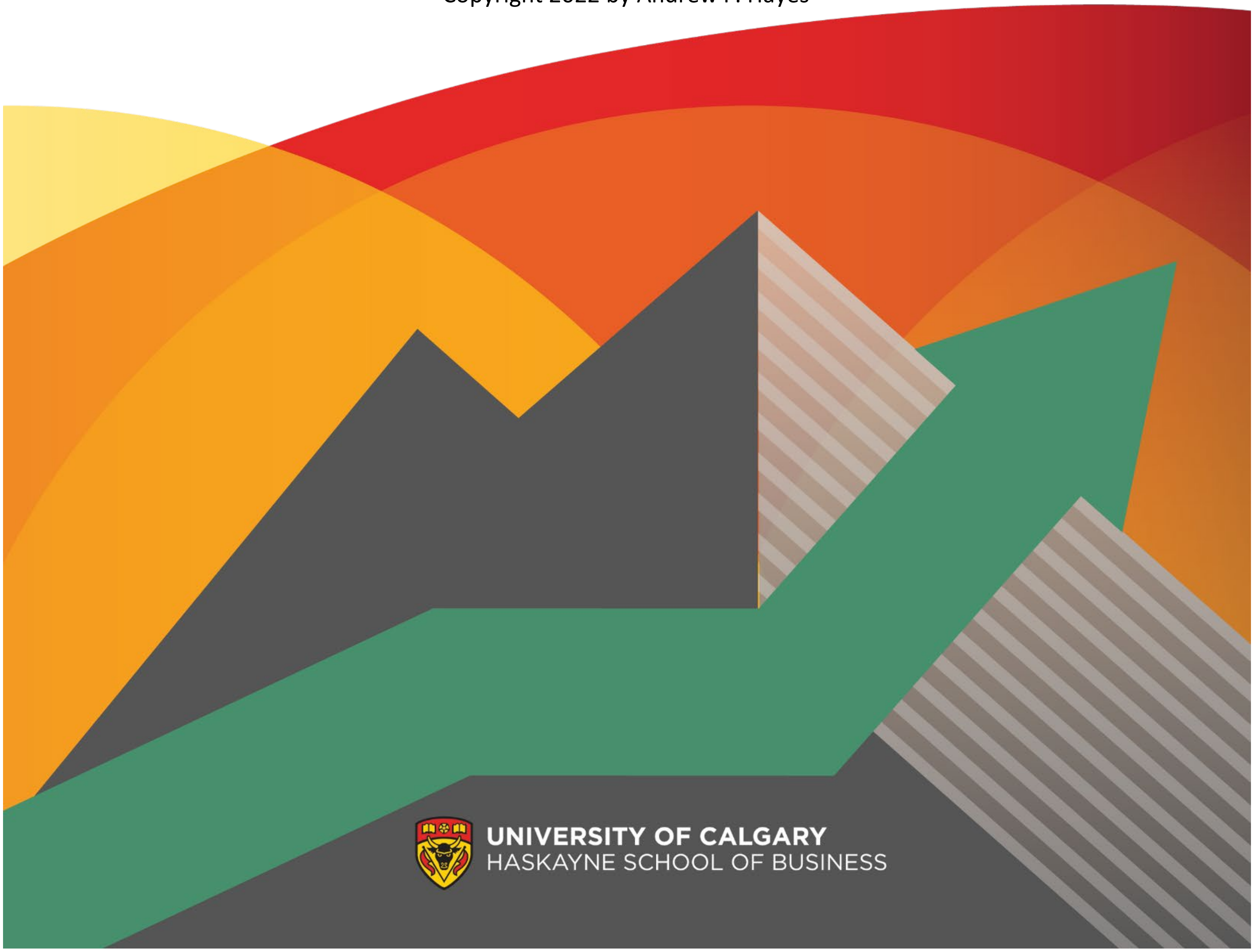
HASKAYNE
CCRAM
CANADIAN CENTRE FOR
RESEARCH ANALYSIS AND METHODS

CCRAM Technical Report 022-03

Changing the Reference Group when Using Indicator Coding for a Multicategorical Variable in PROCESS

Andrew F. Hayes
Haskayne School of Business, University of Calgary



haskayne.ucalgary.ca/CCRAM

A multicategorical variable (a variable representing membership in one of three or more groups) is typically a single variable in a data set containing a set of numerical codes representing which group a case in the data belongs in. Such a variable cannot be used as a predictor in a regression analysis as is. Doing so will usually produce nonsense. The proper approach to including a multicategorical variable representing *k* groups on the right-hand side of a regression equation is to use a set of *k* – 1 variables that code group membership. There are many coding systems that can be used. One of the most popular is *indicator* coding, also called *dummy* coding. With indicator coding, one group is chosen as the *reference group* and cases in that group receives a value of 0 on the *k* – 1 indicator or dummy variables representing the *k* groups. The cases in the remaining *k* – 1 groups are each given a value of 1 on the indicator variable for that group and all other indicator variables for cases in that group are set to 0.

For example, suppose in your data file you have a variable named `group`, with the *k* = 4 groups coded with values 1, 2, 3, and 4 in the data. The table below represents a possible indicator coding system:

Table 1.

group	X1	X2	X3
1	1	0	0
2	0	1	0
3	0	0	1
4	0	0	0

In this system, X1, X2, and X3 are the indicator variables, and group 4 is the reference group because it is represented with zeros on all 3 of the indicator variables. Notice that although there are only three indicator variables, there are four patterns of zeros and ones, one pattern for each group. It is the pattern of values on X1, X2, and X3 that represents group membership.

Using the **`mcx`**, **`mcw`**, or **`mcz`** options in PROCESS, you can tell PROCESS to represent a multicategorical variable specified as X, W, or Z in the model with one of several preprogrammed coding systems. Indicator coding is available by specifying option 1, as discussed in *Introduction to Mediation, Moderation, and Conditional Process Analysis*. When using option 1, PROCESS will always use the group with the numerically *smallest* code as the reference group. For example, in the four-group example above, PROCESS would choose the group coded 1 as the reference group because “1” is the smallest value on the group variable.

To illustrate, we use the data from Chapter 6 of *Introduction to Mediation, Moderation, and Conditional Process Analysis*. In the data file, the three experimental conditions are coded 0, 1, and 2 in the variable named `protest` for the “no protest,” “individual protest,” and “collective protest” conditions, respectively. As the “no protest” group has the numerically smallest code (0) in the `protest` variable, PROCESS will treat it as the reference group when creating the indicator codes. The PROCESS command below (SPSS first, SAS second, R third) estimates evaluation of the attorney from protest condition, specifying protest as a multicategorical variable and using indicator coding to represent the three groups. (Note that this code will only work with PROCESS version 4.1 or later, as model 0 was not available as a model number until version 4.1 released in April 2022).

```
process y=liking/x=protest/model=0/mcx=1.
```

```
%process(data=protest,y=liking,x=protest,model=0,mcx=1)
```

```
process(data=protest,y="liking",x="protest",model=0,mcx=1)
```

This produces as output

```
Model  : 0
    Y  : liking
    X  : protest

Sample
Size:  129

Coding of categorical X variable for analysis:
 protest      X1      X2
    .000    .000    .000
   1.000   1.000    .000
   2.000    .000   1.000

**************************************************************************
OUTCOME VARIABLE:
 liking

Model Summary
          R       R-sq        MSE          F        df1        df2          p
      .2151      .0463     1.0676     3.0552     2.0000   126.0000      .0506

Model
              coeff         se          t          p       LLCI       ULCI
constant     5.3102      .1614    32.9083      .0000     4.9909     5.6296
X1            .5158      .2255     2.2870      .0239      .0695      .9621
X2            .4431      .2231     1.9863      .0492      .0016      .8845
```

Notice at the beginning of the output that PROCESS has produced a table representing the indicator coding system used. The no protest group is the reference group, as X1 and X2 are both 0 for this group. The *F*-ratio in the model summary is identical to the *F*-ratio from a single factor or "one-way" analysis of variance testing the null hypothesis that the group means are the same. Here, $F(2,126) = 3.055$, $p = .051$.

In the data, the group means are 5.310 for the no protest group, 5.826 for the individual protest group, and 5.753 for the collective protest group. The model is $\hat{Y} = 5.310 + 0.516X_1 + 0.443X_2$. When the pattern of indicator codes for each group are plugged into the regression model, the model produces an estimate of *Y* for each group that corresponds to that group's mean on *Y*:

No protest group: $\hat{Y} = 5.310 + 0.516(0) + 0.443(0) = 5.310$

Individual protest group: $\hat{Y} = 5.310 + 0.516(1) + 0.443(0) = 5.826$

Collective protest group: $\hat{Y} = 5.310 + 0.516(0) + 0.443(1) = 5.753$

The regression constant is the mean for the no protest reference group. The regression weight for X1 is the difference between means of the individual protest group and the no protest reference group: 0.516 = 5.826 – 5.310. And the regression weight for X2 is the difference between the means of the collective protest group and the no protest reference group: 0.443 = 5.753 – 5.310. Both the differences are statistically significant. So each of these regression coefficients quantify one group's mean relative to the no protest group mean. That's why the no protest group is called the reference group here.

Changing the Reference Group

But what if you want to a different group as the reference? As PROCESS is programmed to always make the group with the numerically smallest group the reference, you will need to do something to change the reference group. The most obvious solution is to recode the groups so that that the group you desire as the reference has the numerically smallest code. For example, if you want the collective protest group (`protest` = 2) to be the reference, you could swap the 0 and 2 codes for the no protest and collective protest groups. So you don't lose the original coding, I recommend putting the new codes in a different variable.

Suppose the new codes were held in a variable called `protest2` with protest2 = 0, 1, and 2 for the collective protest, individual protest, and no protest groups, respectively. Then the PROCESS command above, but using `protest2` for *X*, generates

```
Model  : 0
    Y  : liking
    X  : protest2

Sample
Size:  129

Coding of categorical X variable for analysis:
 protest2       X1       X2
     .000     .000     .000
    1.000    1.000     .000
    2.000     .000    1.000

**************************************************************************
OUTCOME VARIABLE:
 liking

Model Summary
          R       R-sq        MSE          F        df1        df2          p
      .2151      .0463     1.0676     3.0552     2.0000   126.0000      .0506

Model
              coeff         se          t          p       LLCI       ULCI
constant     5.7533      .1540    37.3530      .0000     5.4485     6.0581
X1            .0727      .2203      .3300      .7419     -.3633      .5088
X2           -.4431      .2231    -1.9863      .0492     -.8845     -.0016
```

Now the regression coefficient for X1 quantifies the difference between the individual protest group mean and the collective protest group mean: 0.073 = 5.826 – 5.753, and the regression coefficient for X2 quantifies the difference between the no protest group mean and the collective protest group mean: –0.443 = 5.310 – 5.753. The regression constant is the mean for the collective protest group, which is our new reference group. The regression model still generates estimates of *Y* for each group that correspond to that group's mean *Y*:

Collective protest group: $\hat{Y} = 5.753 + 0.073(0) - 0.443(0) = 5.753$
Individual protest group: $\hat{Y} = 5.753 + 0.073(1) - 0.443(0) = 5.826$
No protest group: $\hat{Y} = 5.753 + 0.073(0) - 0.443(1) = 5.310$

The test of equality of the three groups means is not affecting by changing the reference group. It is still $F(2,126) = 3.055$, $p = .051$.

An alternative approach to changing the reference group is to program your own indicator coding system using the **xcatcode** option described in Appendix A of *Introduction to Mediation, Moderation, and Conditional Process Analysis*. The desired coding system with the collective protest group as the reference is

Table 2.

	X1	X2
No protest (protest = 0)	1	0
Individual protest (protest = 1)	0	1
Collective protest (protest = 2)	0	0

In Table 2, the groups in the rows are in the same order as the groups are represented with ascending numbers in the variable coding groups that is being used in the analysis. In this case, groups are stored in the protest variable with values 0, 1, and 2. So the first row is the group coded 0, the second row is the group coded 1, and the third row is the group coded 2. Getting the order right is important because PROCESS expects this order when it creates the variable codes in response to your use of the **xcatcode** option.

To tell PROCESS to use the coding system in Table 2, specify **mcx** option 5 in the PROCESS command and then list the numerical codes in the table as a sequence of numbers following the **xcatcode** option, reading the numbers in Table 2 from left to right, top to bottom as you enter the sequence. Thus, in SPSS, SAS, and R, respectively, the PROCESS command is

```
process y=liking/x=protest/mcx=5/model=0/xcatcode=1,0,0,1,0,0.
```

```
d
%process (data=protest,y=liking,x=protest,model=0,mcx=5,xcatcode=1 0 0 1 0 0)
```

```
process (data=protest,y="liking",x="protest",model=0,mcx=5,xcatcode=c(1,0,0,1,0,0))
```

This command generates the output below, which of course is identical to what was generated when we just recoded the groups instead of programming our own indicator codes. Notice that the table toward the top of the output shows that the group coded 2 in `protest` (the collective protest group) is now the reference group. It is important to check this table to make sure that you entered the sequence of numbers correctly when using the **`xcatcode`** option.

```
Model  : 0
    Y  : liking
    X  : protest

Sample
Size:  129

Coding of categorical X variable for analysis:
 protest      X1      X2
    .000   1.000    .000
   1.000    .000   1.000
   2.000    .000    .000

**************************************************************************
OUTCOME VARIABLE:
 liking

Model Summary
          R       R-sq        MSE          F        df1        df2          p
      .2151      .0463     1.0676     3.0552     2.0000   126.0000      .0506

Model
              coeff         se          t          p       LLCI       ULCI
constant     5.7533      .1540    37.3530      .0000     5.4485     6.0581
X1           -.4431      .2231    -1.9863      .0492     -.8845     -.0016
X2            .0727      .2203      .3300      .7419     -.3633      .5088
```

Reversing the Differencing

The regression coefficients for the indicator variables quantify a difference in estimated values of *Y* between one of the groups and the reference group. As discussed earlier, these can be interpreted as a difference between group means, where the reference group mean is subtracted from the mean of the other groups. If the reference group mean is smaller, the regression coefficient is positive. If the reference group mean is larger, the coefficient is negative.

There may be occasions where you might prefer that the subtraction be reversed, such that one of the group means is subtracted from the reference group mean. This won't change the substantive interpretation of the difference between the group means, of course, but it will flip the signs of the corresponding regression coefficient and endpoints of a confidence interval. The standard error for the regression coefficient and the *p*-value for testing the null of no difference will be the same, however.

The reversal of the subtraction and corresponding reversal of the regression coefficients can be accomplished by multiplying all the nonzero values for the indicator codes by -1; in other words, turn the 1s to -1s, as in Table 3 which has the no protest group as the reference. This can be done by programming the matrix, as in the prior example.

Table 3.

	X1	X2
No protest (`protest` = 0)	0	0
Individual protest (`protest` = 1)	-1	0
Collective protest (`protest` = 2)	0	-1

The PROCESS command that accomplishes this is (using the `protest` variable in the data file, which has the groups coded no protest = 0, individual protest = 1, collective protest = 2).

```
process y=liking/x=protest/model=0/mcx=5/xcatcode=0,0,-1,0,0,-1.
```

```
%process (data=protest,y=liking,x=protest,mcx=5,xcatcode=0 0 -1 0 0 -1)
```

```
process (data=protest,y="liking",x="protest",model=0,mcx=5,xcatcode=c(0,0,-1,0,0,-1))
```

The output below results

```
Model  : 0
    Y  : liking
    X  : protest

Sample
Size:  129

Coding of categorical X variable for analysis:
 protest      X1      X2
    .000    .000    .000
   1.000  -1.000    .000
   2.000    .000  -1.000

*************************************************************************
OUTCOME VARIABLE:
 liking

Model Summary
          R       R-sq        MSE          F        df1        df2          p
      .2151      .0463     1.0676     3.0552     2.0000   126.0000      .0506

Model
              coeff         se          t          p       LLCI       ULCI
constant     5.3102      .1614    32.9083      .0000     4.9909     5.6296
X1           -.5158      .2255    -2.2870      .0239     -.9621     -.0695
X2           -.4431      .2231    -1.9863      .0492     -.8845     -.0016
```

As you can see, this output is largely identical to the output when using “1” rather than “-1” for the indicator codes from earlier in this document. However, the signs of the regression coefficients (and the endpoints of the confidence interval) have reversed. The regression coefficient for X1 is now the no protest mean minus the individual protest mean (-0.516 = 5.310 – 5.826) and the regression coefficient for X2 is now the no protest mean minus the individual protest mean (-0.443 = 5.310 – 5.753). The test of equality of the three means is unaffected, $F(2,126) = 3.055$, $p = .051$, and the model still reproduces the group means:

No protest group: $\hat{Y} = 5.310 - 0.516(0) - 0.443(0) = 5.310$

Individual protest group: $\hat{Y} = 5.310 - 0.516(-1) - 0.443(0) = 5.826$

Collective protest group: $\hat{Y} = 5.310 - 0.516(0) - 0.443(-1) = 5.753$

For more information on the coding systems for multicategorical variables available in PROCESS and programming your own system for representing groups, see the PROCESS documentation in *Introduction to Mediation, Moderation, and Conditional Process Analysis*.