

YMC Management Review**Volume 17, No.1, 2024 pp. 9-22****Diagnosing and Mitigating Multicollinearity in Moderated Multiple Regression**

Tsung-Hao Chen *

Associate Professor of Department of Business Administration, Shu-Te University, Kaohsiung city, Taiwan, R.O.C.

* Corresponding author: thchen@stu.edu.tw

Yu-Lun Sun

Graduate School of Business and Administration, Shu-Te University, Kaohsiung city, Taiwan, R.O.C.

Daniel Chan-wei Tsai

Associate Professor of Department of Business Administration, National Pingtung University of Science and Technology, Ping-Tung, Taiwan, R.O.C.

Yi-Mei Huang

Associate Professor of School of Innovation, Design and Technology, Wellington Institute of Technology, Wellington, New Zealand.

Shijun Zhang

Assistant Professor of Department of Business and Commerce, Zhejiang Industry Polytechnic College, Shaoxing, Zhejiang, China.

Abstract

This paper primarily explores the challenges associated with Moderated Multiple Regression Models, particularly how a moderator variable (m) influences the direction or strength of the relationship between an independent variable (x) and a dependent variable (Y). A significant issue arises when there is a high correlation between the independent variable and the moderator, leading to severe multicollinearity. That complicates the accurate estimation of the independent variables' effects on the dependent variable (Myers, 1990).

We develop five moderated multiple regression models with purpose of mitigating the multicollinearity in the analysis. Our empirical findings indicate that three of them perform good tested by the variance inflation factor and condition index. We finally suggest a process of standardizing both independent variable and moderator and

taking the cross multiplication by those two standardized variables before conducting moderated multiple regression analysis.

Keywords: *MMR, VIF, Multicollinearity, CI.*

1. Introduction

The concept of moderation effect refers to the variability in the influence of independent variables (X) on dependent variable (Y) depending on the levels or presence of a moderator variable (M). For example, if the moderator is a continuous variable, the effect of X on Y may change as the values of M increase or decrease. Conversely, if the moderator is a categorical variable, the impact of X on Y will vary according to the specific categories of M considered in the analysis.

In Moderated Multiple Regression Models, the moderator variable (M) is defined as the factor that affects the direction or strength of the relationship between X and Y . This approach to moderated regression analysis, as described by Zedeck (1971), Cohen and Cohen (1983), Aiken and West (1991), and Aguinis (1995), consists of a series of steps for application and interpretation. The multiple regression model is developed as follow:

$$\hat{Y}_i = \beta_0 + \beta_1 \cdot x_i \quad (1)$$

In equation (1), the x represents the independent variable, while the Y stands as the dependent variable. Yet, if there is the other variable m also affects Y , then the equation (1) can be extended as:

$$\hat{Y}_i = \beta_0 + \beta_1 \cdot x_i + \beta_2 \cdot m_i \quad (2)$$

However, if the two variables (x and m) are interacted and affect Y , then the equation (2) is further rewritten as:

$$\hat{Y}_i = \beta_0 + \beta_1 \cdot x_i + \beta_2 \cdot m_i + \beta_3 \cdot x_i \cdot m_i \quad (3)$$

The term $x_i m_i$ signifies the product of the independent variable x and the moderating variable m . If the interaction term significantly impacts the Y , it is interpreted as evidence of a moderating effect (Kleinbanum et al., 1998). However, a high correlation between x and m leads to multicollinearity issues. To circumvent this problem, Aiken and West (1991) recommend individually standardizing x and m before multiplying them, resulting in a modified equation that addresses the multicollinearity concern.

$$\hat{Y}_i = \beta_0 + \beta_1 \cdot x_i + \beta_2 \cdot m_i + \beta_3 \cdot Z_{xi} \cdot Z_{mi} \quad (4)$$

The $Z_{xi} \cdot Z_{mi}$ represents the product of the standardized x and the standardized m . In equation (4), only x and m are standardized before being multiplied, while all other independent, moderator, and dependent variables remain in their original form. This approach suggests further exploration of either standardizing all variables within the independent equation or fully standardizing all variables, including the dependent variable, to compare outcomes and address multicollinearity issues. This study intends to apply the idea of standardization all variables to see if the approach can mitigate the multicollinearity.

2. Research Design

The interaction item, $Z_{xi} \cdot Z_{mi}$, usually leads to severe multicollinearity issues among variables (Myers, 1990). Andy (2009) highlights that multicollinearity arises from a high correlation among independent variables, which complicates the model's ability to make distinct predictions about the dependent variable. Menard (1995) notes that when two or more independent variables are highly correlated, it becomes difficult to accurately estimate their effects on the dependent variable. Such highly correlation among variables leads to a significant increase in the variance inflation factor (VIF) or condition index (CI), or decrease in tolerance levels.

Myers (1990) suggests that multicollinearity can be diagnosed by examining the variance inflation factor (VIF), with a VIF exceeding 10 indicating potential issues. Additionally, a higher condition index (CI) signals more pronounced collinearity. A CI between 30 and 100 suggests moderate to high collinearity in the multiple regression model, while values above 100 indicate severe multicollinearity (Belsley, Kuh, & Welsch, 1980). Tacq (1997) also posits that a CI above 15 may signal collinearity concerns. Subsequent researchers, including Draper and Smith (1998) and Lattin, Carroll, and Green (2003), recommend a CI threshold of over 30 for severe multicollinearity.

The variance inflation factor (VIF) measures the inflation of the variance of the regression coefficients due to collinearity, and tolerance. It is computed as:

$$(\text{VIF})_k = \frac{1}{(1 - R^2)}, K = 1, 2, 3, \dots, p - 1 \quad (5)$$

Here, $(1 - R^2)$ denotes the tolerance, indicating that the variance inflation factor (VIF) and tolerance are inversely related. This stems from the coefficient of determination when the k^{th} variable is regressed against all other variables. A high value of implies low tolerance, suggesting the presence of multicollinearity among the independent variables. This study develops five approaches to test the idea of standardization and examines the performance by the VIF, tolerance, and CI.

2.1 The Original Moderated Multiple Regression Model (MMRM)

The baseline model is the original moderated multiple regression model which is widely applied in the literature.

$$\hat{Y}_i = \beta_0 + \beta_1 \cdot x_i + \beta_2 \cdot m_i + \beta_3 \cdot x_i \cdot m_i \quad (6)$$

This model applies the raw numerical data without considering the correlations among variables as many of literature used. It is expected to encounter the mulitcolinearity issue of Myers (1990), Andy (2009), and Menard (1995).

2.2 The Product of Standardized x and m

Aiken and West (1991) introduce a modified model to address the challenges of multicollinearity, detailed as follows:

$$\hat{Y}_i = \beta_0 + \beta_1 \cdot x_i + \beta_2 \cdot m_i + \beta_3 \cdot Z_{xi} \cdot Z_{mi} \quad (7)$$

They use the x and m in the model as same as those in equation (6). However, the interaction term is formed by the product of standardized x and m .

2.3 Standardization of All Independent Variables

This study goes further to standardize all the independent variables (X) to fit more the assumption of regression model. Therefore, both x and m are all standardized and the model turns to be:

$$\hat{Y}_i = \beta_0 + \beta_1 \cdot Z_{xi} + \beta_2 \cdot Z_{mi} + \beta_3 \cdot Z_{xi} \cdot Z_{mi} \quad (8)$$

The Z_{xi} and Z_{mi} are the standardized x_i and m_i respectively. In this equation (8), all the variables are standardized. The correlation among variables is much lower.

2.4 Standardization of Both Independent and Dependent Variables

More than the standardization of independent variables, this study also standardized the dependent variable. The model is presented as:

$$\widehat{Z}_{yi} = \beta_0 + \beta_1 \cdot Z_{xi} + \beta_2 \cdot Z_{mi} + \beta_3 \cdot Z_{xi} \cdot Z_{mi} \quad (9)$$

Such setting better fit the assumption of regression model that variables are normally distributed. The Z_{yi} is the standardized dependent variable.

2.5 Standardizing Both the Independent and Dependent Variables, with the Interaction Term

An interesting issue is the interaction term should be created before or after the variable standardization. Since the variable distribution of $Z_{xi} Z_{mi}$ does not guarantee a normal form, instead the standardization of $x_i m_i$ does. This study obtains the product of x and m initially and then standardized the product. Our 5th model is written as:

$$\widehat{Z}_{yi} = \beta_0 + \beta_1 \cdot Z_{xi} + \beta_2 \cdot Z_{mi} + \beta_3 \cdot Z_{xi \cdot mi} \quad (10)$$

3. An Illustration of a Moderated Multiple Regression Model

This study uses the demonstration database ex12-5.sav, provided in the textbook 'Advanced Statistical Analysis using SPSS and AMOS' by Kuan Yu, Chen, and Cheng-hua Wang, for a comprehensive explanation. The x represents the landscape imagery average, the m is the switching costs average, and the Y is the loyalty average. The results for our five approaches are reported as follow.

3.1 Model 1: Original Moderated Multiple Regression Equation

The results for the Original Moderated Multiple Regression Equation (equation 6) is as follows:

$$\begin{aligned}\hat{Y}_i &= \beta_0 + \beta_1 x_i + \beta_2 m_i + \beta_3 x_i \cdot m_i \\ &= 0.124 + 0.839x_i + 0.639m_i - 0.097x_i \cdot m_i\end{aligned}$$

Additionally, the standardized equation is presented as follows:

$$\hat{Y}_i = 0.68x_i + 0.759m_i - 0.63x_i \cdot m_i$$

After the estimation, we use both the VIF and CI to identify the multicollinearity. Table 1 reports the VIF and the tolerance for the baseline models.

Table 1
Tests on the tolerance and VIF for baseline model and model 1

model		Unstandardized		Standardized	T	significance	Collinearity Statistics	
		B	standard error	Beta			tolerance	VIF
0	(constant)	2.315	0.361		6.418	0		
	xi	0.355	0.061	0.287	5.849	0	0.992	1.008
	mi	0.203	0.041	0.241	4.898	0	0.992	1.008
1	(constant)	0.124	1.115		0.111	0.912		
	xi	0.839	0.241	0.68	3.482	0.001	0.062	16.048
	mi	0.639	0.214	0.759	2.982	0.003	0.037	27.29
	$xi \cdot mi$	-0.097	0.047	-0.63	-2.075	0.039	0.026	38.772

The model 0 is the model without the interaction term and the model 1 is the equation (6). The interaction term ($xi \cdot mi$) shows a significant effect ($p=0.039<0.05$) on dependent variable, indicating a moderating effect. However, upon adding the interaction term, the VIF for each variable are 16.048, 27.290, and 38.772 for the x , m , and $xi \cdot mi$ respectively. All three VIF values are higher than 10, suggesting a multicollinearity. We further use the CI to diagnose the models in Table 2. The CI value for the interaction term escalates to 78.297, which is higher than the threshold of 30, indicating a severe multicollinearity.

Table 2
Tests on the CI for baseline model and model 1

model	dimension	eigenvalue	Condition index	variance ratio			
				constant	x_i	m_i	$x_i \cdot m_i$
0	1	2.925	1	0	0	0.01	
	2	0.06	6.96	0.01	0.24	0.68	
	3	0.015	13.928	0.99	0.76	0.31	
1	1	3.89	1	0	0	0	0
	2	0.07	7.459	0.01	0.01	0.01	0.01
	3	0.039	9.988	0.02	0.01	0.01	0.02
	4	0.001	78.297	0.97	0.98	0.98	0.97

3.2 The Product of Standardized x and m

The results for the model suggested by Aiken and West (1991), the model 2, to mitigate multicollinearity are as follows.

$$\hat{Y}_i = \beta_0 + \beta_1 x_i + \beta_2 m_i + \beta_3 Z x_i \cdot Z m_i$$

$$= 2.286 + 0.361 x_i + 0.200 m_i - 0.119 Z x_i \cdot Z m_i$$

Additionally, the standardized equation is presented as follows:

$$\hat{Y}_i = 0.293 x_i + 0.238 m_i - 0.101 Z x_i \cdot Z m_i$$

After the estimation, we use both the VIF and CI to identify the multicollinearity. Table 3 reports the VIF and the tolerance for the model 2.

Table 3
Tests on the tolerance and VIF for model 2

model		Unstandardized		Standardized	T	significance	Collinearity Statistics	
		B	standard error	Beta			tolerance	VIF
2	(constant)	2.286	0.359		6.363	0.000		
	x_i	0.361	0.060	0.239	5.975	0.000	0.989	1.011
	m_i	0.200	0.041	0.238	4.861	0.000	0.991	1.009
	$Z x_i \cdot Z m_i$	-0.119	0.057	-0.101	-2.075	0.039	0.996	1.004

When the standardized interaction term is applied, VIF for the respective variables are 1.011, 1.009, and 1.004 for

the x , m , and $Zx_i \cdot mi$ respectively. All three VIF values are lower than 10, indicating the suggested approach of Aiken and West (1991) effectively resolves the multicollinearity. We further use the CI to diagnose the models in Table 4. The CI value for the interaction term escalates to 13.968, which is lower than 30. This indicates a significant improvement in mitigating multicollinearity.

Table 4
Tests on the CI for model 2

model	dimension	eigenvalue	Condition index	variance ratio			
				constant	x_i	mi	$Zx_i \cdot mi$
2	1	2.937	1.000	0.00	0.00	0.01	0.00
	2	0.988	1.724	0.00	0.00	0.00	0.99
	3	0.060	6.984	0.01	0.24	0.68	0.00
	4	0.015	13.968	0.99	0.76	0.31	0.00

3.3 Standardization of All Independent Variables

The results for the model 3 are reported as follows.

$$\begin{aligned}\hat{Y}_i &= \beta_0 + \beta_1 Zx_i + \beta_2 Zm_i + \beta_3 Zx_i \cdot Zm_i \\ &= 4.910 + 0.330Zx_i + 0.268Zm_i - 0.119Zx_i \cdot Zm_i\end{aligned}$$

Additionally, the standardized equation is presented as follows:

$$\hat{Y}_i = 0.293Zx_i + 0.238Zm_i - 0.101Zx_i \cdot Zm_i$$

After the estimation, we use both the VIF and CI to identify the multicollinearity. Table 5 reports the VIF and the tolerance for the model 3.

Table 5
Tests on the tolerance and VIF for model 3

model		Unstandardized		Standardized	T	significance	Collinearity Statistics	
		B	standard error	Beta			tolerance	VIF
3	(constant)	4.910	0.055		89.030	0.000		
	Zx_i	0.330	0.055	0.293	5.975	0.000	0.989	1.011
	Zm_i	0.268	0.055	0.238	4.861	0.000	0.991	1.009
	$Zx_i \cdot Zm_i$	-0.119	0.057	-0.101	-2.075	0.039	0.996	1.004

When the standardized interaction term is applied, VIF for the respective variables are 1.011, 1.009, and 1.004 for the Zx , Zm , and $Zx_i \cdot mi$ respectively. All three VIF values are lower than 10, indicating the tested approach

effectively resolves the multicollinearity. We further use the CI to diagnose the model 3 in Table 6, it is observed that the CI of the interaction term significantly decreases from the 78.297 in Table 2 or the 13.968 in Table 4 down to 1.130. The value is also lower than the threshold of 30. This reduction indicates a notable improvement in resolving the multicollinearity.

Table 6
 Tests on the CI for model 3

model	dimension	eigenvalue	Condition index	variance ratio			
				constant	Z_{xi}	Z_{mi}	$Z_{xi} \cdot Z_{mi}$
3	1	1.138	1.000	0.13	.25	.20	.29
	2	1.054	1.039	0.40	.18	.23	.14
	3	0.917	1.114	0.15	.30	.52	.12
	4	0.891	1.130	0.33	.28	.05	.46

3.4 Standardization of Both Independent and Dependent Variables

The results for the model that standardized the dependent variable (Y) and also the Z_x , Z_m , and $Z_x \cdot Z_m$.

$$\begin{aligned} \hat{Z}_{Yi} &= \beta_0 + \beta_1 Z_{xi} + \beta_2 Z_{mi} + \beta_3 Z_{xi} \cdot Z_{mi} \\ &= -0.009 + 0.293Z_{xi} + 0.238Z_{mi} - 0.105Z_{xi} \cdot Z_{mi} \end{aligned}$$

Additionally, the standardized equation is presented as follows:

$$\hat{Z}_{Yi} = 0.293Z_{xi} + 0.238Z_{mi} - 0.101Z_{xi} \cdot Z_{mi}$$

After the estimation, we use both the VIF and CI to identify the multicollinearity. Table 7 reports the VIF and the tolerance for the model 2.

Table 7
 Tests on the tolerance and VIF for model 4

model		Unstandardized		Standardized	T	significance	Collinearity Statistics	
		B	standard error	Beta			tolerance	VIF
4	(constant)	-.009	.049		-.192	.848		
	Z_{xi}	.293	.049	.293	5.975	.000	.989	1.011
	Z_{mi}	.238	.049	.238	4.861	.000	.991	1.009
	$Z_{xi} Z_{mi}$	-.105	.051	-.101	2.075	.039	.996	1.004

Upon standardizing both the independent variable and the dependent variable, it was observed that the VIF for the x , m , and interaction term are down to 1.011, 1.009, and 1.004. The VIF values for all three variables, similar to

scenarios where only the interaction term or all variables were standardized, they are well below 10. Again, this indicates the approach effectively resolved the multicollinearity. Yet, the CI for model 4 also down to 1.130, which fit better to the assumption of regression model.

Table 8
Tests on the CI for model 4

model	dimension	eigenvalue	Condition index	variance ratio			
				constant	Zx	Zm	$Zxi \cdot Zmi$
4	1	1.138	1.000	.13	.25	.20	.29
	2	1.054	1.039	.40	.18	.23	.14
	3	.917	1.114	.15	.30	.52	.12
	4	.891	1.130	.33	.28	.05	.46

3.5 Standardizing Both the Independent and Dependent Variables, with the Interaction Term

We make a bold attempt that standardizing all the variables in the regression model. The interaction item is standardized after the multiplication of two variables (the x and m). That is, the difference between our model 4 and 5 is the creation process of interaction term. The results for the model 5 are as follows.

$$\begin{aligned}\hat{Z}_{Yi} &= \beta_0 + \beta_1 Zx_i + \beta_2 Zm_i + \beta_3 Zx_i \cdot m_i \\ &= -0.00001 + 0.68x_i + 0.759m_i - 0.63Zx_i \cdot m_i\end{aligned}$$

Additionally, the standardized equation is presented as follows:

$$\hat{Z}_{Yi} = 0.68x_i + 0.759m_i - 0.63Zx_i \cdot m_i$$

After the estimation, we use both the VIF and CI to identify the multicollinearity. Table 9 reports the VIF and the tolerance for the model 5.

Table 9
Tests on the tolerance and VIF for model 5

model		Unstandardized		Standardized	T	significance	Collinearity Statistics	
		B	standard error	Beta			tolerance	VIF
5	(constant)	-3.63E-16	.049		.000	1.000		
	Zxi	.680	.195	.680	3.482	.001	.062	16.048
	Zmi	.759	.255	.759	2.982	.003	.037	27.290
	$Zxi \cdot mi$	-.630	.303	-.630	-2.075	.039	.026	38.772

In table 9, the VIF for the variables x , m , and interaction term are 16.048, 27.290, and 38.772. All three VIF values

have reverted to the levels observed in the original moderated multiple regression model, and all exceeding the threshold of 10. That indicates that this method has not effectively addressed the multicollinearity.

Table 10
Tests on the CI for model 5

model	dimension	eigenvalue	Condition index	variance ratio			
				constant	Z_x	Z_m	$Z_{xi\ mi}$
5	1	1.903	1.000	.00	.00	.01	.01
	2	1.084	1.325	.00	.04	.01	.00
	3	1.000	1.380	1.00	.00	.00	.00
	4	.012	12.390	.00	.96	.98	.99

Upon integrating the interaction of standardized independent variable X with the standardized moderating variable M, it was observed that the CI value for the product term significantly decreased from the initial 78.297 to 12.390, remaining below the threshold of 30, indicating a substantial mitigation of multicollinearity issues. Despite this, the VIF mirrored the original regression equation, suggesting persistent multicollinearity concerns. However, with the CI value dropping to 12.390, applying this method does not simultaneously reduce both multicollinearity indicators, highlighting its limitations compared to other approaches.

4. Conclusion

In summary, our five moderated multiple regression models reveal that Model 1, which is the original regression equation, displays significant multicollinearity as indicated by VIF and CI analyses. To address this issue, we applied four different standardization techniques in Models 2 through 5, leading to the following conclusions:

Initially, the original Moderated Multiple Regression model faced multicollinearity problems, complicating the accurate estimation of the independent variables' effects on the dependent variable. However, the techniques used in Models 2, 3, and 4 effectively resolved these multicollinearity issues. The standardized coefficients from these three models were identical, indicating that replacing the original non-standardized equation with the standardized methods in Models 2, 3, and 4 produced the same results. Conversely, Model 5 did not successfully address the multicollinearity problem.

Second, regarding the diagnosis of multicollinearity through two metrics, Variance Inflation Factor (VIF) and Condition Index (CI), the analysis of models two, three, and four reveals uniform VIF values for the independent variables across these models. However, the CI value for the independent variable in model two

diverges from that observed in models three and four. This discrepancy highlights the VIF method as a more consistent and stable approach for assessing multicollinearity.

Ultimately, it was found that to effectively address multicollinearity, it is sufficient to standardize the independent and moderating variables prior to creating a standardized interaction term through multiplication. Standardizing the product of the original independent and moderating variables after multiplication did not yield effective results. Additionally, there is no need to standardize the dependent variable Y , nor is it necessary to standardize both the independent variable X and the moderating variable m before incorporating them into the moderated multiple regression equation. Thus, the solution to multicollinearity issues lies in standardizing the interaction term formed by the multiplication of the independent and moderating variables.

Reference

1. Agresti, A. 2002. *Categorical data analysis* (2nd ed.). New York: John Wiley.
2. Aguinis, H. 1995. Statistical Power with Moderated Multiple Regression in Management Research. *Journal of Management*, 21(6): 1141–1158.
3. Aiken, L. S. & West, S. G. 1991. *Multiple Regression: Testing and Interpreting Interactions*. Newbury Park, CA: Sage Publications, Inc.
4. Allison, P. D. 1999. *Logistic regression using the SAS system: Theory and application*. Cary, NC: SAS Institute.
5. Andy, F. 2009. *Discoverig Statisticis using SPSS*. London: Sage Publications Inc.
6. Belsley, D. A., Kuh, E. & Welsch, R. E. 1980. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: Wiley.
7. Berry, W. D. & Feldman, S. 1985. *Multiple regressions in practice*. Newbury Park, CA: Sage Publications.
8. Cohen, J. & Cohen, P. 1983. *Applied Multiple Regression Correlation Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.
9. Draper, N. R. & Smith, H. 1998. *Applied Regression Analysis* (3rd ed.). NY: John Wiley and Sons, Inc.
10. Garcia, R. & Kandemir, D. 2006. Illustration of Moderating Effects in Multi-national studies? *International Marketing Review*, 23(4): 371–389.
11. Hosmer, D. W. & Lemeshow, S. 1999. *Applied Survival Analysis Regression Modeling of Time to Event Date*. New York: John Wiley & Sons.
12. Kleinbanum, D. G., Kupper, L. L., & Muller, K. E. 1998. *Applied Regression Analysis and other Multivariable Mehtods* (2th ed.). North Scituate, MA: Duxbury Press.
13. Lattin, J. M., Carrol, J. D. & Green, P. E. 2003. *Analyzing Multivariate Data*. Pacific Grove, CA: Thomson Brooks/Cole.
14. Lawal, B. 2003. *Categorical data analysis with SAS and SPSS applications*. LoMundon: Lawrence Erlbaum Associates.
15. Menard, S. 1995. *Applied logistic regression analysis*. Thousand Oaks, CA: SAGE Publications.
16. Myers, R. H. 1990. *Classical and Modern Regression with Applications*. Boston MA: Duxbury Press.

17. Tacq, J. 1997. *Multivariate Analysis Techniques in Social Science Research*. London: SAGE.
18. Zedeck, S. 1971. Problems with the Use of “Moderator” Variables. *Psychological Bulletin*, 76(4): 295–310.