APPENDIX 17.2: MEAN-CENTERING VARIABLES AND *VIF*

In Chapter 17, we saw that subtracting the mean from a set of scores produces a new set of scores that has a mean of 0. This new set of scores is said to be a *mean-centered* version of the original set of scores. Researchers often mean-center their predictor variables when testing for moderation. Two reasons for this were mentioned. The first is that mean-centering variables makes it somewhat easier to interpret the regression equation when the interaction term is included. The second is that many researchers believe that mean-centering reduces problems with multicollinearity. We will discuss these two issues in turn, making use of the introductory example described in Figure 17.16.

Mean-Centering to Make Sense of the Regression Coefficients

In Chapter 17, we introduced an example in which the regression equation that predicts *y* from *x*, *mod*, and *xm* was

$$\hat{y} = 657.69 - 8.58(x) - 55.65(mod) + 0.96(xm).$$

The partial regression coefficient $b_x = -8.58$ is the slope of the regression line parallel to the *x*-axis when $mod = 0$, and $b_{\text{mod}} = -55.65$ is the slope of the regression line parallel to the *mod*-axis when $x = 0$. These values are not particularly helpful because $x = 0$ and $mod = 0$ are outside the ranges of our predictors.

Mean-centering *x* shifts these scores so that they are centered on 0 rather than m_x, and mean-centering *mod* shifts these scores so that they are centered on 0 rather than m_{mod}. The regression equation that predicts *y* from x_c, mod_c, and $x_c m_c$ is

$$\hat{y} = 203.02 + 1.02(x_c) + 39.88(m_c) + 0.96(x_c m_c).$$

Figure 17.A2.1 shows the effect of mean-centering the scores in *x* and *mod*. If you compare Figure 17.A2.1 with Figure 17.16, you will see that nothing has changed except that all *x* and *mod* scores have been shifted to be centered on $[x_c = 0, mod_c = 0]$, rather than $[m_x = 99.489, m_{mod} = 10]$. The same is true of the regression surface.

The gray lines in Figure 17.A2.1 show the values of the regression equation when $x_c = 0$ and $mod_c = 0$. When $mod_c = 0$, the regression line relating *x* to *y* is

$$\hat{y} = 203.02 + 1.02(x_c).$$

When $x_c = 0$, the regression line relating *mod* to *y* is

$$\hat{y} = 203.02 + 39.88(m_c).$$

Some might see mean-centering as worth the effort because we can now see the slope of the regression equation passing through the mean of *x* and mean of *mod*. However, these slopes could have been calculated as easily from the original regression equation. The slope of the regression equation for any given value of *mod* is

$$b = b_x + b_{xm}(mod).$$

Therefore, to determine the slope of the regression line in the *x* direction at the mean of the moderator, we simply fill in the required quantities (from the original equation) to find

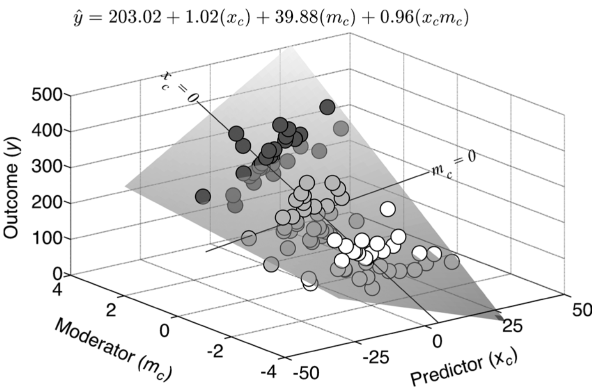
$$b = b_x + b_{xm}(mod\ ) = -8.58 + 0.96(10) = 1.02.$$

Similarly, to determine the slope of the regression plane in the *mod* direction at the mean of the predictor, we simply fill in the required quantities (from the original equation) to find

$$b = b_x + b_{xm}(mod) = -55.65 + 0.96(99.49) = 39.86,$$

which differs from the centered coefficient (39.88) only by rounding error.

FIGURE 17.A2.1 ■ Mean Centering *x* and *mod*



The regression surface with an interaction term ($x_c m_c$) when the predictor (x_c) and moderator (mod_c) have been centered.

Mean-Centering and Collinearity

In Chapter 16, we noted that regression equations might become unstable if the predictor variables are too highly correlated with each other; we called this problem multicollinearity. We saw that SPSS uses *tol* and *VIF* as multicollinearity diagnostics. These reciprocally related measures (*VIF* = 1/*tol*) reflect how much of the variance in a given predictor is explained by the remaining predictors. The question of multicollinearity arises in moderation analysis because the interaction term (*xm*) is always highly correlated with *x* and *mod*, and the multicollinearity diagnostics can go through the roof as a result.

To see the effects of mean-centering (or not) on *tol* and *VIF*, consider the two hierarchical regressions shown in Figure 17.A2.2. Figure 17.A2.2a shows the coefficients table obtained with uncentered predictors and interaction term; *x* and *mod* were entered on step 1 and *xm* was added on step 2. Figure 17.A2.2b shows the coefficients table obtained with centered predictors and interaction term; x_c and mod_c were entered on step 1 and x_cm_c was added on step 2.

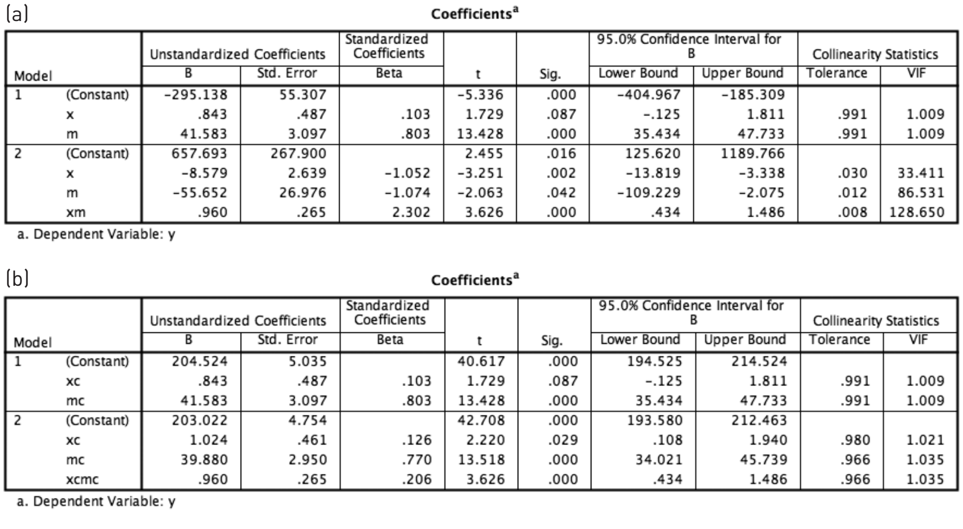
Figure 17.A2.2a shows that on step 1, *tol* and *VIF* were .991 and 1.009 for both *x* and *m*. These indicate no problem with collinearity, and they actually show that there is very little correlation between *x* and *mod*. However, when *xm* is added on step 2, the *tol* and *VIF* diagnostics change dramatically. The *tol* value associated with *x* is .03, which means that the proportion of variability in *x* explained by *mod* and *xm* together is 1 – *tol* = .97. The *tol* values are even smaller for *x* and *xm*.

Figure 17.A2.2b shows that on step 1, *tol* and *VIF* were .991 and 1.009 for both x_c and mod_c. These are the same values seen on step 1 in Figure 17.A2.2a. When x_cm_c is added on step 2, the *tol* and *VIF* diagnostics change very little. The *tol* value associated with x_c is .98. This means that the proportion of variability in x_c explained by mod_c and x_cm_c is 1 – *tol* = .02. Because the *tol* and *VIF* diagnostics do not exceed the rule-of-thumb criteria, mean-centering is often seen as a cure for multicollinearity.

Although centering is widely treated as a cure for multicollinearity, centering usually makes no practical difference to what the regression equation conveys in a moderation analysis. For example, note that in Figure 17.A2.2, the coefficients on *xm* and x_cm_c, their estimated standard errors (0.265), *t*-values (3.626), *p*-values (.000), and confidence intervals [0.434, 1.486] are identical. Therefore, centering has no effect on the interaction term, which is really the main focus of moderation analysis.

In addition, centering has no effect on the model statistics, as shown in Figure 17.A2.3. The Model Summaries in Figure 17.A2.3 correspond to the same

FIGURE 17.A2.2 ■ Regression Coefficients With and Without Mean-Centering



(a)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	-295.138	55.307		-5.336	.000	-404.967	-185.309		
	x	.843	.487	.103	1.729	.087	-.125	1.811	.991	1.009
	m	41.583	3.097	.803	13.428	.000	35.434	47.733	.991	1.009
2	(Constant)	657.693	267.900		2.455	.016	125.620	1189.766		
	x	-8.579	2.639	-1.052	-3.251	.002	-13.819	-3.338	.030	33.411
	m	-55.652	26.976	-1.074	-2.063	.042	-109.229	-2.075	.012	86.531
	xm	.960	.265	2.302	3.626	.000	.434	1.486	.008	128.650

a. Dependent Variable: y

(b)

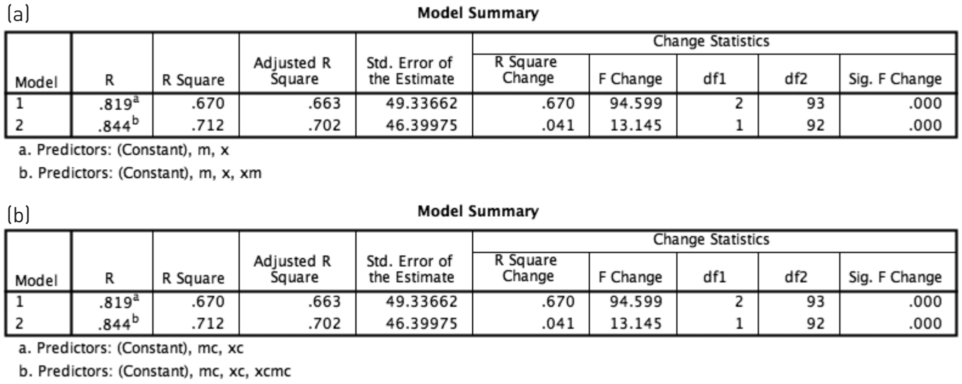
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	204.524	5.035		40.617	.000	194.525	214.524		
	xc	.843	.487	.103	1.729	.087	-.125	1.811	.991	1.009
	mc	41.583	3.097	.803	13.428	.000	35.434	47.733	.991	1.009
2	(Constant)	203.022	4.754		42.708	.000	193.580	212.463		
	xc	1.024	.461	.126	2.220	.029	.108	1.940	.980	1.021
	mc	39.880	2.950	.770	13.518	.000	34.021	45.739	.966	1.035
	xcmc	.960	.265	.206	3.626	.000	.434	1.486	.966	1.035

a. Dependent Variable: y

Hierarchical regressions (a) with and (b) without mean-centering of *x* and *mod*. Mean-centering of *x* and *mod* makes no difference to the statistics of the interaction term (*x* or x_cm_c).

FIGURE 17.A2.3 ■ Model Summaries With and Without Mean-Centering



(a)

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	.819[a]	.670	.663	49.33662	.670	94.599	2	93	.000
2	.844[b]	.712	.702	46.39975	.041	13.145	1	92	.000

a. Predictors: (Constant), m, x

b. Predictors: (Constant), m, x, xm

(b)

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	.819[a]	.670	.663	49.33662	.670	94.599	2	93	.000
2	.844[b]	.712	.702	46.39975	.041	13.145	1	92	.000

a. Predictors: (Constant), mc, xc

b. Predictors: (Constant), mc, xc, xcmc

Hierarchical regressions (a) with and (b) without mean-centering of x and m. Mean-centering of x and m makes no difference to the statistics of the models.

two hierarchical regressions shown in Figure 17.A2.2. The Model Summary in Figure 17.A2.3a was obtained with x, *mod*, and xm, whereas the Model Summary in Figure 17.A2.3b was obtained with x_c, mod_c, and $x_c m_c$. A quick look at these two tables shows that they are identical. Therefore, if our focus is on whether there is evidence of an interaction between x and *mod*, mean-centering the predictors makes absolutely no difference to our conclusions.

Reconsidering *VIF*

Statistics texts commonly caution about a regression coefficient for which *VIF* exceeds 10. The worry is that the inflation of the estimated variance associated with the regression coefficient ($s_{b_i}^2$) makes the estimate unreliable; small changes in the predictor variables may cause large changes in the regression equation. This unreliability would be reflected in a large value of $s_{b_i}^2$, which would mean the estimate of β_i is very imprecise. However, O'Brien (2007) pointed out that *VIF* is not the only thing that affects $s_{b_i}^2$. In fact, $s_{b_i}^2$ may be quite reasonable (small) even if *VIF* is many times greater than 10.

Equation 17.A2.1 shows how $s_{b_i}^2$ can be expressed in terms of *VIF*, as shown in Chapter 17:

$$s_{b_i}^2 = \frac{s_{\text{est}}^2}{ss_{x_i}} VIF. \tag{17.A2.1}$$

VIF tells us how much $s_{b_i}^2$ is inflated relative to what it would be if predictor x_i shared no variance with the other predictors. Notice, however, that $s_{b_i}^2$ is also affected by ss_{x_i}. Because ss_{x_i} is a sum of squared deviations, it will increase as sample size increases. Therefore, ss_{x_i} and *VIF* have opposite effects on $s_{b_i}^2$. As a consequence, $s_{b_i}^2$ may be quite small even if *VIF* exceeds the *rule-of-thumb* criterion of 10.

The inflation of $s_{b_i}^2$ is also counteracted by R^2 itself. As R^2 increases, all things being equal, $s_{b_i}^2$ will get smaller. This can be seen in equation 17.A2.2. The first line shows $s_{b_i}^2$ defined in terms of *VIF*. The second line replaces s_{est}^2 with its definition, which includes the term $1 - R^2$. In the last line of equation 17.A2.2, $1 - R^2$ has been moved beside *VIF* for clarity.

$$\begin{aligned} s_{b_i}^2 &= \frac{s_{\text{est}}^2}{ss_{x_i}} VIF \\ &= \frac{(1-R^2)s_y^2 \dfrac{n-1}{n-k-1}}{ss_{x_i}} VIF \\ &= \frac{s_y^2 \dfrac{n-1}{n-k-1}}{ss_{x_i}} VIF(1-R^2). \end{aligned} \tag{17.A2.2}$$

Equation 17.A2.2 makes clear that $1 - R^2$ and *VIF* have opposite effects on $s_{b_i}^2$; as R^2 increases, $1 - R^2$ decreases.

These two considerations show that the *rule-of-thumb* criterion of 10 is a blunt instrument (like $p < .05$) that should not be used indiscriminately.

Finally, let's return to the question of mean-centering variables. Although many researchers use mean centering in moderation to reduce multicollinearity and make regression equations more stable, Echambadi and Hess (2007) show that mean-centering variables does not actually achieve this. Echambadi and

Hess point out that multicollinearity is related to the covariance matrix for the predictor variables. As we saw in Chapter 13, the covariance for variables x_1 and x_2 is

$$\mathrm{cov}_{x_1 x_2} = \frac{\Sigma(x_1 - m_{x_1})(x_2 - m_{x_2})}{n-1}. \qquad (17.\mathrm{A}2.3)$$

The covariance matrix is a square matrix, like a correlation matrix, whose entries are covariances rather than correlations. The degree to which columns of a covariance matrix can be predicted from each other gives rise to a measure called the determinant (or det). If the determinant is very small, then the matrix is said to be singular. Echambadi and Hess show that the determinant for the covariance matrix for centered and uncentered variables is the same, so mean-centering does not address the question of multicollinearity. They conclude their article with the following comment:

> As shown in our analytical results, compared with uncentered models, mean-centering does not change the computational precision of parameters, the sampling accuracy of the main effects, simple effects, interaction effects, or the overall model R^2. Therefore, it is clear that mean-centering does not alleviate collinearity problems in moderated regression models. (p. 443)

In summary, *VIF* greater than 10 does not necessarily imply an unstable regression equation, and mean-centering variables does not provide more insight into the nature of the moderator.

References

Echambadi, R., & Hess, J. D. (2007). Mean-centering does not alleviate collinearity problems in moderated multiple regression models. *Marketing Science*, *26*(3), 438–445. doi:10.1287/mksc.1060.0263

O'Brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity*, *41*(5), 673–690. doi:10.1007/s11135-006-9018-6