Data_analysis.calm

# On the Meaning of Regression Coefficients for Categorical and Continuous Variables: Model I and Model II; Effect Coding and Dummy Coding

R. C. Gardner
Department of Psychology

This describes the simple case where there is one categorical variable (with 3 levels) and one continuous variable, though the observations can be generalized to any number of variables and levels. In this situation, we could consider the 3 groups and the results of bivariate regressions computed on each one. This would yield the following:

| Gp1 | Gp2 | Gp3 |
|---|---|---|
| $Y = b_{01} + b_{11}C$ | $Y = b_{02} + b_{12}C$ | $Y = b_{03} + b_{13}C$ |

Using multiple regression with Effect or Dummy Coding, we could generate regression equations for Model I or Model II. Model I is the unique sum of squares approach where all variables are entered into the regression equation simultaneously. Model II is a hierarchical approach where the "main" effects are entered first and the two way interaction is entered on the next step. This could be extended for more variables with no loss in generality.

Examples of the coding would be as follows. In this example, two vectors (A1, and A2) are needed to define the three groups, C is the continuous variable (centered or not), and A1C and A2C would represent the products of the A and C vectors. Only 1 observation is shown in each group (and no value of C is shown because this could change for each observation in each group), but obviously there would be more. The generalizations that follow apply to both equal and unequal sample sizes.

| Group | Effect Coding | | | | | Dummy coding | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | A1 | A2 | C | A1C | A2C | A1 | A2 | C | A1C | A2C |
| 1 | 1 | 0 | . | C | 0 | 1 | 0 | . | C | 0 |
| 2 | 0 | 1 | . | 0 | C | 0 | 1 | . | 0 | C |
| 3 | -1 | -1 | . | -C | -C | 0 | 0 | . | 0 | 0 |

**Model I**

For Model I, all terms are entered at once, and the regression equation is:

$$Y = b_0 + b_1 A1 + b_2 A2 + b_3 C + b_4 A1C + b_5 A2C$$

The following table illustrates the precise meaning of the regression coefficients (in terms of the univariate regression coefficients shown for each group in the first table presented) when both Effect coding and Dummy coding is used. This is done for explanatory purposes. In either

case, the multiple correlation will be identical but the regression coefficients will take on different values, and the tests of significance of the squared multiple semi-partial correlations for the A vectors and the continuous variable will also differ. In fact, Dummy coding should not be used for Model I because it produces incorrect tests of significance for the main effects of the categorical factor (i.e., the F-ratio for the squared multiple semi-partial correlation).

| Coefficient | Effect coding | Dummy Coding |
|---|---|---|
| $b_0$ | $\frac{b01+b02+b03}{3} = \overline{b0}$ | $b_{03}$ |
| $b_1$ | $b_{01} - b_0$ | $b_{01} - b_{03}$ |
| $b_2$ | $b_{02} - b_0$ | $b_{02} - b_{03}$ |
| $b_3$ | $\frac{b11+b12+b13}{3} = \overline{b1}$ | $b_{13}$ |
| $b_4$ | $b_{11} - b_3$ | $b_{11} - b_{13}$ |
| $b_5$ | $b_{12} - b_3$ | $b_{12} - b_{13}$ |

It will be noted that the regression coefficients for Effect coding describe effects, while those for Dummy coding describe contrasts with the group coded with all 0's (group 3 in this case).

**Model II**

For Model II, the effects are entered hierarchically, beginning with the main effects only. Thus, the first step in Model II, would produce the following regression equation:

$$Y = b_0 + b_1A1 + b_2A2 + b_3C$$

Note in this case, there is no product term involving A and C, hence only the meanings for the regression coefficients for $b_0$ $b_1$ and $b_2$ given in the above table apply. That is, the same definitions are applicable but the meaning of the terms $b_{01}$, $b_{02}$, and $b_{03}$ change, and the value of $b_3$ is the within cells regression coefficient ($b_w$) for the 3 groups. This value is:

$$bw = \frac{\sum_a \sum_n (C_{ia} - \overline{C}_a)(Y_{ia} - \overline{Y}_a)}{\sum_a \sum_n (C_{ia} - \overline{C}_a)^2} = \frac{\sum_a b_a S^2_a (n-1)}{\sum_a S^2_a (n-1)}$$

This stage of the analysis is simply an analysis of covariance where group is the treatment variable and C is the covariate. Thus:

$b_{01}$ = the adjusted mean of Y for group 1
$b_{02}$ = the adjusted mean of Y for group 2
$b_{03}$ = the adjusted mean of Y for group 3

As a consequence, the values of the regression coefficients will change between Effect coding and Dummy coding, but the tests of significance will be the same.

The second step in Model II adds the product terms, so that the definition of the regression coefficients at this stage are identical to those given earlier for Model I. For obvious reasons, the values of the regression coefficients for $b_0$, $b_1$, $b_2$, and $b_3$ are different from what they were in the equation generated by step 1.

**Numerical Illustration - Model I**

The following example is based on data from 3 groups with a dependent variable, Y, and a centered continuous variable, A. Bivariate regression analysis for each of the groups yields the following statistics.

| Group | Y mean | A mean | Intercept | Slope | n | Standard Deviation |
|---|---|---|---|---|---|---|
| 1 | 6.333 | - .1667 | 6.380 | .282 | 6 | 6.43169 |
| 2 | 8.625 | .8750 | 8.565 | .069 | 8 | 6.91659 |
| 3 | 8.300 | -1.2000 | 9.012 | .593 | 5 | 4.60435 |

The following tables present the regression coefficients for Model I for both Effect coding and Dummy coding. These results are used to illustrate the formulae given previously.

Model I Effect Coding

**Coefficients[a]**

| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. |
|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | | |
| 1 | (Constant) | 7.986 | .377 | | 21.154 | .000 |
| | a | .315 | .074 | .796 | 4.256 | .001 |
| | b1 | -1.605 | .531 | -.531 | -3.023 | .010 |
| | b2 | .579 | .499 | .205 | 1.160 | .267 |
| | ab1 | -.033 | .098 | -.056 | -.335 | .743 |
| | ab2 | -.246 | .089 | -.506 | -2.753 | .016 |

a. Dependent Variable: x

Model I Dummy Coding

**Coefficients[a]**

| Model | | Unstandardized Coefficients | | Standardized Coefficients | | |
|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | t | Sig. |
| 1 | (Constant) | 9.012 | .738 | | 12.213 | .000 |
| | a | .593 | .172 | 1.500 | 3.448 | .004 |
| | d1 | -2.631 | .981 | -.533 | -2.681 | .019 |
| | d2 | -.447 | .929 | -.096 | -.481 | .638 |
| | ad1 | -.311 | .204 | -.447 | -1.523 | .152 |
| | ad2 | -.524 | .193 | -.964 | -2.721 | .017 |

a. Dependent Variable: x

Close inspection of these tables will demonstrate that although both analyses yield the same squared multiple correlation (.674), the regression statistics generated by Effect coding are different from those for Dummy coding. Following illustrates the precise meaning of these statistics in terms of the regression coefficients obtained in the bivariate solutions.

| Coefficient | Effect coding | Dummy Coding |
|---|---|---|
| $b_0$ | $\frac{6.380+8.565+9.012}{3} = 7.986$ | 9.012 |
| $b_1$ | 6.380 - 7.986 = -1.606 | 6.380 - 9.012 = -2.632 |
| $b_2$ | 8.565 -7.986 = .579 | 8.565 - 9.012 = -.447 |
| $b_3$ | $\frac{.282+.069+.593}{3} = .315$ | .593 |
| $b_4$ | .282 - .315 = -.033 | .282 - .593 = -.311 |
| $b_5$ | .069 - .315 = -.246 | .069 - .593 = -.524 |

In the regression table for effect coding presented earlier, significant values were obtained for the constant (i.e., $b_0$), the continuous variable, a, (i.e., b3), the first categorical variable, b1 (i.e., $b_1$), and the product, ab2 (i.e., $b_5$). Thus, these values suggest that:

1. the intercept, 7.986 (the value predicted when a = 0) is significantly different from 0,
2. the difference between the intercept for group 1 and the mean of the intercepts (i.e., $b_1$ = 6.380 - 7.986 = -1.606) is significantly different from 0,
3. the mean slope (i.e., $b_3$ = .315) is significantly different from 0, and
4. the slope for group 2 differs significantly from the mean of the slopes (i.e., .069 - .315 = -.246.

The results for Dummy coding also indicate significant effects for the constant, a, d1, and ad2, but the values are different and their meanings are very different. These results suggest that:

1. the intercept for group 3, 9.012 is significantly different from 0,
2. the difference between the intercept for group 1 and the intercept for group 3 (i.e., $b_1$ = 6.380 - 9.012 = -2.632) is significantly different from 0,
3. the slope for group 3 (i.e., $b_3$ = .593) is significantly greater than 0, and
4. the slope for group 2 is significantly less than the slope for group 3 (i.e., .069 - .593 = -.524.

**Model II**

The following tables present the regression coefficients for Model II for both Effect coding and Dummy coding.

Model II Effect Coding

**Coefficients[a]**

| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. |
|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | | |
| 1 | (Constant) | 7.787 | .439 | | 17.751 | .000 |
| | a | .210 | .075 | .532 | 2.806 | .013 |
| | b1 | -1.419 | .623 | -.469 | -2.278 | .038 |
| | b2 | .654 | .587 | .231 | 1.113 | .283 |
| 2 | (Constant) | 7.986 | .377 | | 21.154 | .000 |
| | a | .315 | .074 | .796 | 4.256 | .001 |
| | b1 | -1.605 | .531 | -.531 | -3.023 | .010 |
| | b2 | .579 | .499 | .205 | 1.160 | .267 |
| | ab1 | -.033 | .098 | -.056 | -.335 | .743 |
| | ab2 | -.246 | .089 | -.506 | -2.753 | .016 |

a. Dependent Variable: x

Model II Dummy Coding

**Coefficients[a]**

| Model | | Unstandardized Coefficients | | Standardized Coefficients | | |
|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | t | Sig. |
| 1 | (Constant) | 8.552 | .844 | | 10.135 | .000 |
| | a | .210 | .075 | .532 | 2.806 | .013 |
| | d1 | -2.184 | 1.139 | -.442 | -1.918 | .074 |
| | d2 | -.111 | 1.081 | -.024 | -.103 | .919 |
| 2 | (Constant) | 9.012 | .738 | | 12.213 | .000 |
| | a | .593 | .172 | 1.500 | 3.448 | .004 |
| | d1 | -2.631 | .981 | -.533 | -2.681 | .019 |
| | d2 | -.447 | .929 | -.096 | -.481 | .638 |
| | ad1 | -.311 | .204 | -.447 | -1.523 | .152 |
| | ad2 | -.524 | .193 | -.964 | -2.721 | .017 |

a. Dependent Variable: x

For Model II, the regression coefficients for the first step are different than those in Model I, even though the meanings are the same as demonstrated above. The difference, of course, is because the initial estimates are different. That is, the regression coefficient for the continuous independent variable (a) is the within cells regression coefficient. The value is:

$$\frac{(.282)(5)(6.43169^2)+(.069)(7)(6.91659^2)+(.593)(4)(4.60435^2)}{(5)(6.43169^2)+(7)(6.91659^2)+(4)(4.60435^2)} = .210$$

and the intercepts for each group are the adjusted means for that group. That is:

$b_{01}$ = the adjusted mean of Y for group 1 = 6.333 - .210 (-.1667-0) = 6.368
$b_{02}$ = the adjusted mean of Y for group 2 = 8.625 - .210 (.8750 - 0) = 8.441
$b_{03}$ = the adjusted mean of Y for group 3 = 8.300 - .210 (-1.200 - 0) = 8.552

Using these values in the defining formula given earlier for $b_0$, $b_1$, $b_2$, and $b_3$ yields the following table.

| Coefficient | Effect coding | Dummy Coding |
|---|---|---|
| $b_0$ | $\frac{6.368+8.441+8.552}{3} = 7.787$ | 8.552 |
| $b_1$ | 6.368 - 7.787 = -1.419 | 6.368 - 8.552 = -2.184 |
| $b_2$ | 8.441 - 7.787 = .654 | 8.441 - 8.552 = -.111 |
| $b_3$ | .210 | .210 |

In this case, the tests of significance for Effect coding at the first step indicate that:

1. the mean intercept (i.e., the mean of the adjusted means (7.787) differs significantly from 0,
2. the intercept for group 1 differs significantly from the mean of the intercepts (i.e., b1 = 6.368 - 7.787 = -1.419),
3. the within cells regression coefficient (i.e., $b_3$ = .210) differs significantly from 0.

On the other hand, the results for Dummy coding at the first step indicate that:

1. the intercept (i.e., the adjusted mean for group 3 = 8.552) differs significantly from 0, and
2. the within cells regression coefficient (i.e., $b_3$ = .210) differs significantly from 0. It will be noted that this is the only place where the two analyses describe the same result.

Finally, it will be noted that the results at step 2 have values identical to those obtained with Model I.