MARKETING SCIENCE
Vol. 26, No. 3, May–June 2007, pp. 438–445
ISSN 0732-2399 | EISSN 1526-548X | 07 | 2603 | 0438

DOI 10.1287/mksc.1060.0263

Mean-Centering Does Not Alleviate Collinearity Problems in Moderated Multiple Regression Models

Raj Echambadi
Department of Marketing, College of Business Administration, University of Central Florida, P.O. Box 161400, Orlando, Florida 32816-1400, rechambadi@bus.ucf.edu

James D. Hess
Department of Marketing and Entrepreneurship, C.T. Bauer College of Business, 375H Melcher Hall, University of Houston, Houston, Texas 77204, jhess@uh.edu

The cross-product term in moderated regression may be collinear with its constituent parts, making it difficult to detect main, simple, and interaction effects. The literature shows that mean-centering can reduce the covariance between the linear and the interaction terms, thereby suggesting that it reduces collinearity. We analytically prove that mean-centering neither changes the computational precision of parameters, the sampling accuracy of main effects, simple effects, interaction effects, nor the R^2. We also show that the determinants of the cross product matrix $X'X$ are identical for uncentered and mean-centered data, so the collinearity problem in the moderated regression is unchanged by mean-centering. Many empirical marketing researchers commonly mean-center their moderated regression data hoping that this will improve the precision of estimates from ill conditioned, collinear data, but unfortunately, this hope is futile. Therefore, researchers using moderated regression models should not mean-center in a specious attempt to mitigate collinearity between the linear and the interaction terms. Of course, researchers may wish to mean-center for interpretive purposes and other reasons.

Key words: moderated regression; mean-centering; collinearity
History: This paper was received September 21, 2005, and was with the authors 3 months for 2 revisions; processed by Dawn Iacobucci.

1. Introduction

Multiple regression models with interactions, also known as moderated models, are widely used in marketing and have been the subject of much scholarly discussion (Sharma et al. 1981, Irwin and McClelland 2001). The interaction (or moderator) effect in a moderated regression model is estimated by including a cross-product term as an additional exogenous variable as in

$$y = \alpha_1' x_1 + \alpha_2' x_2 + x_1' \alpha_3 x_2 + \alpha_0 + \alpha_c' x_c + \varepsilon, \tag{1}$$

where α_i and x_i are $k_i \times 1$ column vectors for $i = 1, 2$, α_3 is a $k_1 \times k_2$ matrix of coefficients that determine the interaction terms, and x_c plays the role of other covariates that are not part of the moderated element. The moderator term, $x_1'\alpha_3 x_2$, is likely to covary to some degree with the variable x_1 (and with the variable x_2). This relationship has been interpreted as a form of multicollinearity, and collinearity makes it difficult to distinguish the separate effects of the linear and interaction terms involving x_1 and x_2.

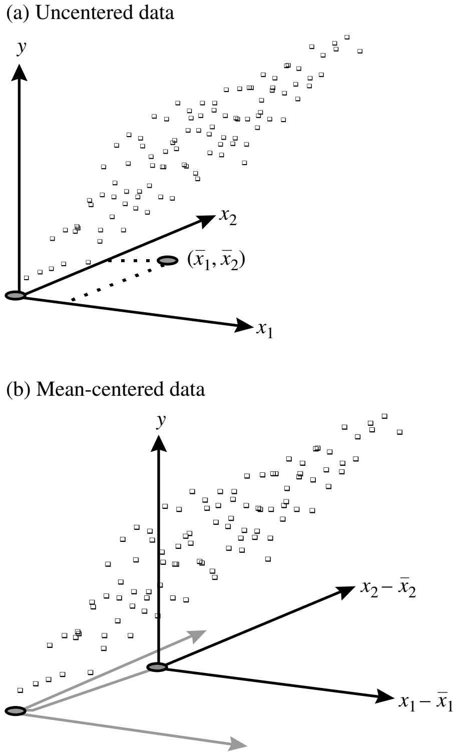
In response to this problem, various researchers including Aiken and West (1991), Cronbach (1987), and Jaccard et al. (1990) recommend mean centering the variables x_1 and x_2 as an approach to alleviating collinearity related concerns. Mean centering (1) gives the following:

$$y = \beta_1'(x_1 - \bar{x}_1) + \beta_2'(x_2 - \bar{x}_2) + (x_1 - \bar{x}_1)'\beta_3(x_2 - \bar{x}_2) + \beta_0 + \beta_c' x_c + \upsilon. \tag{2}$$

In comparison to Equation (1), the linear term $x_1 - \bar{x}_1$ in Equation (2) will typically have smaller covariance with the interaction term because the multiplier of $x_1 - \bar{x}_1$ in the interaction term, $\beta_3(x_2 - \bar{x}_2)$, is zero on average.

This practice of mean-centering has become routine in the social sciences. It is common to see statements from marketing researchers such as, "we mean-centered all independent variables that constituted an interaction term to mitigate the potential threat of multicollinearity" (cf. Kopalle and Lehmann 2006). Can such a simple shift in the location of the origin really help us see the pattern between variables? We use a hypothetical example to suggest an answer to this question. Let the true model for this simulated data be: $y = x_1 + (1/2)x_1x_2 + \varepsilon$ where $\varepsilon \sim$N(0, 0.1). In Figure 1(a), we graph the relationship between y and

Figure 1 Graphical Representation of Uncentered and Mean-Centered Data in 3D Variable Space


(a) Uncentered data

(b) Mean-centered data

uncentered (x_1, x_2). In Figure 1(b), we see the relationship between y and mean-centered (x_1, x_2). Obviously, the same pattern of data is seen in both the graphs, since shifting the origin of the exogenous variables x_1 and x_2 does not change the relative position of any of the data points. Intuitive geometric sense tells us that looking for statistical patterns in the centered data will not be easier or harder than in the uncentered data.

In this paper, we will demonstrate analytically that the geometric intuition is correct: mean-centering in moderated regression does not help in alleviating collinearity. Although Belsley (1984) has shown that mean-centering does not help in additive models, to our knowledge, this is the first time anyone has analytically demonstrated that mean-centering does not alleviate collinearity problems in multiplicative models. Specifically, we demonstrate that (1) in contrast to Aiken and West's (1991) suggestion, mean-centering does not improve the accuracy of numerical computation of statistical parameters, (2) it does not change the sampling accuracy of main effects, simple effects, and/or interaction effects (point estimates and standard errors are identical with or without mean-centering), and (3) it does not change overall measures of fit such as R^2 and adjusted-R^2. It does not hurt, but it does not help, not one iota.

The rest of the paper is organized as follows. We prove analytically that mean centering neither improves computational accuracy nor changes the ability to detect relationships between variables in moderated regression. Next, using data from a study of brand extensions, we illustrate the equivalency of the uncentered and the mean-centered models and demonstrate how one set of coefficients and their standard errors can be recovered from the other. Finally, we discuss the reasons why so many marketing scholars mean-center their variables and the conditions under which mean-centering may be appropriate.

2. Mean Centering Neither Helps Nor Hurts

Collinearity can be viewed as a particular form of near linear dependencies among a set of observed variates (Belsley 1991). Increased levels of collinearity in a data set may cause (1) computational problems in the estimation, (2) sampling stability problems wherein insignificant shifts in data may produce significant relative shifts in the estimates, and (3) statistical problems that may exhibit themselves in terms of large variances of the estimates (Cohen and Cohen 1983, p. 115). Considering the first problem, Aiken and West (1991, p. 182) recommend mean-centering because it may help avoid computational problems by reducing roundoff errors in inverting the product matrix.[1] However, McCullough (1999) has demonstrated that even for complex linear regression problems, roundoff errors are extremely rare in modern double-precision, singular value decomposition statistical computing. Addressing the other two problems, Aiken and West (1991) also imply that mean-centering reduces the covariance between the linear and interaction terms, thereby increasing the determinant of $X'X$. This viewpoint that collinearity can be eliminated by centering the variables, thereby reducing the correlations between the simple effects and their multiplicative interaction terms is echoed by Irwin and McClelland (2001, p. 109). We will show that this is incorrect.

Straight-forward algebraic manipulation of Equation (1) shows that it is equivalent to

$$\begin{aligned} y = {} & (\alpha_1 + \alpha_3\bar{x}_2)'(x_1 - \bar{x}_1) + (\alpha_2 + \alpha_3'\bar{x}_1)'(x_2 - \bar{x}_2) \\ & + (x_1 - \bar{x}_1)'\alpha_3(x_2 - \bar{x}_2) + \alpha_0 + \alpha_1'\bar{x}_1 \\ & + \alpha_2'\bar{x}_2 + \bar{x}_1'\alpha_3\bar{x}_2 + \alpha_c'x_c + \varepsilon. \end{aligned} \tag{3}$$

[1] When the determinant of $X'X$ is near zero as it might be with collinear data, the computation of $(X'X)^{-1}$ will eventually lead to division by almost zero (recall $A^{-1} = \text{adj}(A)/|A|$ for a square matrix A), which produces rounding errors that might make estimates computationally unstable. Each computation done at double-precision on a modern computer will be accurate to at least 15 digits of accuracy, but repeated computations can cause the errors to accumulate.

Comparing (2) and (3), there is a linear relationship between the $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ parameter vectors; for example, $\beta_1 = \alpha_1 + \alpha_3\bar{x}_2$. Since α_3 is a matrix, we need to vectorize it to establish the linear relationship. The expression $\text{vec}(A)$ is the vectorization operator that stacks columns on top of one another to form a column vector, and the Kronecker product is denoted $A \otimes B$. A fundamental identity in vectorization is $\text{vec}(ABC) = (C' \otimes A)\,\text{vec}(B)$. Apply this to an interaction term to get $x_1'\alpha_3 x_2 = \text{vec}(x_1'\alpha_3 x_2) = (x_2 \otimes x_1)'\,\text{vec}(\alpha_3) = \text{vec}(\alpha_3)'(x_2 \otimes x_1)$, and apply it to $\alpha_3\bar{x}_2$, to get $\text{vec}(\alpha_3\bar{x}_2) = \text{vec}(I\alpha_3\bar{x}_2) = (\bar{x}_2 \otimes I)'\,\text{vec}(\alpha_3)$.

As a result, the relationship between $\boldsymbol{\beta}$ from the mean-centered model and $\boldsymbol{\alpha}$ from the uncentered model is

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \text{vec}(\beta_3) \\ \beta_0 \\ \beta_c \end{bmatrix} = \begin{bmatrix} I_{k_1} & 0 & \bar{x}_2' \otimes I_{k_1} & 0 & 0 \\ 0 & I_{k_2} & \bar{x}_1' \otimes I_{k_2} & 0 & 0 \\ 0 & 0 & I_{k_1 k_2} & 0 & 0 \\ \bar{x}_1' & \bar{x}_2' & \bar{x}_2' \otimes \bar{x}_1' & 1 & 0 \\ 0 & 0 & 0 & 0 & I \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \text{vec}(\alpha_3) \\ \alpha_0 \\ \alpha_c \end{bmatrix} = W\boldsymbol{\alpha}. \quad (4)$$

where I_{k_i} is an identity matrix of dimensions $k_i \times k_i$. We use I for identity matrices when the dimensions are implied by the context. Reversing roles, $\boldsymbol{\alpha} = W^{-1}\boldsymbol{\beta}$, where

$$W^{-1} = \begin{bmatrix} I & 0 & -\bar{x}_2' \otimes I & 0 & 0 \\ 0 & I & -\bar{x}_1' \otimes I & 0 & 0 \\ 0 & 0 & I & 0 & 0 \\ -\bar{x}_1' & -\bar{x}_2' & \bar{x}_2' \otimes \bar{x}_1' & 1 & 0 \\ 0 & 0 & 0 & 0 & I \end{bmatrix}. \quad (5)$$

Note, that because in general $(A \otimes B)(M \otimes N) = (AM) \otimes (BN)$, it must be that $-\bar{x}_1'(\bar{x}_2' \otimes I) = -([1] \otimes \bar{x}_1')(\bar{x}_2' \otimes I) = -[1]\bar{x}_2' \otimes \bar{x}_1' I = -\bar{x}_2' \otimes \bar{x}_1'$. With this observation it is easy to see that $WW^{-1} = I$, so (5) is the inverse of W. Suppose a data set consists of an $n \times 5$ matrix of explanatory variable values $X \equiv [X_1 \vdots X_2 \vdots X_1 * X_2 \vdots \mathbf{1} \vdots X_c]$, where X_j is a column n-vector of observations of the jth variable, $X_1 * X_2$ is an n-vector whose typical component is $X_{i1}X_{i2}$, and $\mathbf{1}$ is a vector of ones. The empirical version of (1) is therefore $Y = X\boldsymbol{\alpha} + \varepsilon$. This is equivalent to $Y = XW^{-1}W\boldsymbol{\alpha} + \varepsilon = XW^{-1}\boldsymbol{\beta} + \varepsilon$. It is easily seen that $XW^{-1} \equiv [X_1 - \bar{x}_1\mathbf{1} \vdots X_2 - \bar{x}_2\mathbf{1} \vdots (X_1 - \bar{x}_1\mathbf{1}) * (X_2 - \bar{x}_2\mathbf{1}) \vdots \mathbf{1} \vdots X_c]$, the mean-centered version of the data.

An immediate conclusion is that ordinary least squares (OLS) estimates of (1) and (2) produce identical estimated residuals e, and because the residuals are identical, the R^2 for both formulations are identical. This is consistent with Kromrey and Foster-Johnson's (1998) findings based on Monte Carlo simulations that the two models are equivalent. OLS estimators $\mathbf{a} = (X'X)^{-1}X'Y$ and $\mathbf{b} = ((XW^{-1})'(XW^{-1}))^{-1}(XW^{-1})'Y$ are related to each other by $\mathbf{b} = W\mathbf{a}$. Finally, the variance-covariance of the uncentered and centered OLS estimators are $S_{\mathbf{a}} = s^2(X'X)^{-1}$ and $S_{\mathbf{b}} = s^2(W'^{-1}X'XW^{-1})^{-1} = s^2W(X'X)^{-1}W'$, where the estimator of σ^2 is $s^2 = e'e/(n-5)$.

Is the claim, "mean-centering increases the determinant of the $X'X$ matrix, thereby reducing collinearity" true? In the uncentered data, we must invert $X'X$ and in the centered data we must invert $W'^{-1}X'XW^{-1}$. However, mean-centering not only reduces the off-diagonal elements (such as $X_1'X_1 * X_2$), but it also reduces the elements on the main diagonal (such as $X_1 * X_2'X_1 * X_2$), and it has no effect whatsoever on the determinant. Therefore, while Aiken and West (1991) and Irwin and McClelland (2001) show that mean-centering normal random variables reduces the magnitude of the correlations between the simple effects and their interaction terms, the determinants are identical for both the centered and uncentered cases.[2] In other words, there is no new information added to the estimation by mean-centering (Kam and Franzese 2005, p. 58) and hence the collinearity is not reduced or eliminated in mean-centered models.

Theorem 1. *The determinant of the uncentered data product matrix $X'X$ equals the determinant of the centered data product matrix $W'^{-1}X'XW^{-1}$.*

(Proofs of all theorems are relegated to the appendix.) Since the source of computational problems in inverting these matrices is a small determinant, the same computational problems exist for mean-centered data as for uncentered data.

Also, assuming that the random variable ε is normally distributed, the OLS $\mathbf{a}$ is normally distributed with a mean $\boldsymbol{\alpha}$ and variance-covariance matrix $\sigma^2(X'X)^{-1}$. Since $\mathbf{b}$ is a linear combination of these,

[2] A point not often made is that the magnitude of the covariance between X_1 and X_1X_2 can increase with mean-centering of nonsymmetric random variables. The magnitude of this covariance with uncentered data is $|\mu_1\,\text{Cov}(X_1, X_2) + \mu_2\,\text{Var}(X_1) + \text{E}[(X_1 - \mu_1)^2 \cdot (X_2 - \mu_2)]|$, while for mean-centered data its magnitude is $|\text{E}[(X_1 - \mu_1)^2(X_2 - \mu_2)]|$. For normal distributions, third order moments are zero, so by mean-centering the magnitude of the covariance goes from a positive value down to zero. However, for skewed distributions, it is very easy for the magnitude of the covariance to increase. For example, if $E[(X_1 - \mu_1)^2(X_2 - \mu_2)]$ and $\text{Cov}(X_1, X_2)$ are positive but μ_1 and μ_2 are negative, when the variables are mean-centered, the magnitude of the covariance of X_1 and X_1X_2 could increase from near zero to a positive value.

$W\mathbf{a}$, it must be normal with mean $W\boldsymbol{\alpha}$ and an estimated variance-covariance matrix $WS_{\mathbf{a}}W'$. As Aiken and West (1991) have shown, estimation of the interaction term is identical for uncentered and centered data; we repeat this for completeness sake.

Theorem 2. *The OLS estimates of the interaction terms α_3 and β_3, a_3 for* (1) *and b_3 for* (2), *have identical point estimates and standard errors.*

This result generalizes to all other effects as seen in the next three theorems.

Theorem 3. *The main effect of x_1 (β_1 from Equation* (2) *or $\alpha_1 + \alpha_3\bar{x}_2$ from Equation* (3)) *as measured by the OLS estimate b_1 or by the OLS estimate $a_1 + a_3\bar{x}_2$ have identical point estimates and standard errors.*

Note that the coefficient α_1 in Equation (1) is not the main effect of x_1; the "main effect" means the "average effect" of x_1 across all values of x_2, namely $\alpha_1 + \alpha_3\bar{x}_2$. Instead, the coefficient α_1 is the simple effect of x_1 when $x_2 = \mathbf{0}$.[3] Algebraic rearrangement of (4) states that this simple effect can also be measured from the main effects found in the mean-centered Equation (2) since $a_1 = b_1 - b_3\bar{x}_2$.

Theorem 4. *The simple effect of x_1 when $x_2 = 0$ is either α_1 in Equation* (1) *or $\beta_1 - \beta_3\bar{x}_2$ from Equation* (2) *and the OLS estimates of each of these (a_1 for* (1) *and $b_1 - b_3\bar{x}_2$ for* (2)) *have identical point estimates and standard errors.*

Theorem 5. *The simple effect of x_1 when $x_2 = \mathbf{1}$ is either $\alpha_1 + \alpha_3\mathbf{1}$ in Equation* (1) *or $\beta_1 - \beta_3(\mathbf{1} - \bar{x}_2)$ from Equation* (2) *and the OLS estimates of each of these ($a_1 + a_3\mathbf{1}$ for* (1) *and $b_1 - b_3(\mathbf{1} - \bar{x}_2)$ for* (2)) *have identical point estimates and standard errors.*

In summary, while some researchers may believe that mean-centering variables in moderator regression will reduce collinearity between the interaction term and linear terms and will miraculously improve their computational or statistical conclusions, this is not so. We have demonstrated that mean-centering does not improve computational accuracy or change the ability to detect relationships between variables in moderated regression. Therefore, it is evident that if collinearity plagues uncentered data, it will also affect the estimates and standard errors of the mean-centered data, as well. The cure for collinearity with mean-centering is illusory.

3. An Illustration from the Brand Extension Literature

The decision to extend a brand, wherein a current brand name is used to enter a completely different product class, is a strategically critical decision for many firms. As a result, marketing scholars have attempted to explain consumer evaluations of brand extensions to glean insights on why some brand extensions succeed and others fail (see Bottomley and Holden (2001) for a comprehensive review of the brand extension literature). It is believed that brand extension evaluations are based primarily on the interaction of the "perceived quality" of the parent brand with the degree of "perceived fit" between the parent and the extension product categories (Echambadi et al. 2006).

Although the literature has considered three separate measures of perceived fit, we use one such fit measure: perceived substitutability, defined as the extent to which consumers view two product classes as substitutes, for expositional simplicity. Specifically, we test the contingent role of substitutability on the relationship between parent brand quality and consumer evaluations of brand extensions. Based on the findings from the prior literature, we expect that the linear effects of both parent brand quality and substitutability would increase brand extension evaluations. We further expect that, at higher levels of substitutability, the positive relationship between quality and extension evaluations would be further strengthened.

We use this example from Bottomley and Holden's (2001) study to illustrate the equivalency of the uncentered and mean-centered regressions.[4] The estimated equation is

$$\begin{aligned}\text{Evaluations} = {} & \alpha_1\text{Substitute} + \alpha_2\text{Quality} \\ & + \alpha_3\text{Substitute} \times \text{Quality} + \alpha_c X_c + \varepsilon, \qquad (6)\end{aligned}$$

where Evaluations is operationalized by a composite two-item measure of perceived overall quality of the brand extension and the likelihood of trying the extension; Substitute refers to the perceived substitutability; Quality refers to the perceived quality of the parent brand; finally, X_c is a vector of control variables. All variables are measured on a 7-point scale. Similar to Bottomley and Holden (2001), we use OLS to estimate the model.

Table 1 shows the results of the uncentered and mean-centered regression models from a sample of $n = 10,203$ observations. The table has been graphically annotated to demonstrate how one could compute the mean-centered (main-effect) estimates and standard errors for Substitute using only the statistics from the uncentered (simple effect) regression. As

[3] Bold **0** is a vector of all zeroes and bold **1** is the unit-vector of all ones.

[4] We thank Stephen Holden for providing us with seven of the eight data sets used in the analysis by Bottomley and Holden (2001). The data are available at http://mktsci.pubs.informs.org. For details of the data sets used, please see Echambadi et al. (2006).

Table 1 **How to Compute Statistics for Main Effects from an Uncentered Moderated Regression: An Annotated Example of Brand Extension Evaluations**

$$a_1 + \bar{x}_2 a_3 \equiv -\mathbf{0.047} + \mathbf{5.208} * \mathbf{0.034} = \mathbf{0.132} \text{ Main effect of substitute}$$

	Uncentered estimate				Mean-centered estimate		
	Substitute a_1	Quality a_2	Substitute × Quality a_3		Substitute b_1	Quality b_2	Substitute × Quality b_3
Estimate	−**0.04655**	0.250	**0.03426**	Estimate	**0.132**	0.349	0.034
(SE_i)	(0.026)	(0.015)	(0.005)	(SE_i)	**(0.008)**	(0.009)	(0.005)
VIF	13.69	2.90	16.05	VIF	1.31	1.03	1.02
VAR-COV$_a$				VAR-COV$_b$			
a	**0.000656**	—	—	b_1	0.000063	—	—
a_2	0.0003	0.00023	—	b_2	−0.00001	0.000085	—
a_3	−**0.00011**	−0.00005	**0.000021**	b_3	−0.000003	0.000003	0.00002
Variable mean							
$\bar{x}_i$	2.898	**5.2077**					
R^2	0.30			R^2	0.30		

$$\text{VAR}_{a_1+\bar{x}_2 a_3} = \text{VAR}_{a_1} + 2\bar{x}_2 \text{COV}_{a_1 a_3} + \bar{x}_2^2 \text{VAR}_{a_3}$$
$$\equiv \mathbf{0.000656} + \mathbf{2} \times 5.208 \times \mathbf{(-0.00011)} + 5.2077^2 \times \mathbf{0.000021} = \mathbf{0.000062} = \mathbf{0.008^2}$$

Notes. Brand Evaluation $= \alpha_1$ Substitute $= \alpha_2$ Quality $+ \alpha_3$ Substitute × Quality + Controls.

has been proved above, this is completely general. Conversely, estimates and standard errors for the uncentered model could be computed using only the statistics from the mean-centered regression. Both models are equally precise.

As seen from Table 1, both the mean-centered and the uncentered models provided an identical fit to the data, and yielded the same model R^2. As noted by Aiken and West (1991) and shown above, the coefficient (0.034) and the standard error (0.005) of the interaction (highest order) term, and hence the t statistics of this term, are identical with or without mean-centering.

An examination of the linear terms of mean-centered and uncentered models from Table 1 reveals an apparent conflicting story. Results from the uncentered model show that the coefficient of Substitute is significantly negative ($a_1 = -0.047$), implying that higher levels of perceived substitutability of a brand extension leads to lowered brand extension evaluations. This runs counter to the a priori expectation of a positive relationship between substitutability and brand extension evaluations. The researcher might note the large variance inflation factors (*VIF*) in Table 1 and the high correlation (0.91) between Substitute and the Substitute × Quality interaction term in the left portion of Table 2 and conclude that multicollinearity is the cause of this peculiar finding. If the variables were mean-centered, the correlation between Substitute and the interaction term is much lower (0.06), suggesting reduced collinearity.

When the mean-centered variables are used, the estimates confirm the prior expectation that substitutability increases brand extension evaluations ($b_1 = 0.132$). The researcher might believe that this improvement is due to alleviating the ill effects of collinearity, since the correlation between the Substitute and the Substitute × Quality entry is reduced from 0.91 to 0.06 by mean-centering. This explanation, although intuitively appealing, is simply false.

The effects for substitutability measured in these two models are vastly different (simple effects from the uncentered models vis-à-vis main effects from the mean-centered models) and hence, direct comparisons of the corresponding coefficients are inappropriate. The infamous "comparison of apples and oranges" metaphor is appropriate. In the uncentered regression model, the coefficients represent *simple* effects of the exogenous variables; i.e., the effects of each variable when the other variables are at zero. The coefficient for Substitute in this model should therefore be interpreted as the change in brand extension evaluations due to substitutability in the complete absence of parent brand quality. This makes a lot of sense. A highly substitutable extension brand with zero parent brand quality is bound to elicit negative evaluations.

However, when data are mean-centered, the coefficients represent *main* effects of these variables; i.e.,

Table 2 **Correlations Between Variables**

	Uncentered variables			Mean-centered variables		
Substitute	1.00			1.00		
Quality	0.11	1.00		0.11	1.00	
Substitute × Quality	0.91	0.43	1.00	0.06	−0.07	1.00

the effects of each variable when the other variables are at their mean values. The coefficient for Substitute in the mean-centered model should be interpreted as the change in brand extension evaluations due to substitutability for parent brands with average quality. The change in estimates and reduction in standard errors is based entirely on the fact that the coefficients have different substantive meanings.[5] As proved in Theorems 2–5 and demonstrated in Table 1, the same main effect estimates and standard errors could be computed from estimates based upon the uncentered data.

In summary, the uncentered and mean-centered models are statistically equivalent. Using Equation (1) and $\mathbf{b} = W\mathbf{a}$, we can recover an equally accurate measure of the main effect in the mean-centered model from the uncentered data. Similarly, using Equation (2) and $\mathbf{a} = W^{-1}\mathbf{b}$, we can recover an equally accurate measure of the simple effect in the uncentered model from the centered data. As demonstrated in Table 1, the standard errors of coefficients in an uncentered model can also be recovered from the mean-centered data and vice versa.

4. Comments

Why do so many researchers mean-center their moderated variables? Clearly, there is a fear that by including a term x_1x_2 in the regressors, they will create collinearity with the main regressor, such as x_1, so that it will become difficult to distinguish the separate effects of x_1 and x_1x_2 on y. If we make x_1's multiplier in the interaction term, x_2, closer to zero on average, then we reduce the covariance and correlation, and one simple way to do this is to replace the multiplier x_2 by $x_2 - \bar{x}_2$. By subtracting the mean, the typical value of the multiplier is zero and hence the covariance between the regressor and the interaction terms is typically smaller. This appears to reduce the "potential threat of multicollinearity" and hopefully improves the ability to distinguish the effect of changes in x_1 from changes in x_1x_2.

This logic seems plausible, but it is incorrect. The complete analysis of mean-centering indicates that mean-centering often reduces covariances between the linear and the interaction terms; however, it does not add any new information to the estimation. 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.

In the light of these results, it is evident that the decision to mean-center the moderated variable data should be made independently from the specious rationale of trying to alleviate multicollinearity. Could we mean-center to obtain a better interpretation? One might argue that the main effects are more meaningful because they characterize the overall relationships better, so the data should be mean-centered. However, one might also argue that the simple effects are preferable because they provide a more fine-grained understanding of the patterns, so the data should be uncentered. Because both models are mathematically equivalent and the results for the uncentered case can be obtained from the mean-centered model and vice versa, mean-centering does not necessarily provide a better interpretation of the data. It just provides a different interpretation. Of course, recovery of the proper standard errors does require computing the diagonal elements of matrices such as WS_aW' in the uncentered context, or $W^{-1}S_bW^{-1'}$ in the mean-centered case, and this may be accomplished easier by reversing the data-centering decision with a few mouse clicks.

Mean-centering does not hurt, so there is no need to re-evaluate the conclusions of the many published papers that have used mean-centering so long as the researchers are clear about the proper interpretation of the linear terms. The correct interpretation of mean-centered coefficients in moderated regression models has often been overlooked in many marketing papers. For example, a content analysis of papers that employed mean-centering approaches shows that only four of the 70 published papers published in nine major marketing journals from 1994 to 2004 explicitly mentioned the different interpretation for the mean-centered coefficients.

How do we diagnose collinearity? Grewal et al. (2004) caution researchers to be careful about relying on available diagnostics for what constitutes different levels of collinearity. Our content analysis of select marketing journals reveals that bivariate correlations and VIFs are currently the most commonly used tools to diagnose multicollinearity. One drawback of the commonly used collinearity diagnostics is that they evaluate regressors one by one and hence may not be appropriate diagnostics of collinear situations by themselves (Coenders and Saez 2000). High VIFs are sufficient but not necessary to collinearity (Belsley 1991). It is, therefore, imperative that researchers use multiple diagnostic tools to diagnose potential collinearity problems in moderated regression. This reliance on multiple tools may reduce false alarms raised by empirical researchers about the presence of collinearity. If collinearity problems are indeed suspected, Echambadi et al. (2006) suggest randomly estimating subsets of the data to test the plausibility and

[5] For an exposition of the distinction between simple effects and main effects, please refer to Irwin and McClelland (2001).

stability of coefficients. Unstable coefficients across these random subsets of data may confirm the presence of collinearity problems.

A casual check of our journals also reveals that researchers use collinearity diagnostics on mean-centered data to confirm the absence of collinearity problems. However, because mean-centering typically masks the role of the constant term in any underlying dependencies, these diagnostic measures of collinearity, such as the VIF and the correlation matrix, produce meaningless diagnostic information (Belsley 1984, p. 73). Partial and semi-partial correlations advocated by Cohen and Cohen (1983) also suffer from the same failing when applied on mean-centered data. Therefore, as Belsley (1984) suggests, researchers must always use uncentered data for assessing collinearity.

Since mean-centering does not mitigate collinearity in moderated regression, one might ask, "What else can be done in the face of collinearity?" One alternative is to use the residual-centering method proposed by Lance (1988), but this is a distinctly bad idea. Echambadi et al. (2006) show that residual-centering makes the linear effects biased and inconsistent, which is undesirable. Morris et al.'s (1986) principal component regression procedure has also been deemed unacceptable (Cronbach 1987).

Because collinearity problems cannot be remedied after the data has been collected in most cases, we endorse Grewal et al.'s suggestion that researchers carefully design their studies prior to collecting their data. If feasible, one can address it by using a data collection scheme that isolates the interaction effect (for example, a factorial design). Likewise, if feasible, one can address the loss of power associated with multicollinearity by increasing the sample size; in this regard, Woolridge (2001) notes that the effects of multicollinearity are indistinguishable from the effects of micronumerosity, or small sample sizes. More and better data always helps in reducing collinearity (Judge et al. 1988, p. 874).

Summary

Whether we estimate the uncentered moderated regression Equation (1) or the mean-centered Equation (2), all the point estimates, standard errors and t statistics of the main effects, all simple effects, and interaction effects are identical and will be computed with the same accuracy by modern double-precision statistical packages. This is also true of the overall measures of accuracy such as R^2 and adjusted-R^2.

Acknowledgments

The names of the authors are listed alphabetically. This is a fully collaborative work. We thank Inigo Arroniz, Edward A. Blair, Pavan Chennamaneni, and Tran Van Thanh for their many helpful suggestions in crafting this article. Vishal Bindroo's technical assistance is gratefully acknowledged.

Appendix

Proof of Theorem 1. By doing a Laplace expansion down the last column of W,

$$W=\begin{bmatrix} I_{k_1} & 0 & \bar{x}_2'\otimes I_{k_1} & 0 & 0\\ 0 & I_{k_2} & \bar{x}_1'\otimes I_{k_2} & 0 & 0\\ 0 & 0 & I_{k_1k_2} & 0 & 0\\ \bar{x}_1' & \bar{x}_2' & \bar{x}_2'\otimes\bar{x}_1' & 1 & 0\\ 0 & 0 & 0 & 0 & I \end{bmatrix},$$

then the fourth set of columns, then the third set of rows, it is clear that the determinant $|W|=1$. Consequently, $\det(W'^{-1}X'XW^{-1})=\det(W'^{-1})\det(X'X)\det(W^{-1})=\det(W^{-1\prime})\det(X'X)\det(W^{-1})=\det(X'X)$. Q.E.D.

Note: the interaction terms are made into a k_1k_2-vector $x_2\otimes x_1$. Thus in the variance covariance matrix S, entries like S_{31} will have k_1k_2 rows.

Proof of Theorem 2. From the third row of (4), $\text{vec}(b_3)=\text{vec}(a_3)$, so $b_3=a_3$. In this appendix, we will denote $S_{\mathbf{a}}$ by S. Using matrix multiplication of (4), the third set of columns of SW' is

$$\begin{bmatrix} S_{31}\\ S_{32}\\ S_{33}\\ S_{30}\\ S_{3c} \end{bmatrix}.$$

The third set of rows of W is $[0\ 0\ I\ 0\ 0]$, so the 3rd set of row × 3rd set of columns of WSW' is S_{33}. That is, $\text{Var}(\text{vec}(b_3))=\text{Var}(\text{vec}(a_3))$. Q.E.D.

Proof of Theorem 3. From the first row of (4), the point estimates are equal. The first column of SW' is

$$\begin{bmatrix} S_{11}+S_{13}(\bar{x}_2\otimes I)\\ S_{21}+S_{23}(\bar{x}_2\otimes I)\\ S_{31}+S_{33}(\bar{x}_2\otimes I)\\ S_{01}+S_{03}(\bar{x}_2\otimes I)\\ S_{c1}+S_{c3}(\bar{x}_2\otimes I) \end{bmatrix}.$$

Notice that $(\bar{x}_2\otimes I)$ is $k_1k_2\times k_1$. The first row of W is $\lfloor I_{k_1}\ \ 0\ \ \bar{x}_2'\otimes I_{k_1}\ \ 0\ \ 0\rfloor$, so the variance of b_1 (the 1st row × 1st column of WSW') equals $S_{11}+2S_{13}(\bar{x}_2\otimes I)+(\bar{x}_2\otimes I)'S_{33}(\bar{x}_2\otimes I)$. The variance of $a_1+a_3\bar{x}_2=a_1+(\bar{x}_2\otimes I)'\cdot\text{vec}(a_3)$ is $\text{Var}(a_1)+2\text{Cov}(a_1,\ \text{vec}(a_3))(\bar{x}_2\otimes I)+(\bar{x}_2\otimes I)'\cdot\text{Var}(\text{vec}(a_3))(\bar{x}_2\otimes I)=S_{11}+2S_{13}(\bar{x}_2\otimes I)+(\bar{x}_2\otimes I)'S_{33}(\bar{x}_2\otimes I)$. That is, $\text{Var}(b_1)=\text{Var}(a_1+a_3\bar{x}_2)$. Q.E.D.

Proof of Theorem 4. The variance of $b_1-b_3\bar{x}_2=b_1-(\bar{x}_2\otimes I)'\,\text{vec}(b_3)$ equals $\text{Var}(b_1)-2\text{Cov}(b_1,\text{vec}(b_3))(\bar{x}_2\otimes I)+(\bar{x}_2\otimes I)'\,\text{Var}(\text{vec}(b_3))(\bar{x}_2\otimes I)$. From the proof of Theorems 2 and 3 we know that $\mathrm{var}(\mathrm{vec}(b_3)) = S_{33}$ and $\mathrm{var}(b_1) = S_{11} + 2S_{13}(\bar{x}_2 \otimes I) + (\bar{x}_2 \otimes I)' S_{33}(\bar{x}_2 \otimes I)$. The first column of SW' is

$$\begin{bmatrix} S_{11} + S_{13}(\bar{x}_2 \otimes I) \\ S_{21} + S_{23}(\bar{x}_2 \otimes I) \\ S_{31} + S_{33}(\bar{x}_2 \otimes I) \\ S_{01} + S_{03}(\bar{x}_2 \otimes I) \\ S_{c1} + S_{c3}(\bar{x}_2 \otimes I) \end{bmatrix}$$

and the third row of W is $[0\ 0\ I\ 0\ 0]$, so the covariance of b_1 and $\mathrm{vec}(b_3)$ (the 3rd row × 1st column of WSW') is $S_{31} + S_{33}(\bar{x}_2 \otimes I)$. Hence the variance of $b_1 - b_3\bar{x}_2$ equals $S_{11} + 2S_{13}(\bar{x}_2 \otimes I) + (\bar{x}_2 \otimes I)' S_{33}(\bar{x}_2 \otimes I) - 2(S_{31} + S_{33}(\bar{x}_2 \otimes I))(\bar{x}_2 \otimes I) + (\bar{x}_2 \otimes I)' S_{33}(\bar{x}_2 \otimes I) = S_{11.}$ That is, $\mathrm{Var}(a_1) = \mathrm{Var}(b_1 - b_3\bar{x}_2)$. Q.E.D.

Proof of Theorem 5. A variant of the above.

References

Aiken, L. S., S. G. West. 1991. *Multiple Regression: Testing and Interpreting Interactions*. Sage Publications, Newbury Park, CA.

Belsley, D. A. 1984. Demeaning conditioning diagnostics through centering. *Amer. Statistician* **38**(2) 73–77.

Belsley, D. A. 1991. *Conditioning Diagnostics: Collinearity and Weak Data in Regression*. John Wiley & Sons, New York.

Bottomley, P., S. Holden. 2001. Do we really know how consumers evaluate brand extensions? Empirical generalizations based on secondary analysis of eight studies. *J. Marketing Res.* **38**(4) 494–500.

Coenders, G., M. Saez. 2000. Collinearity, heteroscedasticity and outlier diagnostics in regression. Do they always offer what they claim? A. Ferligoj, A. Mrvar, eds. *New Approaches in Applied Statistics, Metodološki Zvezki*. Vol. 16. Faculty of Social Sciences, Ljubljana, Slovenia, 79–94.

Cohen, J., P. Cohen. 1983. *Applied Multiple Regression: Correlation Analysis for Behavioral Sciences*, 2nd ed. Lawrence Erlbaum Associates, Hillsdale, NJ.

Cronbach, L. J. 1987. Statistical tests for moderator variables: flaws in analyses recently proposed. *Psych. Bull.* **102**(3) 414–417.

Echambadi, R., I. Arroniz, W. Reinartz, J. Lee. 2006. Empirical generalizations from brand extension research: how sure are we? *Internat. J. Res. Marketing* **23**(3) 253–261.

Grewal, R., J. A. Cote, H. Baumgartner. 2004. Multicollinearity and measurement error in structural equation models: implications for theory testing. *Marketing Sci.* **23**(4) 519–529.

Irwin, J. R., G. H. McClelland. 2001. Misleading heuristics and moderated multiple regression models. *J. Marketing Res.* **38**(1) 100–109.

Jaccard, J. R., R. Turrisi, C. K. Wan. 1990. *Interaction Effects in Multiple Regression*. Sage Publications, Newbury Park, CA.

Judge, G. G., W. E. Griffiths, R. C. Hill, H. Lutkepohl, T.-C. Lee. 1988. *Introduction to the Theory and Practice of Econometrics*, 2nd ed. John Wiley & Sons, New York.

Kam, C. D., R. J. Franzese. 2005. *Modeling and Interpreting Interactive Hypotheses in Regression Analysis: A Refresher and Some Practical Advice*. University of Michigan Press, Ann Arbor, MI.

Kopalle, P. K., D. R. Lehmann. 2006. Setting quality expectations when entering a market: what should the promise be? *Marketing Sci.* **25**(1) 8–24.

Kromrey, J. D., L. Foster-Johnson. 1998. Mean centering in moderated multiple regression: much ado about nothing. *Ed. Psych. Measurement* **58**(1) 42–67.

Lance, C. E. 1988. Residual centering, exploratory and confirmatory moderator analysis, and decomposition of effects in path models containing interactions. *Appl. Psych. Measurement* **12**(June) 163–175.

McCullough, B. D. 1999. Assessing the reliability of statistical software: Part II. *Amer. Statistician* **53**(2) 149–159.

Morris, J. H., J. D. Sherman, E. R. Mansfield. 1986. Failures to detect moderating effects with ordinary least squares-moderated multiple regression: some reasons and a remedy. *Psych. Bull.* **99** 282–288.

Sharma, S., R. M. Durand, O. Gur-Arie. 1981. Identification and analysis of moderator variables. *J. Marketing Res.* **18**(3) 291–300.

Woolridge, J. M. 2001. *Econometric Analysis of Cross Section and Panel Data*. MIT Press, Cambridge, MA.