ULTIPLE LINEAR REGRESSION VIEWPOINTS
LUME 14, NUMBER 1, SPRING 1985

Multiple Comparisons Via Multiple Linear Regression: Learning the Obvious Takes Time

John D. Williams

The University of North Dakota

Perhaps a best starting point is at the beginning--the beginning of my involvement in multiple linear regression ala Ward, Bottenberg and Jennings. A presession to the AERA annual meeting in New York in 1967 was my first exposure to this type of analysis. I must admit something less than being fully enthralled with their ideas at the time. Despite computer accessibility for the five day workshop, I didn't actually run any programs. To me it was just a new fad. When getting back to Grand Forks (N.D.) I did feel some pangs of conscience and tried running a simple ANOVA by regression. The problem was a three group situation; I was trying to run:

$$Y = b_1X_1 + b_2X_2 + b_3X_3 + e_1 \qquad [1]$$

where

X_1 = 1 if a member of group 1, 0 otherwise,

X_2 = 1 if a member of group 2, 0 otherwise,

X_3 = 1 if a member of group 3, 0 otherwise,

b_1, b_2, b_3 are regression coefficients,

Y = the criterion score, and

e_1 = the error in prediction with this model.

The program used at the presession was DATRAN, a forerunner of LINEAR (which of course, I didn't actually use). The program available to me back

Paper presented at the American Educational Research Association, Chicago, April, 1985 in North Dakota was a stock IBM program; in retrospect, such stock programs typically have automatic inclusion of a unit vector (or constant). Well, what happened next is both a descriptor of something about my personality (stubborn) or possibly lack of intelligence (slow). On a daily basis for seven weeks, (that's 35 times) I unsuccessfully tried running the program exactly as shown in equation 1 without any change. I thought possibly there was something wrong with the computer or the program; never did it cross my mind that I might have made a conceptual error. Finally, I started monkeying with the input (I was convinced the stuff in Bottenberg and Ward, 1963, was wrong). Well, I finally made the right mistake, and the program actually worked correctly. One form of that mistake is as follows:

$$Y = b_0 + b_1X_1 + b_2X_2 + e_1. \qquad [2]$$

The difference between equation 2 and equation 1 ostensibly is the exclusion of b_0 in equation 1 and the exclusion of b_3X_3 in equation 2.

Also, I now know that equations 1 and 2 are reparameterizations of one another. There are also some other "obvious" things about equation 2; it took me only four years to discover some of the obvious.

Equation 2 can allow not only a simple ANOVA, but also describes some important aspects of Dunnett's (1955) test (Williams, 1971); b_0 is not just a constant, but is equal to $\overline{Y}_3$, the so-called left out group. Also, $b_1 = \overline{Y}_1 - \overline{Y}_3$ and $b_2 = \overline{Y}_2 - \overline{Y}_3$. Equation 2 could be rewritten as:

$$Y = \overline{Y}_3 + (\overline{Y}_1 - \overline{Y}_3)X_1 + (\overline{Y}_2 - \overline{Y}_3)X_2 + e_1. \qquad [3]$$

The tests of the regression coefficients $b_1 = \overline{Y}_1 - \overline{Y}_3$ and $b_2 = \overline{Y}_2 - \overline{Y}_3$ are identically equal to the t values in Dunnett's test.

In addition to an ANOVA, other simple designs can be shown in a regression lay-out, such as the analysis of covariance, the t test, and treatments x subjects designs. The use of equations such as equation 2 to complete these designs was shown in Williams (1970). As usual, I had no idea at the time of the relationship to multiple comparisons. In some ways, the relationships are so simple and direct that it gives me cause for some degree of humility to remember how long it took me to discern the obvious again.

Through the use of full and restricted models, a process to test comparisons equivalent to Tukey's (1953) test was shown (Williams, 1974a). With three groups, beginning with equation 1, $Y = b_1X_1 + b_2X_2 + b_3X_3 + e_1$. Now suppose the test of $\overline{Y}_2 = \overline{Y}_3$ is of interest. In terms of the regression coefficients $b_2 = b_3$ is the appropriate restriction. Then $Y = b_1X_1 + b_2X_2 + b_2X_3 + e_2$ or

$$Y = b_1X_2 + b_2(X_2 + X_3) + e_2.$$

Let $V_1 = X_2 + X_3$; then

$$Y = b_1X_1 + b_2V_1 + e_2. \qquad [4]$$

Equation 4 can be reparameterized so that the unit vector (constant term) is reintroduced by excluding either X_1 or V_1. Excluding X_1 yields:

$$Y = b_0 + b_2V_2 + e_2. \qquad [5]$$

Testing $t = \sqrt{F} = \sqrt{\dfrac{(R_2^2 - R_5^2)/1}{(1 - R_2^2)/(N - K)}}$ yields a t appropriate to testing $\overline{Y}_2$ to $\overline{Y}_3$.

On the other hand, there is an easy way to run Tukey's test by regression. All that is necessary is the <u>set</u> of reparameterizations of equation 1:

$$Y = b_0 + b_1X_1 + b_2X_2 + e_1, \qquad [2]$$

$$Y = b_0 + b_1X_1 + b_3X_3 + e_1, \qquad [6]$$

and $Y = b_0 + b_2X_2 + b_3X_3 + e_1. \qquad [7]$

Here, the test of the computed t values is identical to a similar test for Tukey's test. (It took a full three years after doing the same thing with Dunnett's test to realize that Tukey's test could be accomplished through successive psuedo-Dunnett's tests). One complication is that most published studentized range tables are in terms of q, rather than in terms of testing the regression coefficients for significance. A table showing a direct solution using tests on the (partial) regression weights is given in Williams (1976, 1980).

In that I routinely would find all simple reparameterizations of an equation for an ANOVA solution, taking seven years to discover the obvious says something.

Two-Way Disproportionate ANOVAs

The two-way analysis of variance with disproportionate cell frequencies has been discussed in many different publications; Bottenberg and Ward (1963) showed a regression solution for the general case, and Jennings (1967) concentrated on the disproportionate situation. To be honest, I had a lot of trouble understanding the Jennings article, so I tried to go about doing what I could understand from the original Bottenberg and Ward presentation. One aspect of Bottenberg, Ward and Jennings in their various writings is a concern for explicitly stating exactly the hypothesis being tested through the use of a restriction on the regression coefficients. This aspect has been both a blessing and a curse; it is a blessing in the sense that the approach allows a precise methodology. It is a curse in that users are often at a disadvantage because of the cognitive completixity and relative mathematical sophistication required in comparison to traditional analysis of variance methodologies. It could be arqued that a middle ground can be attempted; to some degree, that middle ground was something I tried to do (Williams, 1974b).

As an example of a two-way ANOVA with disproportionate cell frequencies the following data set was originally published in Williams (1972):

Data for Disproportionate Two-Way Analysis of Variance

Effect	Effect		
	B_1	B_2	B_3
A_1	8 6 4	1 1	6 2
A_2	10	7 5 4 4 3	10 9 7 5 4

The solution given (1972) that was meant to simplify the process was to form four models:

$$Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + b_5X_5 + e_3; \qquad [8]$$

where

X_1 = 1 if from an individual in cell 1 (row 1, column 1), 0 otherwise;

X_2 = 1 if from an individual in cell 2 (row 1, column 2), 0 otherwise;

X_3 = 1 if from an individual in cell 3 (row 1, column 3), 0 otherwise;

X_4 = 1 if from an individual in cell 4 (row 2, column 1), 0 otherwise;

X_5 = 1 if from an individual in cell 5 (row 2, column 2), 0 otherwise;

and b_0 to b_5 are regression coefficients for this model.

$$Y = b_6 + b_7X_7 + e_4; \qquad [9]$$

where

X_7 = 1 from an individual in row 1, 0 otherwise and

b_6, b_7 are regression coefficients for this model.

$$Y = b_8 + b_9X_9 + b_{10}X_{10} + e_5; \qquad [10]$$

where

$X_9 = 1$ if from an individual in column 1, 0 otherwise;

$X_{10} = 1$ if from an individual in column 2, 0 otherwise; and

b_8, b_9 and b_{10} are regression coefficients for this model.

$$Y = b_{11} + b_{12}X_7 + b_{13}X_9 + b_{14}X_{10} + e_6. \qquad [11]$$

Now a solution in terms of sums of squares can be given as follows:

From: equation 8, $SS_{ATTRIBUTABLE} = 80.80$;

$SS_{DEVIATION} = 51.20$;

equation 9, $SS_{ATTRIBUTABLE} = 20.36$;

equation 10, $SS_{ATTRIBUTABLE} = 37.43$ and

equation 11, $SS_{ATTRIBUTABLE} = 80.25$.

This information could be used to construct a fitting contants solution or a hierarchical solution (Cohen, 1968) or the solution described by Jennings (1967); although Jennings laboriously goes through the process of testing hypotheses through restrictions on a reparameterization of the full model:

$$Y = b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + b_5X_5 + b_6X_6 + e_3. \qquad [12]$$

This model corresponds to equation 8, except that the unit vector is omitted (b_0) and the sixth cell is represented through b_6X_6. Because my solution, while it coincides with Jennings, can be addressed without adjusting the sums of squares as must be done for a fitting constants solution or a hierarchical solution, I called this solution the "unadjusted main effects" solution--in retrospect, a poor choice of names. It was called this because of the means of extracting the sums of squares--but its usefulness is because it corresponds to the Jennings solution. That by the way, is another story--I spent an hour and a half convincing Earl that my solution gave the same results as his; at first he was skeptical. Finally, he accepted that, "computationally, their respective sums of squares was the same," but thought only people such as myself who understand both approaches and used my approach as a computational short cut should use it; if you didn't know what hypotheses were being tested, you probably shouldn't use it. I thought Earl was being a little harsh back in 1972, but today I'm coming closer to agreement with that position.

In particular, it could be noted that the so-called "full rank model" as described by Timm and Carlson (1975), and which in fact they describe using my (1972) data set, has no better claim to being a full rank model solution than Jennings (1967); the hypotheses tested by these and other approaches are considered in Williams (1977a). It is unfortunate that the Timm and Carlson (1975) solution might be seen by some as "standard practice" or "state of the art". The issue really is, which hypotheses are of greatest interest? If the Timm and Carlson hypotheses are truly of the greatest interest, they can be addressed via the Bottenberg and Ward approach.

A summary table that computationally tests hypotheses proportional to cell frequencies such as proposed by Jennings can easily be formed from the information from equations 8, 9, 10 and 11:

$SS_{ROWS} = 20.36$; $SS_{COLS} = 37.43$;

$SS_{RC} = 80.80 - 80.25 = .55$;

$SS_{within} = 51.20$. The summary table is as follows:

Table 1

Summary Table for Two-Way
Disproportionate Cell Frequencies

Source of Variation	df	SS	MS	F
Rows	1	20.36	20.36	4.77
Columns	2	37.43	18.72	4.38
R X C	2	.55	.28	.07
Within	12	51.20	4.27	

In regard to multiple comparisons in a two-way layout, equation 12 is an appropriate starting point. The number and type of comparisons (contrasts) would be important for deciding on the type of test (Dunnett's, Tukey's, Scheffe's, 1959, and Dunn's, 1961). As an example of constructing contrast to test a hypothesis of interest, suppose the researcher wants to compare column 1 to column 2, weighing the cells by their size, the hypothesis, in terms of sample means, is:

$$\frac{3\bar{Y}_1 + 1\bar{Y}_4}{4} = \frac{2\bar{Y}_2 + 5\bar{Y}_5}{7} .$$

In terms of the regression coefficients,

$$\frac{3b_1 + b_4}{4} = \frac{2b_2 + 5b_2}{7}$$

Unraveling and solving for b_1 yields: $b_1 = 8/21b_2 + 20/21b_5 - 7/21b_4$.

Substituting this restriction into equation 12 yields:

$$Y = (8/21b_2 + 20/21b_5 - 7/21b_4)X_1 + b_2X_2 + b_3X_3 + b_4X_4 + b_5X_5 + b_6X_6 + e_7; \qquad [13]$$

or

$$Y = b_2(X_2 + 8/21X_1) + b_3X_3 + b_4(X_4 - 7/21X_1) + b_5(X_5 + 20/21X_1) + b_6X_6 + e_7. \qquad [14]$$

Reparameterization with $b_6 = 0$ yields:

$$Y = b_0 + b_2(X_2 + 8/21X_1) + b_3X_3 + b_4(X_4 - 7/21X_1) + b_5(X_5 + 20/21X_1) + e_7. \qquad [15]$$

Equation 14 can be used in programs where unit vectors can be ommitted. Its reparameterization, equation 15, is useful when a unit vector is automatically incorporated into a regression solution. Equations 8 and 12 (full models) yield $R_F^2 = .61212$. Equations 14 and 15 (restricted models) yield $R_F^2 = .38544$. Then:

$$t = \sqrt{F} = \sqrt{\frac{(R_F^2 - R_R^2)/1}{(1 - R_F^2)/12}} = 2.648.$$

This t value should be tested against an appropriate table depending upon the type and number of total comparisons considered by the researcher.

This approach to multiple comparisons is probably much closer to the approach of Jennings and Bottenberg and Ward than I would have considered 10 to 15 years ago. Additional considerations regarding multiple comparisons in the two-way analysis of variance ban be found in Williams (1980).

Multiple Comparisons in the Analysis of Covariance

Students would often ask questions such as, "How do you do multiple comparisons on adjusted means in the analysis of covariance?" I've often been impressed with questions students ask; I'm sure they've been less impressed with at least some of my answers. Well, for several years, I didn't have any good answer to the aforementioned question (other than, "That's a good question.") and as the answer finally came to me, there was far more embarrassment than awe. The "answer" had been on the printouts that I'd been using for years. In a nutshell, it was simply the test of signifiance for the group partial regression weights in a full model. An example of a solution for this problem was taken from Williams (1979).

Table 2 is taken from Williams (1974b, p. 104 and 109). In Table 2, X_1 is a binary variable for membership in group 1, X_2 is a binary variable for membership in group 2 and X_3 is similarly a binary variable for membership in group 3 and X_4 represents a pretest score; the Y value represents a posttest score.

Table 2

Data for the Analysis of Covariance

Y	X_1	X_2	X_3	X_4
35	1	0	0	12
27	1	0	0	17
32	1	0	0	13
29	1	0	0	10
27	1	0	0	8
38	0	1	0	29
25	0	1	0	12
36	0	1	0	17
25	0	1	0	22
31	0	1	0	15
27	0	0	1	17
35	0	0	1	22
19	0	0	1	10
17	0	0	1	8
32	0	0	1	13

Under the assumption of a single regression line on the covariate (the pretest, X_4) an analysis of covariance can be accomplished with two linear models:

$$Y = b_0 + b_1X_1 + b_2X_2 + b_4X_4 + e_8, \qquad [16]$$

and

$$Y = b_0 + b_4X_4 + e_9. \qquad [17]$$

In that a large part of the print-out regarding equation 16 is useful, the print-out is reproduced in Table 3.

The usual analysis of covariance can be completed by using:

$$F = \frac{(R_F^2 - R_R^2)/(g - 1)}{(1 - R_2^2)/(N - C - g)} = \frac{(.61959 - .47476)/2}{(1 - .61959)/11} = 2.09,$$

which for df = 2, 11, $p > .05$.

In equation 16 the X_3 variable has been omitted. Thus $b_1 = \overline{Y}_1\text{adj} - \overline{Y}_3\text{adj}$ and $b_2 = \overline{Y}_2\text{adj} - \overline{Y}_3\text{adj}$. To find the adjusted means, the following equations can be used:

$$\overline{Y}_3\text{adj} = b_0 + b_4X_4 = 15.36 + .76(15) = 26.76;$$

$$\overline{Y}_1\text{adj} = b_1 + Y_3\text{adj} = 5.52 + 26.76 = 32.28; \text{ and}$$

$$\overline{Y}_2\text{adj} = b_2 + Y_3\text{adj} = 3.20 + 27.76 = 29.96.$$

The adjusted values agree with those originally given by Williams (1974b, p. 106), though the method shown here is simplified somewhat.

More importantly, the standard error of the regression coefficients corresponding to X_1 and X_2 are respectively equal to the standard errors for comparing $\overline{Y}_1\text{adj}$ to $\overline{Y}_3\text{adj}$ and $\overline{Y}_2\text{adj}$ to $\overline{Y}_3\text{adj}$. Thus, the computed t values given in Table 3 are directly usable in whichever multiple comparison procedure the researcher prefers. The use of Dunnett's (1955), Tukey's (1953), Dunn's (1961) and Scheffe's (1959) tests are described in a regression format using computed t values in Williams (1976, 1980). Were there interest in comparing $\overline{Y}_1\text{adj}$ to $\overline{Y}_2\text{adj}$, a model of the form:

$$Y = b_0 + b_1X_1 + b_3X_3 + b_4X_4 + e_8 \qquad [18]$$

could be used, with focus on the computed t value for the X_1 variable.

Table 3

Print-Out for Equation 16

Variable	Mean	Standard Deviation	Correlation X vs Y	Regression Coefficient	Std. Error of Reg. Coef.	Computed T Value
4	15.00	5.85	0.689	0.76	0.22783	3.33582
1	0.33	0.48	0.039	5.52	2.73396	2.01905
2	0.33	0.48	0.398	3.20	2.92653	1.09345
Dependent Y	29.66	6.12				
INTERCEPT	15.36					

MULTIPLE CORRELATION 0.78714

STD. ERROR OF ESTIMATE 4.26230

MULTIPLE CORRELATION SQUARED 0.61959

ONE MINUS MULTIPLE CORRELATION SQD. 0.38041

Analysis of Variance for the Regression

Source of Variation	Degrees Of Freedom	Sum of Squares	Mean Squares	F Value
Attributable to Regression	3	325.49	108.497	5.972
Deviation from Regression	11	199.84	18.167	
Total	14	525.33		

f course, multiple covariates and/or more complex comparisons can be ncorporated; multiple covariates can be incorporated without adding too uch complexity to the solution. The remarkable thing is that the solution to multiple comparisons for the analysis of covariance is easily chieved.

Multiple Comparisons in Repeated Measure Designs

Again, the impetus (to me) for interest in multiple comparisons in epeated measures designs in general, and treatments x subjects designs n particular comes from students. Students would ask, "O.K., so now e can do a treatments x subjects design by regression. How do we run nultiple comparisons?" Since they asked the question long before I had any suitable answer, a question might be asked, "What answer did I give?" To quote both the famous and infamous (e.g. Steve Martin and John Mitchell), "I forgot." Considering that that answer can be as simple as, "It's right there on your printout," I won't dwell anymore on why it took so long.

Multiple Comparisons for Treatments X Subjects Designs

To consider multiple comparisons for treatments x subjects designs (or repeated measure designs) an example taken from Chapter 7 of Williams (1974b, p. 56) is used; see Table 4.

Table 4

Three Treatment Methods of Paired-Associate Learning with Educable Mentally Retarded Subjects

Subject	Treatment One	Treatment Two	Treatment Three
1	18	27	15
2	17	24	14
3	14	13	12
4	5	8	6
5	11	14	10
6	9	12	8
7	14	16	15
8	12	17	9
9	22	21	16
10	10	18	15

The information in Table 4 can be placed in a tabular form suitable for use in regression format; see Table 5.

Table 5

Illustration of Design Matrix for Treatments X Subjects Designs

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}
3	1	0	0	1	0	0	0	0	0	0	0	0	0	60
7	0	1	0	1	0	0	0	0	0	0	0	0	0	60
5	0	0	1	1	0	0	0	0	0	0	0	0	0	60
7	1	0	0	0	1	0	0	0	0	0	0	0	0	55
4	0	1	0	0	1	0	0	0	0	0	0	0	0	55
4	0	0	1	0	1	0	0	0	0	0	0	0	0	55
14	1	0	0	0	0	1	0	0	0	0	0	0	0	39
13	0	1	0	0	0	1	0	0	0	0	0	0	0	39
12	0	0	1	0	0	1	0	0	0	0	0	0	0	39
5	1	0	0	0	0	0	1	0	0	0	0	0	0	19
8	0	1	0	0	0	0	1	0	0	0	0	0	0	19
6	0	0	1	0	0	0	1	0	0	0	0	0	0	19
11	1	0	0	0	0	0	0	1	0	0	0	0	0	35
14	0	1	0	0	0	0	0	1	0	0	0	0	0	35
10	0	0	1	0	0	0	0	1	0	0	0	0	0	35
9	1	0	0	0	0	0	0	0	1	0	0	0	0	29
12	0	1	0	0	0	0	0	0	1	0	0	0	0	29
8	0	0	1	0	0	0	0	0	1	0	0	0	0	29
14	1	0	0	0	0	0	0	0	0	1	0	0	0	45
16	0	1	0	0	0	0	0	0	0	1	0	0	0	45
15	0	0	1	0	0	0	0	0	0	1	0	0	0	45
12	1	0	0	0	0	0	0	0	0	0	1	0	0	38
17	0	1	0	0	0	0	0	0	0	0	1	0	0	38
9	0	0	1	0	0	0	0	0	0	0	1	0	0	38
22	1	0	0	0	0	0	0	0	0	0	0	1	0	59
21	0	1	0	0	0	0	0	0	0	0	0	1	0	59
16	0	0	1	0	0	0	0	0	0	0	0	1	0	59
10	1	0	0	0	0	0	0	0	0	0	0	0	1	43
18	0	1	0	0	0	0	0	0	0	0	0	0	1	43
15	0	0	1	0	0	0	0	0	0	0	0	0	1	43

The values in Table 5 are defined as follows:

Y = the criterion score;

X_1 = 1 if the score corresponds to Treatment 1, 0 otherwise;

X_2 = 1 if the score corresponds to Treatment 2, 0 otherwise;

X_3 = 1 if the score corresponds to Treatment 3, 0 otherwise;

X_4 = 1 if the score is obtained from Subject 1, 0 otherwise;

X_5 = 1 if the score is obtained from Subject 2, 0 otherwise;

X_6 = 1 if the score is obtained from Subject 3, 0 otherwise;

X_7 = 1 if the score is obtained from Subject 4, 0 otherwise;

X_8 = 1 if the score is obtained from Subject 5, 0 otherwise;

X_9 = 1 if the score is obtained from Subject 6, 0 otherwise;

X_{10} = 1 if the score if obtained from Subject 7, 0 otherwise;

X_{11} = 1 if the score is obtained from Subject 8, 0 otherwise;

X_{12} = 1 if the score is obtained from Subject 9, 0 otherwise;

X_{13} = 1 if the score is obtained from Subject 10, 0 otherwise; and

X_{14} = the sum of the criterion scores for each subject separately.

A full model for this data could be given as:

$$Y = b_0 + b_1X_1 + b_2X_2 + b_4X_4 + b_5X_5 + b_6X_6 + b_7X_7 + b_8X_8 + b_9X_9 + b_{10}X_{10} + b_{11}X_{11} + b_{12}X_{12} + e_{10}; \qquad [19]$$

an alternative model would be:

$$Y = b_0 + b_1X_1 + b_3X_3 + b_4X_4 + b_5X_5 + b_6X_6 + b_7X_7 + b_8X_8 + b_9X_9 + b_{10}X_{10} + b_{11}X_{11} + b_{12}X_{12} + e_{10}. \qquad [20]$$

See Table 6 for a printout using equation 19.

From Table 6, it can be seen that $t_1 = 1.10362$ and $t_2 = 4.59846$; that t values are respectively the tests regarding comparing $\overline{Y}_1$ to $\overline{Y}_3$ and $\overline{Y}_2$ to $\overline{Y}_3$, taking into account that the subjects serve as their own controls. A similar printout could be generated using a model corresponding to equation 20. Values from this printout show $t_1 = -3.49484$, $t_3 = -4.59847$; these t values correspond to comparing $\overline{Y}_1$ to $\overline{Y}_2$ and $\overline{Y}_3$ to $\overline{Y}_2$. Also, the corresponding means are $\overline{Y}_1 = 13.20$, $\overline{Y}_2 = 17.00$ and $\overline{Y}_3 = 12.00$. These computed t values should be compared to an appropriate multiple comparison table for significance.

Table 6

Output of Full Model for Treatments X Subjects Design

Variable No.	Mean	Standard Deviation	Correlation X vs Y	Regression Coefficient	Std. Error Of Reg. Coef.	Computed T Value	Beta
1	0.33333	0.47946	-0.12145	1.19998	1.08732	1.10362	0.11210
2	0.33333	0.47946	0.41105	4.99997	1.08732	4.59846	0.46710
4	0.10000	0.30513	0.39195	5.66663	1.98515	2.85451	0.33690
5	0.10000	0.30513	0.28185	4.00001	1.98515	2.01496	0.23781
6	0.10000	0.30513	-0.07046	-1.33331	1.98515	-0.67164	-0.07927
7	0.10000	0.30153	-0.51085	-7.99992	1.98515	-4.12987	-0.47562
8	0.10000	0.30153	-0.15854	-2.66665	1.98515	-1.34329	-0.15854
9	0.10000	0.30153	-0.29066	-4.66664	1.98515	-2.35077	-0.27745
10	0.10000	0.30153	0.06166	0.66668	1.98515	0.33583	0.03964
11	0.10000	0.30153	-0.09248	-1.66665	1.98515	-0.83956	-0.09909
12	0.10000	0.30153	0.36993	5.33332	1.98515	2.68661	0.31708
Dependent Y	14.06667	5.13226					

INTERCEPT	12.26667	
MULTIPLE CORRELATION	0.92774	
STD. ERROR OF ESTIMATE	2.43131	
MULTIPLE CORRELATION SQUARED		0.86070
ONE MINUS MULTIPLE CORRELATION SQD.		0.13930

Analysis of Variance for the Regression

Source of Variation	Degrees Of Freedom	Sum of Squares	Mean Squares	F Value
Attributable to Regression	11	657.46021	59.76910	10.11102
Deviation from Regression	18	106.40308	5.91128	
Total	29	763.86328		

Using the Shortcut Method

The solution just given in the last section presumed that each subject (except one) is separately coded using a binary coding scheme. Clearly, if the number of subjects is at all large, the coding procedure described in Williams (1977b) and using:

$$Y = b_0 + b_1X_1 + b_2X_2 + b_{14} + e_{10} \qquad [21]$$

might be preferrable. However, one difficulty with using this shortcut procedure is that the standard error of the regression coefficients for X_1 and X_2 are too small due to the degrees of freedom, as generated by the computer program, not being accurate for deviation from regression. These t values could be adjusted by multiplying by an appropriate constant. The appropriate constant is: $c = \sqrt{\frac{MS_{w_{21}}}{MS_{w_{19}}}}$ where $MS_{w_{21}}$ is the mean square within (or deviation from regression) for equation 21 and $MS_{w_{19}}$ is the mean square within for equation 19. The $MS_{w_{21}}$ is 4.09225 and $MS_{w_{19}}$ is 5.91125. Thus, c = .83203. The values generated by equation 21 for t_1 and t_2 (comparing $\overline{Y}_1$ to $\overline{Y}_3$ and $\overline{Y}_2$ to $\overline{Y}_3$) are $t_1 = 1.32641$ and $t_2 = 5.52678$. Multiplying t_1 and t_2 by c yields corrected $t_1 = 1.10361$ and corrected $t_2 = 4.59845$, within rounding error of the values found earlier. Of course, $MS_{w_{19}}$ would not be available were the researcher using the shortcut method. However, $MS_{w_{19}} = \frac{SS_w}{N-S-g+1}$ where N is the total number of scores, S is the number of subjects and g is the number of groups. The denominator can also be found as (S-1)(g-1).

Repeated Measures Designs

Multiple comparisons also can be relatively routinized for large data sets involving repeated measures. Williams and Williams (1984) showed research application of a hypotheses testing process for k groups ɜasured at three times for large N. More recently, they showed in press) the same solutions to the problem done earlier in Williams 1980); a 3 x 4 repeated measure design with five entries per cell was ıade to show a problem that was not solvable in a regression format; ·ortunately (or unfortunately) a solution was found, so the chapter vas entitled, "Problems less amenable to a regression solution." In ɑpplying this solution to the larger data set, two progressively easier solutions were found; the preferred solution (i.e., easiest to accomplish) is embarrassingly close to a simple Bottenberg and Ward/Ward and Jennings (1973) solution.

Perhaps the point of all of this is to give some comfort to those who have struggled within the use of regression as a technique to address research questions, particularly as they look over their shoulders and think they may never master the process. Insofar as I might be seen as one who has mastered this process, let me point out, I'm still learning!

References

Bottenberg, R. A., & Ward, J. H. (1963). Applied multiple linear regression. Lackland Air Force Base, Texas: Personnel Research Laboratory PRL-TDR-63-6.

Cohen, J. (1968). Multiple regression as a general data-analytic system. Psychological Bulletin, 70, 526-543.

Dunn, O. J. (1961). Multiple comparisons among means. Journal of the American Statistical Association, 56, 52-64.

Dunnett, C. W. (1955). A multiple comparison procedure for comparing several treatments with a control. Journal of the American Statistical Association, 50, 1086-1121.

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

Scheffe, H. (1959). The analysis of variance. New York: Wiley.

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

Tukey, J. W. (1953). The problem of multiple comparisons. Dittoed, Princeton University.

Ward, J. W., & Jennings, E. E. (1973). Introduction to linear models. Englewood Cliffs, New Jersey: Prentice-Hall.

Williams, J. D. (1970). A regression approach to experimental design. Journal of Experimental Education, 39(1), 83-90.

Williams, J. D. (1971). A multiple regression approach to multiple comparisons for comparing several treatments with a control. Journal of Experimental Education, 39(3), 93-96.

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

Williams, J. D. (1974a) A simplified regression formulation of Tukey's test. Journal of Experimental Education, 42(4), 80-82.

Williams, J. D. (1974b) Regression analysis in educational research. New York: MSS Information Corporation.

Williams, J. D. (1976). Multiple comparisons by multiple linear regression. Multiple Linear Regression Viewpoints Monograph Series-2, 7(1).

Williams, J. D. (1977a). Full rank and non-full rank models with contrast and binary coding systems for two-way disproportionate cell frequency analyses. Multiple Linear Regression Viewpoints, 8(1), 1-18.

ms, J. D. (1977b). A note on coding the subjects effects in treat-
ents x subjects designs. Multiple Linear Regression Viewpoints, 8(1),
2-35.

ms, J. D. (1979). Contrasts with unequal N by multiple linear
egression. Multiple Linear Regression Viewpoints, 9(3), 1-7.

ams, J. D. (1980). Multiple comparisons in higher dimensional designs.
Multiple Linear Regression Viewpoints Monograph Series #5, 10(3).

ams, J. D., & Williams, J. A. (1984). Testing hypotheses in a repeated
measures design on employees attitudes with large samples. Paper
presented at the Annual Meeting of the American Educational Research
Association, New Orleans, April.

iams, J. D., & Williams, J. A. (in press). Testing hypotheses in a
repeated measures design-An example. Multiple Linear Regression Viewpoints.