

2X2 Analysis of Variance and Multiple Regression: Coding Does Make a Difference
R. C. Gardner
Department of Psychology

It is generally well-known that analysis of variance can be performed using multiple regression (cf., Cohen & Cohen, 1983). Often this approach is used when at least one of the factors is continuous, but as discussed by Gardner (2007, 2008) one can perform even these analyses using SPSS GLM. The purpose of this report is to comment on the use of multiple regression to perform factorial design analyses of variance when factors are categorical, and to highlight some of the issues involved when considering the two approaches equivalent. For the sake of simplicity, this presentation will provide numerical examples for the special case of a 2x2 completely randomized factorial design with equal sample sizes in the cells. The observations made, however, generalize to cases where the sample sizes are not equal, where there are more than two levels of any given factor(s), and/or where there are more than two factors (Gardner, 2006a; 2006b).

Following is the SPSS data file that will be used to illustrate the points made. The file contains the coding necessary to perform GLM UNIVARIATE as well as the effect coding and the dummy coding needed for the multiple regression runs.

	a	b	x	ea	eb	eaeb	da	db	dad
1	1.00	1.00	11.00	1.00	1.00	1.00	1.00	1.00	1.00
2	1.00	1.00	14.00	1.00	1.00	1.00	1.00	1.00	1.00
3	1.00	1.00	15.00	1.00	1.00	1.00	1.00	1.00	1.00
4	1.00	1.00	13.00	1.00	1.00	1.00	1.00	1.00	1.00
5	1.00	1.00	17.00	1.00	1.00	1.00	1.00	1.00	1.00
6	1.00	2.00	16.00	1.00	-1.00	-1.00	1.00	.00	.00
7	1.00	2.00	12.00	1.00	-1.00	-1.00	1.00	1.00	.00
8	1.00	2.00	17.00	1.00	-1.00	-1.00	1.00	1.00	.00
9	1.00	2.00	18.00	1.00	-1.00	-1.00	1.00	1.00	.00
10	1.00	2.00	18.00	1.00	-1.00	-1.00	1.00	1.00	.00
11	2.00	1.00	17.00	-1.00	1.00	-1.00	.00	1.00	1.00
12	2.00	1.00	16.00	-1.00	1.00	-1.00	.00	1.00	1.00
13	2.00	1.00	15.00	-1.00	1.00	-1.00	.00	1.00	1.00
14	2.00	1.00	15.00	-1.00	1.00	-1.00	.00	-1.00	1.00
15	2.00	1.00	17.00	-1.00	1.00	-1.00	.00	1.00	1.00
16	2.00	2.00	13.00	-1.00	-1.00	1.00	.00	.00	.00
17	2.00	2.00	10.00	-1.00	-1.00	1.00	.00	.00	.00
18	2.00	2.00	16.00	-1.00	-1.00	1.00	.00	.00	.00
19	2.00	2.00	14.00	-1.00	-1.00	1.00	.00	.00	.00
20	2.00	2.00	12.00	-1.00	-1.00	1.00	.00	-1.00	.00

The analysis of variance produces the following summary table. Note, that in this analysis, the main effects of A and B are not significant but the interaction is ($p < .013$).

Tests of Between-Subjects Effects

Dependent Variable: x

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	36.400 ^a	3	12.133	2.822	.072
Intercept	4380.800	1	4380.800	1018.791	.000
a	1.800	1	1.800	.419	.527
b	.800	1	.800	.186	.672
a * b	33.800	1	33.800	7.860	.013
Error	68.800	16	4.300		
Total	4486.000	20			
Corrected Total	105.200	19			

a. R Squared = .346 (Adjusted R Squared = .223)

If requested, GLM UNIVARIATE will output Parameter estimates for this analysis. As we shall see, these are the regression coefficients for Dummy coding. We shall discuss this further in a latter section.

Parameter Estimates

Dependent Variable: x

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	13.00000	.92736	14.01826	.00000	11.03408	14.96592
[a=1.00]	3.20000	1.31149	2.43998	.02670	.41977	5.98023
[a=2.00]	0 ^a
[b=1.00]	3.00000	1.31149	2.28748	.03612	.21977	5.78023
[b=2.00]	0 ^a
[a=1.00] * [b=1.00]	-5.20000	1.85472	-2.80365	.01274	-9.13184	-1.26816
[a=1.00] * [b=2.00]	0 ^a
[a=2.00] * [b=1.00]	0 ^a
[a=2.00] * [b=2.00]	0 ^a

a. This parameter is set to zero because it is redundant.

The means for this analysis are shown in the next table, and as can be seen there are minimal differences in the main effects of both A and B, but a clear interaction between the two.

	B1	B2	A Means
A1	14.0	16.2	15.1
A2	16.0	13.0	14.5
B Means	15.0	14.6	14.8

We could perform the same analysis using multiple regression. With an AxB design, this would require (a-1) vectors for the A factor, (b-1) vectors for the B factor and (a-1)(b-1) vectors for the interaction. Most presentations deal with Effect coding or Dummy coding, but in fact any type of coding, even nonsense coding, will yield the same multiple correlations and the same estimates of the cell values 14, 16.2, 16, and 13, as long as the codes are unique for the four groupings. Thus, one might say that the type of coding doesn't matter, however, as shown below the type of coding does determine the values of the regression coefficients and thus the interpretation of the associated tests of significance.

The effect coding is shown in the data editor as ea, eb, and eab respectively. If effect coding is used to perform the multiple regression, the following four multiple correlations can be computed. They are:

$$R^2_{A,B, AB} = .34601 \quad R^2_{B,AB} = .32890 \quad R^2_{A,AB} = .33840 \quad R^2_{A,B} = .02471$$

and from these, we can compute the squared multiple semipartial correlations as follows:

$$R^2_A = .34601 - .32890 = .01711$$

$$R^2_B = .34601 - .33840 = .00761$$

$$R^2_{AB} = .34601 - .02471 = .32130$$

The corresponding F-ratios are:

$$F_A = R^2_A / ((1 - R^2_{A,B,AB}) / (N - 4)) = .01711 / (.65399 / 16) = .01711 / .04087 = .41864$$

$$F_B = R^2_B / ((1 - R^2_{A,B,AB}) / (N - 4)) = .00761 / .04087 = .18620$$

$$F_{AB} = R^2_{AB} / ((1 - R^2_{A,B,AB}) / (N - 4)) = .32130 / .04087 = 7.86151$$

It will be noted that these F-ratios agree with those obtained in the analysis of variance.

The regression coefficients for the full model are:

Coefficients^a

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1.000 (Constant)	14.80000	.46368		31.91850	.00000
ea	.30000	.46368	.13081	.64700	.52681
eb	.20000	.46368	.08720	.43133	.67198
eaeb	-1.30000	.46368	-.56683	-2.80365	.01274

a. Dependent Variable: x

Note that these regression coefficients are different from the Parameter Estimates obtained in the analysis of variance. This is because these are regression coefficients for Effect coding while the parameter Estimates are regression coefficients for Dummy coding. Note further that these coefficients are directly related to the means in the analysis of variance. Thus:

$$\text{Constant} = 14.80000 = \text{Grand Mean} = 14.8.$$

$$b_{ea} = .30000 = \text{Mean}_{1.} - \text{Grand Mean} = 15.1 - 14.8 = .3$$

$$b_{eb} = .20000 = \text{Mean}_{1.} - \text{Grand Mean} = 15.0 - 14.8 = .2$$

$$b_{eaeb} = -1.30000 = \text{Mean}_{11} - \text{Mean}_{1.} - \text{Mean}_{.1} + \text{Grand Mean} = 14.0 - 15.1 - 15 + 14.8 = -1.3.$$

Note too that the definition of the three b -coefficients are effects, hence the name for the coding, and that the t -tests of significance of all the coefficients are the square roots of the F -ratios in the analysis of variance summary table.

Following are the results using Dummy coding. The vectors in the data file are identified as da, db, and dadb respectively. The four multiple correlations are:

$$R^2_{A,B, AB} = .34601 \quad R^2_{B,AB} = .10266 \quad R^2_{A,AB} = .13213 \quad R^2_{A,B} = .02471$$

and from these, we can compute the squared multiple semipartial correlations as follows:

$$R^2_A = .34601 - .10266 = .24335$$

$$R^2_B = .34601 - .13213 = .21388$$

$$R^2_{AB} = .34601 - .02471 = .32130$$

The corresponding F-ratios are:

$$F_A = R^2_A / ((1 - R^2_{A,B,AB}) / (N - 4)) = .24335 / (.65399 / 16) = .24335 / .04087 = 5.95425$$

$$F_A = R^2_B / ((1 - R^2_{A,B,AB}) / (N - 4)) = .21388 / .04087 = 5.23318$$

$$F_A = R^2_{AB} / ((1 - R^2_{A,B,AB}) / (N - 4)) = .3213 / .04087 = 7.86151$$

It will be noted that the F -ratios for A and B are not the same as those in the analysis of variance table, or those obtained with Effect coding, while that for the highest effect, the AxB interaction, is the same as those obtained previously.¹

The regression coefficients for the full model using Dummy coding are:

Coefficients^a

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1.000 (Constant)	13.00000	.92736		14.01826	.00000
da	3.20000	1.31149	.69763	2.43998	.02670
db	3.00000	1.31149	.65403	2.28748	.03612
dadb	-5.20000	1.85472	-.98177	-2.80365	.01274

a. Dependent Variable: x

Note that these regression coefficients are the same as those in the Parameter Estimates obtained in the analysis of variance. This is because these are both coefficients associated with Dummy coding. Note further that these coefficients are directly related to the means in the analysis of variance, though in a manner that is different from that for Effect coding. Thus:

$$\text{Constant} = 13.0000 = \text{Mean}_{22} = 13.0$$

$$b_{da} = 3.20000 = \text{Mean}_{12} - \text{Mean}_{22} = 16.2 - 13.0 = 3.2$$

$$b_{db} = 3.00000 = \text{Mean}_{21} - \text{Mean}_{22} = 16.0 - 13.0 = 3.0$$

$$b_{dadb} = -5.20000 = \text{Mean}_{11} - \text{Mean}_{12} - \text{Mean}_{21} + \text{Mean}_{22} = 14.0 - 16.2 - 16.0 + 13.0 = -5.2$$

Note that the definition of the three b -coefficients are contrasts involving the cell means, and are not effects. The values of 3.2 and 3.0 are differences between cell means and the cell mean for the group that is coded as 0 on both the A and the B vector, and the value of -5.2 is a contrast/contrast interaction. Furthermore, the t -tests of significance for each of the b -

¹It may seem curious that the GLM Univariate program uses Dummy coding and obtains the correct F -ratios, while the F -ratios obtained from multiple regression using Dummy coding are not the same. This is because although both procedures are instances of the general linear model, SPSS GLM Univariate solves a series of matrix equations and does not calculate a set of multiple correlations as done here.

coefficients are the square roots of the F -ratios calculated above using Dummy coding.

In summary, although any type of coding of categorical variables will generate the same value of the multiple correlation for the full model, each one has different definitions of the associated regression coefficients. This report has focused on the distinction between Effect coding and Dummy coding and the F -ratios obtained using Model I (the unique sums of squares approach, or SPSS Type 3). Different results for the main effects would be obtained if Model II or Model III were used (and Effect coding and Dummy coding would yield identical F -ratios), but the final set of regression coefficients, and thus their interpretation, would be comparable in definition to those presented above. As I often say, attempting to interpret regression coefficients, particularly when categorical variables and interactions are involved, is a mug's game and the individual who is not aware of the precise definitions of the regression coefficients should beware!

References

- Cohen, J. & Cohen, P. (1975). *Applied multiple regression/correlation analysis for the behavioral sciences*. New York: Erlbaum.
- Gardner, R. C. (2006a). Analysis of variance with categorical and continuous factors: Beware the landmines. Unpublished manuscript, Department of Psychology, University of Western Ontario.
- Gardner, R. C. (2006b). On the meaning of regression coefficients for categorical and continuous variables: Model I and Model II: Effect coding and Dummy coding. Unpublished manuscript, Department of Psychology, University of Western Ontario.
- Gardner, R. C. (2007). Performing analyses of variance with continuous and categorical factors: The easy way. Unpublished manuscript, Department of Psychology, University of Western Ontario.
- Gardner, R. C. (2008). Three factor completely randomized design with one continuous factor: Using SPSS GLM Univariate. Unpublished manuscript, Department of Psychology, University of Western Ontario.