques of simple linear gression to multiple linear regression frequently results in confusion and isunderstanding for students in the social sciences. The major problem ea concerns the understanding of the regression coefficients when regres- s are moderately correlated. In most texts on regression analysis (Cohen d Cohen, 1975; Draper and Smith, 1966; Kerlinger and Pedhazur, 1973) e extension from simple to multiple regression is discussed via the special se where the regressors are uncorrelated. Pedagogically, this is appro- riate since it requires the introduction of a minimum of new concepts. owever, in the actual analysis of data in the social sciences, correlated gressors are far more the rule than the exception. Unfortunately, it is the conceptual leap from independent regressors to correlated regressors at there exists the greatest lack of clarity in explanation. An example of is confusion is the belief displayed not only be beginning students, but practicing researchers, that the order of entry of the variables into a

This work was supported in part by a grant from the Rutgers University Research Council. The authors appreciate editorial comments from an anonymous reviewer.

stepwise regression procedure affects the resultant regression weights when the full model is estimated! This confusion is exacerbated by the treatment of stepwise regression output in statistical computer packages such as SPSS (although, to the credit of the SPSS authors, they provide the best explanation we have found to date on the problem of correlated regressors) (Nie, et al., 1975).

The purpose of this paper is to present a lucid explanation of partial regression coefficients in the presence of correlation among the regressors. Our goal is to bridge the gap between a purely verbal explanation such as ". . . the increase in Y for a unit increase in X holding all other variables constant. . ." and a purely mathematical explanation such as:

$$B_{Y1.2} = \frac{r_{Y1} - r_{Y2}r_{12}}{1 - r_{12}^2} \left(\frac{S_Y}{S_1}\right)$$

Although both of these approaches are technically correct, neither provides a particularly good intuitive understanding of what is involved in multiple regression with correlated regressors.

Simple and Partial Regression Coefficients

It is our experience that the simple regression coefficient is readily comprehended by students as they approach multiple regression, and that an explanation of the partial regression coefficient in terms of a simple regression coefficient is heuristically appealing to students. Such a transition is clear and direct in the case of mutually orthogonal regressors. This multiple regression setting reduces to a series of simple regression equations (as in Draper and Smith, 1966, pp. 107-115). That is, the partial regression coefficient is identical to what it would be in a simple regression.

Our purpose here is to show that a similar reduction can be used even when regressors are correlated. The presentation below demonstrates how this would be done. It might reasonably follow the mutually orthogonal setting in a regression course.

Consider a regression with dependent variable Y, and three moderately correlated regressors X_1, X_2, and X_3:

(1) $$\hat{Y} = B_1X_1 + B_2X_2 + B_3X_3 + B_0$$

Since the regressors are correlated, it is obvious that the coefficient B_1 will not have the same value as a simple regression coefficient from the regression of Y on X_1 (alone). However, B_1 will be identical to the coefficient obtained from a simple regression of Y on the residuals of X_1 (say, X_1') after the collinearity with X_2 and X_3 has been removed. This can be accomplished by regressing X_1 on X_2 and X_3:

$$\hat{X}_1 = a_2X_2 + a_3X_3 + a_0$$

then $$\hat{X}_{1i}' = X_{1i} - \hat{X}_{1i} .$$

The same procedure is followed for X_2 and X_3. A new equation:

(2) $$\hat{Y} = B_1'X_1' + B_2'X_2' + B_3'X_3' + B_0'$$

can be shown to yield exactly the same regression coefficients as equation (1). That is, $\underset{\sim}{B}' = \underset{\sim}{B}$. The pedagogical advantage gained by creating equation (2) is that the $\tilde{X}_1$'s are mutually orthogonal and the B_1's can be understood as in the mutually orthogonal case. Thus, the partial regression coefficient is the simple regression coefficient of Y on the residuals of X_1 after the effects of X_2 and X_3 have been removed from X_1.

The utility of this approach to understanding regression coefficients is that it allows the student to link his comprehension of the partial regression coefficient to the firmer ground of the simple regression coefficient. This is particularly useful when such concepts as suppressor variables, multicollinearity, and shrinkage in r-squared are discussed.

An Example

An example of this approach with three regressors using the SAS statistical package is presented below:

```
(JCL)

DATA SAMPA;

INPUT  Y  X1  X2  X3;

CARDS;

   (insert data)

PROC GLM; MODEL  Y = X1  X2  X3;

PROC GLM; MODEL  X1 = X2  X3;
```

```
OUTPUT OUT = SAMPB  RESIDUAL = RESID;
DATA SAMPC; MERGE SAMPA SAMPB;
PROC GLM; MODEL Y = RESID;
DATA SAMPA;
PROC GLM; MODEL X2 = X1  X3;
OUTPUT OUT = SAMPD  RESIDUAL = RESID;
DATA SAMPE;  MERGE SAMPA  SAMPD;
PROC  GLM;  MODEL  Y = RESID;
DATA SAMPA;
PROC  GLM;  MODEL  X3 = X1 X2;
OUTPUT OUT = SAMPF  RESIDUAL = RESID;
DATA SAMPG; MERGE SAMPA  SAMPF;
PROC GLM;  MODEL  Y = RESID;
//
```

The first PROC GLM statement results in the standard multiple regression output for the full model. The second PROC GLM regresses X_1 on the remaining independent variables and calculates the residuals, while the third PROC GLM performs the regression of Y on the residual of X_1. Students can now verify that the regression coefficient for residuals is identical to that for X_1 in the original model. The remaining PROC GLM statements calculate the coefficients for X_2 and X_3 in the same manner. Although the layout for calculating all regression coefficients is presented here for completeness, calculation of only one or two of these may be sufficient for instruction.

References

n, J. and Cohen, P. (1975) Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

er, N.R. and Smith, H. (1966) . Applied Regression Analysis. New York: John Wiley and Sons.

nger, F.N. and Pedhazur, E.J. (1973) Multiple Regression in Behavioral Research. New York: Holt, Rinehart, and Winston, Inc.

N.H. et al. (1975) Statistical Package for the Social Sciences, Second Edition. New York: McGraw-Hill Book Company.