Chapter 7: Categorical & Continuous Variables

Conducting Separate Regressions as a Follow-up for a Statistically Significant Interaction: A Caveat

In this chapter I suggested that when you encounter a statistically significant interaction (cross product) between a categorical and continuous variable that you should: 1. Graph it to understand the nature of the interaction, and 2. Follow-up with separate regressions by groups to determine whether the continuous variable is statistically significant in all groups. Let me illustrate a minor problem with this second step.

Note on p. 146, Figure 7.15, the coefficient associated with CBM_CEN in the lower half of the table of coefficients. This is the regression coefficient for the continuous variable with the cross product in the regression equation. You know from the explanation of this table of coefficients that this coefficient is also equal to the coefficient you would get for the group scored zero (girls, in this case) if you were to conduct a separate regression for girls and boys. The value (.058, or 5.84E-02, as shown in the Figure) is the same in this table as it is in the table of coefficients for the separate regressions shown in Figure 7.17 (the top table in this Figure, the regression for girls). The coefficients match.

Note, however, that the standard errors associated with this coefficient differ in the two tables. It is .103 in Figure 7.15 versus .100 in Figure 7.17. The difference is slight, but it could make a difference in whether the coefficient is statistically significant or not.

Which value is correct? The first one is; the second is incorrect because the degrees of freedom are incorrect. When we conducted separate regressions we told SPSS to just analyze the data for girls, and then just analyze the data for boys, and thus for each regression only half the data were used. As a result, the *df* used in calculating the standard errors (*N*-*k*-1) were wrong. This does not really matter for the girls regression, because we have the correct *SE* shown in Figure 7.15. But the value for the boys' regression shown in Figure 7.17 is also incorrect, and our original regression did not provide the correct standard error for boys.

This last conundrum—the fact that we do not have correct SE for the group coded 1 in the dummy variable—also provides the clue to how we can obtain it: We simply redo the regression but this time code girls 1 and boys 0. The table of coefficients for such a regression is shown in Figure 1 (at the end of this note). The row for cbm_c in the second half of the table (with the cross product in the equation) shows the value for the regression of CAT scores regressed on CBM scores for the group coded zero. In this regression boys were coded zero, so the value for the regression of CAT on CBM for boys is .974, with a *SE* of .142. Thus, as expected, the table for boys shown in Figure 7.17 in the book shows the correct value for the regression coefficient for boys, but the *SE* is slightly off (.146 instead of .142). You know now how to get the correct *SE*s.

Another way to obtain the regression coefficients, and which also provides the correct *SE*s, is shown in the section below "Testing Separate Slopes in a Single Regression."

Other Methods of Coding Categorical Variables to Create Cross Products

As noted in Chapter 6, there are numerous ways of coding categorical variables into variables that can be analyzed in MR. We covered three: dummy coding, effect coding, and criterion scaling. In Chapter 7 we used dummy coding as the basis for the creation of cross products (interaction terms), but we could have used other coding methods. Two such methods are discussed here.

Testing Interactions Using Effect Coding of Categorical Variables

Effect coding was presented in Chapter 6, and like dummy coding, it could be used as the basis for the creation of cross products to test for interactions (moderation) in multiple regression. Although effect coding and cross products seem less useful when a categorical variable has only two categories (as opposed to three or more), I will use it here with the Kranzler et al simulated data because those are the data we used to illustrate most completely interactions between categorical and continuous variables.

Recall that with effect coding, one group is assigned a value of 1 for the effect-coded variable, others are assigned a value of zero, and one group is assigned a value of -1 on all effect-coded variables. In Chapter 6 we assigned values of -1 to the contrast group, the group that was assigned values of zero across variables when we used dummy coding.

With only two groups (boys and girls) in the test bias example, we would assign values of 1 to one group and values of -1 to the other group. I created such a code in the Kranzler et al simulated data and named it boyse; boys were assigned a value of 1 and girls a value of -1. Figure 2 compares this effect-coded sex variable compared to a dummy-coded version. The cross product variable, created by multiplying boyse x cbm_c (centered CBM scores) is named cbm_boyse_c.

Figure 3 shows the regression results using the effect-coded sex variable and the cross product of that variable and the centered CBM variable. Compare these results with those in the chapter (based on dummy coding). Note that the ΔR^2 associated with the cross product is identical to the value shown in the chapter. It does not matter which method is chosen for coding categorical variables and creating cross products. If we do it correctly, the ΔR^2 associated with the cross-products and the statistical significance of this block will always be the same.

In the table of coefficients, however, note that the *b* values differ from those in the chapter. This makes sense, because the effect coding and the resulting cross product make different comparisons than does dummy coding. Note also that one of the *t* values and its level of statistical significance differs. The take-home lesson is that different coefficients may be significant or not in the table of coefficients (because, in part, different comparisons are being made), but that the statistical significance of the ΔR^2 should remain the same across coding methods, and no matter how many cross-products there are. So, for

example, if we had a three-level categorical variable we would have two dummy or effect-coded variables (*g*-1), and thus two cross-product terms would be needed to test for an interaction. As long as we added both of those cross-products in the second block of the regression, the ΔR^2 should remain the same across coding methods.

What do the coefficients represent? Recall that effect coding produces results that are consistent with the general linear model, with comparisons to the grand mean or the mean of means. And although the various coefficients can be interpreted (see, for example Cohen et al., 2003 for more detail), the interpretation is not as straightforward as when dummy coding is used.

Testing Separate Slopes in a Single Regression

Cohen and colleagues (2003) showed a neat trick that allows both the calculation of the regression equations for the separate groups (which we did when we used dummy coding for the Sex variable) *and* the statistical significance of the slopes for the separate groups. Previously in order to determine the statistical significance of these separate slopes we have conducted post hoc regressions for each group separately (and at the beginning of this note we saw how the *SE*s from these separate regressions are not quite correct)

It is a little tricky to describe, but I hope this description combined with an illustration will make the method clear. In our methodology so far, we have been creating *g*-1 dummy or effect-coded categorical variables. When we multiply those times the centered continuous variables, we also have *g*-1 cross products. In the first block in the regression, we add the coded categorical variable(s) and the centered continuous variable. In block 2, we have added the cross product, or when there are more than two categories to the categorical variable, multiple cross products.

What the Cohen et al. "simple slopes" method does is essentially gets rid of the continuous variable in block 2, but adds *g* (not *g*-1) cross products that include a combination of the cross products and the continuous predictor variable. The first simple slope variable has the same values as the centered continuous variable for the first group, but values of zero for every other group. The second simple slope variable has the same value as the centered continuous variable for the second group, but values of zero for every other group, and so on. Again, the regression includes only the coded categorical variables and the simple slope variables (not the centered continuous variable). The resulting unstandardized coefficients for the simple slopes variables show the coefficients (and the correct *SE*s and statistical significance) for the separate regressions for each group.

Here is how it would work using the Kranzler et al. simulated data. Figure 4 shows a portion of the data with these two new simple slope variables included; these are labeled "boy_cbm" and "girl_cbm." Notice, as described, that for the boys, the values for the boy_cbm variable are equal to the centered CBM variable for boys, but equal to zero for girls. And note that the girl_cbm variable has values equal to the centered CBM variable for girls, but values of zero for the boys.

Figure 5 shows some of the results of the multiple regression of CAT Reading Comprehension test on the dummy coded Sex variable, boy_cbm, and girl_cbm. Note that the *R* and the R^2 are identical to the value

for block 2 of the MR shown in the Chapter (Figure 7.14), .874 and .763, respectively. We could legitimately calculate the ΔR^2 and statistical significance of the interaction term by comparing this value with the value from block 1 of the sequential regression shown in Figure 7.14. It doesn't matter how you enter the interaction terms, if you enter them as a block (and do it correctly), the ΔR^2, *F*, etc will be the same.

The figure also shows the table of coefficients. In this table,

1. The intercept (as in the original dummy-coded analysis) represents the intercept for the group coded zero on the dummy-coded Sex variable (Girls).
2. The *b* for Sex is the difference in intercepts for the group coded 1 on the Sex variable, Boys. Thus the boy intercept for the separate regression equations is 632.07 + 156.11 = 788.18. We got this same information, of course, from the initial analysis.
3. The *b* coefficient for boy_cbm is equal to the value we would obtain for the slope of the regression line if we were to do a separate regression for boys. The table also shows that this value is statistically significant. Note that the SE for this slope is correct (compare it to the value shown in Figure 1 of this note).
4. The *b* coefficient for girl_cbm is equal to the value we would obtain for the slope of the regression line if we were to do a separate regression for girls. Note also that it is not statistically significant.

Once again, we could and did figure out the regression coefficients (intercepts and *b*'s) for the separate regressions for boys and girls from the values shown in the original regression (Figure 7.14), but to get the statistical significance we would have needed to conduct separate regressions for boys and girls (or to have redone the analysis using boys as the reference group).

Cohen and colleagues note that this method is useful when researchers want to know whether or not a particular variable is a statistically significant predictor in every group. I expect that in most cases researchers will still want to conduct the original sequential regression analysis to determine whether the interaction (cross product or cross products with more than two groups) is statistically significant. Thus I expect most of us would use the method as a follow-up test. Alternatively, one could conduct the MR shown in block 1 in Figure 7.14, and then calculate the ΔR^2 (and significance) associated with this MR. Or, instead, one can conduct the separate regressions as a follow-up. Still, it is an elegant method.

Figure 1. Table of coefficients from the regression with boys coded zero and girls coded 1. The row for cbm_c in the lower half of the table shows the coefficient, standard error, etc for The table provides the correct standard error for the coefficient for

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	789.673	7.812		101.091	.000	774.169	805.176
	girlsd girl coded 1 dummy var	-156.826	11.052	-.795	-14.190	.000	-178.761	-134.890
	cbm_c	.372	.094	.222	3.968	.000	.186	.559
2	(Constant)	788.176	6.937		113.622	.000	774.406	801.945
	girlsd girl coded 1 dummy var	-156.110	9.807	-.791	-15.918	.000	-175.577	-136.644
	cbm_c	.974	.142	.581	6.848	.000	.691	1.256
	cbm_girld_c	-.915	.175	-.442	-5.217	.000	-1.263	-.567

a. Dependent Variable: cat California Achievement Test, Reading Comprehension

Figure 2. Effect coding of the Sex variable compared to dummy coding in the Kranzler et al. simulated data.

boysd * boyse boys, effect coded Crosstabulation

Count

		boyse boys, effect coded		
		-1.00	1.00	Total
boysd	.00	50	0	50
	1.00	0	50	50
Total		50	50	100

Figure 3. Regression results with effect coding used as the basis for creating cross products.

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	boyse boys, effect coded, cbm_c[b]	.	Enter
2	cbm_boyse_c[b]	.	Enter

a. Dependent Variable: cat California Achievement Test, Reading Comprehension

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.834[a]	.696	.690	55.21114	.696	111.123	2	97	.000
2	.874[b]	.763	.756	48.98666	.067	27.217	1	96	.000

a. Predictors: (Constant), boyse boys, effect coded, cbm_c

b. Predictors: (Constant), boyse boys, effect coded, cbm_c, cbm_boyse_c

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	711.260	5.521		128.825	.000
	cbm_c	.372	.094	.222	3.968	.000
	boyse boys, effect coded	78.413	5.526	.795	14.190	.000
2	(Constant)	710.121	4.904		144.818	.000
	cbm_c	.516	.088	.308	5.883	.000
	boyse boys, effect coded	78.055	4.904	.791	15.918	.000
	cbm_boyse_c	.458	.088	.273	5.217	.000

a. Dependent Variable: cat California Achievement Test, Reading Comprehension

Figure 4. A portion of the Kranzler et al simulated data with the addition of the new "simple slope" variables.

File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help

1 : sex .0

	cbm	sex	cat	cbm_cen	boy_cbm	girl_cbm	var	var
31	248.00	.00	602.00	126.71	.00	126.71		
32	3.00	.00	622.00	-118.29	.00	-118.29		
33	163.00	.00	673.00	41.71	.00	41.71		
34	100.00	.00	644.00	-21.29	.00	-21.29		
35	7.00	.00	620.00	-114.29	.00	-114.29		
36	81.00	.00	578.00	-40.29	.00	-40.29		
37	148.00	.00	637.00	26.71	.00	26.71		
38	120.00	.00	688.00	-1.29	.00	-1.29		
39	83.00	.00	606.00	-38.29	.00	-38.29		
40	174.00	.00	632.00	52.71	.00	52.71		
41	106.00	.00	615.00	-15.29	.00	-15.29		
42	190.00	.00	732.00	68.71	.00	68.71		
43	60.00	.00	653.00	-61.29	.00	-61.29		
44	63.00	.00	682.00	-58.29	.00	-58.29		
45	191.00	.00	652.00	69.71	.00	69.71		
46	97.00	.00	655.00	-24.29	.00	-24.29		
47	193.00	.00	499.00	71.71	.00	71.71		
48	64.00	.00	570.00	-57.29	.00	-57.29		
49	41.00	.00	612.00	-80.29	.00	-80.29		
50	108.00	.00	658.00	-13.29	.00	-13.29		
51	116.00	1.00	809.00	-5.29	-5.29	.00		
52	147.00	1.00	817.00	25.71	25.71	.00		
53	31.00	1.00	760.00	-90.29	-90.29	.00		
54	123.00	1.00	820.00	1.71	1.71	.00		
55	149.00	1.00	892.00	27.71	27.71	.00		
56	137.00	1.00	781.00	15.71	15.71	.00		
57	200.00	1.00	810.00	78.71	78.71	.00		
58	117.00	1.00	779.00	-4.29	-4.29	.00		
59	112.00	1.00	772.00	-9.29	-9.29	.00		
60	131.00	1.00	821.00	9.71	9.71	.00		
61	29.00	1.00	736.00	-92.29	-92.29	.00		
62	107.00	1.00	784.00	-14.29	-14.29	.00		
63	145.00	1.00	781.00	23.71	23.71	.00		
64	75.00	1.00	811.00	-46.29	-46.29	.00		
65	161.00	1.00	796.00	39.71	39.71	.00		
66	182.00	1.00	910.00	60.71	60.71	.00		
67	110.00	1.00	792.00	-11.29	-11.29	.00		

Data View Variable View

Figure 5. Regression results using the simple slopes method.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.874[a]	.763	.756	48.98666

a. Predictors: (Constant), sex, girl_cbm girl simple slope, boy_cbm boy simple slope

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	632.065	6.932		91.174	.000	618.305	645.826
	boy_cbm boy simple slope	.974	.142	.340	6.848	.000	.691	1.256
	girl_cbm girl simple slope	.058	.103	.028	.568	.571	-.146	.262
	sex	156.110	9.807	.791	15.918	.000	136.644	175.577

a. Dependent Variable: cat California Achievement Test, Reading Comprehension