Chapter 7: Categorical & Continuous Variables

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. For comparison purposes, the relevant regression results from the chapter (with dummy coding) are shown in Figure 1.

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 girls_eff; girls were assigned a value of 1 and boys 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 girls_eff x cbm_cen (centered CBM scores) is named cbm_girleff.

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.

Follow-up for a Statistically Significant Interaction

In the chapter I suggested that when you encounter a statistically significant interaction (cross product) between a categorical and continuous variable that you should graph it to understand the nature of the interaction. You may also want to determine whether the continuous variable is statistically significant in all groups. As noted in the text, we can get the correct regression coefficients for all the groups from the overall regression, but the SEs and statistical significance are incorrect for the group coded 1 on the dummy variable. I suggested that if this information was needed for follow-up that a simple way of obtaining that information was to simply redo the regression and recode the dummy variable in the opposite direction.

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

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. Because the overall regression tells you the correct coefficients for both groups, but the correct SEs only for the group coded 0, I suggested in the text that if you want to determine the statistical significance of these separate slopes the easiest way to do so was to redo the regression, with a reversed dummy variable and a new cross-product. This "simple slopes" method from Cohen and colleagues is a more elegant method of gaining that information.

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 "cbm_boy" and "cbm_girl." Notice, as described, that for the boys, the values for the cbm_boy variable are equal to the centered CBM variable for boys, but equal to zero for girls. And note that the cbm_girl 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, cbm_boy, and cbm_girl. Note that the *R* and the R^2 are identical to the value for block 2 of the MR shown in the Chapter (Figure 7.15), .556 and .309, 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 7.15. 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 (Boys).
2. The *b* for Girl is the difference in intercepts for the group coded 1 on the Girl variable (Girls). Thus the girl intercept for the separate regression equations is 675.571 – 20.014 = 655.557. We got this same information, of course, from the initial analysis.
3. The *b* coefficient for cbm_boy 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 not statistically significant. Note that the SE for this slope is correct. Note also that this value and its SE are the same as shown in the text (because boys were coded 0 in the original anlaysis).
4. The *b* coefficient for cbm_girl 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. The SE is different, however (and is correct in the present analysis). Note also that it is 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 7.15), but to get the statistical significance we would have needed to redo the analysis using girls 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. Still, it is an elegant method.

Figure 1. Table of coefficients from the regression with boys coded zero and girls coded 1 (from Chapter 7). The row for cbm_cen in the lower half of the table shows the coefficient, standard error, etc for group coded 0 on the dummy variable (boys). The table provides the correct standard error for the coefficient for boys, but not for girls.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	677.176	5.125		132.140	.000	667.005	687.347
	cbm_cen	.317	.064	.446	4.941	.000	.189	.444
	Girl Sex, girls=1	-19.675	7.289	-.244	-2.699	.008	-34.141	-5.209
2	(Constant)	675.571	4.891		138.117	.000	665.862	685.280
	cbm_cen	.129	.082	.182	1.570	.120	-.034	.292
	Girl Sex, girls=1	-20.014	6.925	-.248	-2.890	.005	-33.760	-6.268
	sex_cbm	.414	.122	.391	3.389	.001	.172	.657

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

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

Girl Sex, girls=1 * Girl_eff Girl, effect coded Crosstabulation

Count

		Girl_eff Girl, effect coded		
		-1.00	1.00	Total
Girl Sex, girls=1	.00 Boys	50	0	50
	1.00 Girls	0	50	50
Total		50	50	100

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

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	.475[a]	.226	.210	36.02877	.226	14.163	2	97	.000
2	.556[b]	.309	.287	34.22655	.083	11.484	1	96	.001

a. Predictors: (Constant), cbm_cen, Girl_eff Girl, effect coded

b. Predictors: (Constant), cbm_cen, Girl_eff Girl, effect coded, CBM_girleff

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	36770.411	2	18385.206	14.163	.000[b]
	Residual	125912.997	97	1298.072		
	Total	162683.408	99			
2	Regression	50223.551	3	16741.184	14.291	.000[c]
	Residual	112459.857	96	1171.457		
	Total	162683.408	99			

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

b. Predictors: (Constant), cbm_cen, Girl_eff Girl, effect coded

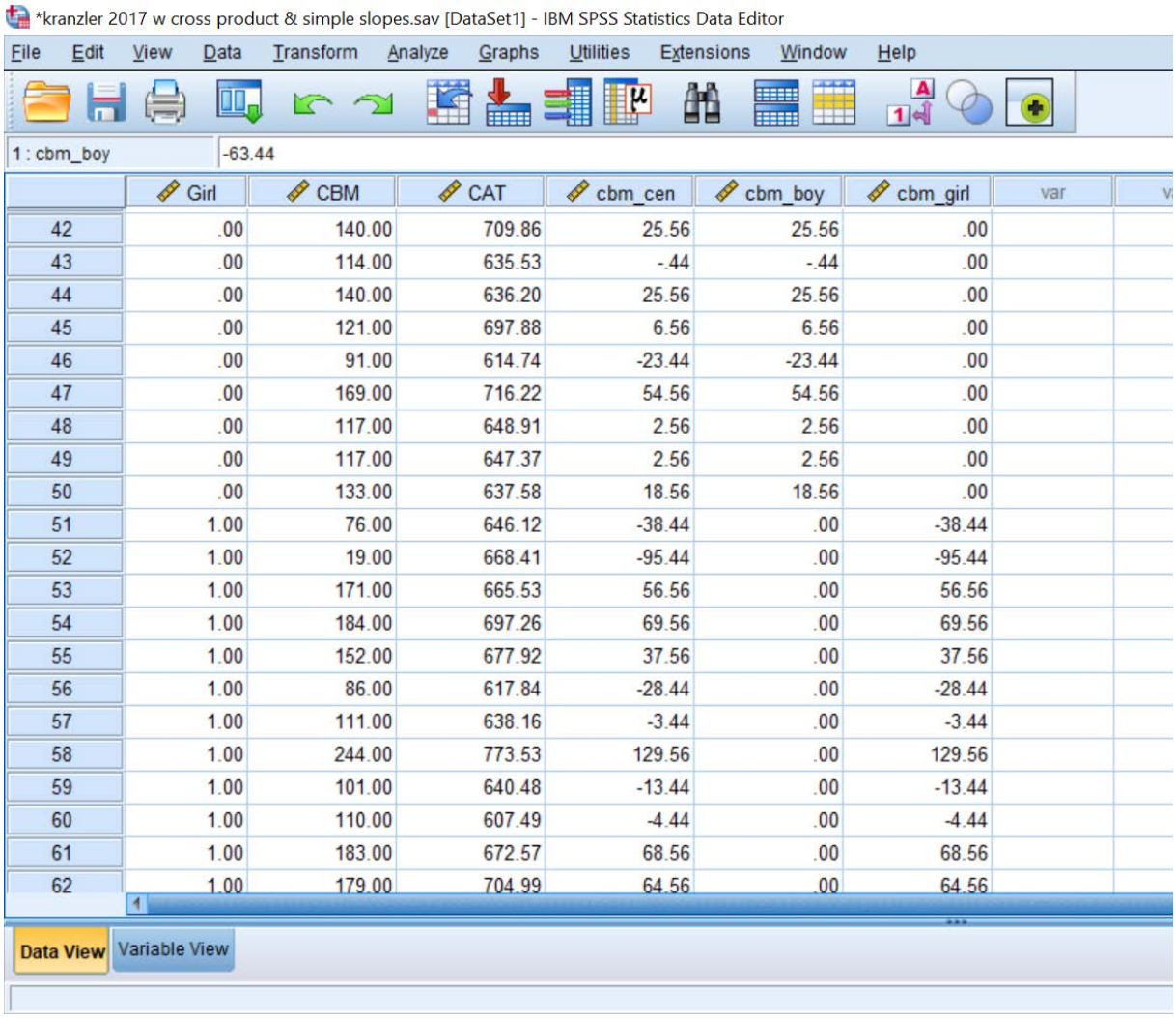
c. Predictors: (Constant), cbm_cen, Girl_eff Girl, effect coded, CBM_girleff

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	667.338	3.603		185.224	.000
	Girl_eff Girl, effect coded	-9.838	3.644	-.244	-2.699	.008
	cbm_cen	.317	.064	.446	4.941	.000
2	(Constant)	665.564	3.462		192.223	.000
	Girl_eff Girl, effect coded	-10.007	3.462	-.248	-2.890	.005
	cbm_cen	.336	.061	.474	5.501	.000
	CBM_girleff	.207	.061	.289	3.389	.001

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.



	Girl	CBM	CAT	cbm_cen	cbm_boy	cbm_girl	var
42	.00	140.00	709.86	25.56	25.56	.00	
43	.00	114.00	635.53	-.44	-.44	.00	
44	.00	140.00	636.20	25.56	25.56	.00	
45	.00	121.00	697.88	6.56	6.56	.00	
46	.00	91.00	614.74	-23.44	-23.44	.00	
47	.00	169.00	716.22	54.56	54.56	.00	
48	.00	117.00	648.91	2.56	2.56	.00	
49	.00	117.00	647.37	2.56	2.56	.00	
50	.00	133.00	637.58	18.56	18.56	.00	
51	1.00	76.00	646.12	-38.44	.00	-38.44	
52	1.00	19.00	668.41	-95.44	.00	-95.44	
53	1.00	171.00	665.53	56.56	.00	56.56	
54	1.00	184.00	697.26	69.56	.00	69.56	
55	1.00	152.00	677.92	37.56	.00	37.56	
56	1.00	86.00	617.84	-28.44	.00	-28.44	
57	1.00	111.00	638.16	-3.44	.00	-3.44	
58	1.00	244.00	773.53	129.56	.00	129.56	
59	1.00	101.00	640.48	-13.44	.00	-13.44	
60	1.00	110.00	607.49	-4.44	.00	-4.44	
61	1.00	183.00	672.57	68.56	.00	68.56	
62	1.00	179.00	704.99	64.56	.00	64.56	

Figure 5. Regression results using the simple slopes method.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.556[a]	.309	.287	34.22655

a. Predictors: (Constant), cbm_girl, cbm_boy, Girl

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)	675.571	4.891		138.117	.000	665.862	685.280
	Girl	-20.014	6.925	-.248	-2.890	.005	-33.760	-6.268
	cbm_boy	.129	.082	.134	1.570	.120	-.034	.292
	cbm_girl	.544	.091	.513	6.006	.000	.364	.723

a. Dependent Variable: CAT