

Second Order Regression with Two Predictor Variables Centered on Mean in an Ill Conditioned Model

Ijomah Maxwell Azubuike

Department of Maths/Statistics, University of Port Harcourt, Nigeria

Abstract It has been recognized that centering can reduce collinearity among explanatory variables in a linear regression models. However, efficiency of centering as a solution to multicollinearity highly depends on correlation structure among predictive variables. In this paper, simulation study was performed in a polynomial model to examine the effect of centering at various level of collinearity. The results empirically verify that centering first dramatically reduces the collinearity whereas under severe collinearity, centering provides only a small improvement over no centering at all. Therefore application of centering as a solution to multicollinearity problem should be discouraged under severe collinearity.

Keywords Centering, Regression model, Polynomial Regression, Ill conditioned model

1. Introduction

Regression analysis is a statistical method widely used in many fields such as Statistics, economics, technology, social sciences and finance. A linear regression model is constructed to describe the relationship between the dependent variable and one or several independent variables. All procedures used and conclusions drawn in a regression analysis depend on assumptions of a regression model. The most used model is the classic linear regression model and the most used method for estimating classic model parameters is the Ordinary Least Squares (OLS).

The linear regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \varepsilon \quad (1)$$

The matrix form can be expressed as

$$\underline{y} = \underline{X}\underline{\beta} + \underline{\varepsilon} \quad (2)$$

where

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad x = \begin{bmatrix} x_{11} & \dots & x_{1p} \\ x_{21} & \dots & x_{2p} \\ \vdots & & \vdots \\ x_{n1} & \dots & x_{np} \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix} \quad e = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Normal equation in matrix form. Given n $\underline{y} = \underline{X}\underline{\beta} + \underline{\varepsilon}$ for one predictor variable (say x_1)

$$y = \beta_0 + \beta_1 x_1 + \varepsilon \quad (3)$$

The sample estimate of $\underline{\hat{\beta}}$ can be written as

$$\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = \begin{bmatrix} n & \Sigma x_1 \\ \Sigma x_1 & \Sigma x_1^2 \end{bmatrix}^{-1} \begin{bmatrix} \Sigma Y \\ \Sigma X_1 Y \end{bmatrix} \quad (4)$$

This is expressed as $\hat{\beta} = (X'X)^{-1} X'Y$

Under the classic assumptions, the OLS method has some attractive statistical properties that have made it one of the most powerful and popular methods of regression analysis. However, OLS is not appropriate if the explanatory variables exhibit strong pair wise and/or simultaneous correlation (multicollinearity), causing the design matrix to become non-orthogonal or worse, ill-conditioned. Once the design matrix is ill-conditioned, the least squares estimates are seriously affected, e.g., instability of parameter estimates, reversal of expected signs of the coefficients, masking of the true behavior the linear model being explored, etc. Furthermore, it reveals that small change in the data may lead to large differences in regression coefficients, and causes a loss in power and makes interpretation more difficult since there is a lot of common variation in the variables (Vasu and Elmore 1975; Belsley 1976; Stewart 1987; Dohoo et al., 1996; Tu et al., 2005).

The problem of multicollinearity commonly exists among economic indicators that are influenced by similar policies that lead their simultaneous movement along similar directions.

* Corresponding author:

maxwell.ijomah@uniport.edu.ng (Ijomah Maxwell Azubuike)

Published online at <http://journal.sapub.org/statistics>

Copyright © 2019 The Author(s). Published by Scientific & Academic Publishing

This work is licensed under the Creative Commons Attribution International

License (CC BY). <http://creativecommons.org/licenses/by/4.0/>

Whether co-integration exists or not among the predictors, simultaneous drifting away in some directions especially among time series that exhibit non-stationary behavior is common.

There are many solutions proposed in the literature to address this problem. Among them are Dropping of variables (Carnes and Slade, 1988), the general shrinkage estimators (Hoerl and Kennard 1980; McDonald and Galarneau, 1975; George and Oman, 1996; McDonald, 1980), principal component regression (Butler and Denham, 2000), then centering (Aiken & West 1991, Cronbach 1987, Irwin & McClelland 2001). However, there is a growing debate on whether or not to involve centering as a solution for a collinear regression. It has been argued that the source of any discrepancies among statistical findings based on regression analyses in absence of centering is not mysterious; they can always be explained and resolved. Various researchers including Aiken and West (1991); Cronbach (1987), and Jaccard, Wan & Turrisi (1990); Irwin & McClelland, (2001); and Smith & Sasaki, (1979) recommend mean centering the variables x_1 and x_2 as an approach to alleviating collinearity related concerns. Aiken, West and others further recommend that one centre only in the presence of interactions. Centering aids interpretation and reduces the potential for multicollinearity (Aiken and West 1991). It is therefore a strategy to prevent errors in statistical inference. 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). Centering is recommended in order to eliminate collinearities which are due to the origins of the predictor variables and it can often provide computational benefits when small storage or low precision prevail (Marquardt and Snee, 1975).

However, in contrast to Cronbach's injunction, other authors (Glantz & Slinker, 2001; Kromrey & Foster-Johnson, 1998; Belsley, 1984; Echambadi & Hess 2007) take the stand that centering doesn't usually change the statistical results, is necessary only in certain circumstances, and can thus easily be. Few authors like Hocking (1984), Snee (1983), Belsley (1984b), have attempted to address this problem but did not extend to the polynomial regression especially the two predictor variables. The problem of centering is therefore an issue which is still not completely resolved. In this article, we clarify the issues and reconcile the discrepancy. Therefore the aim of this paper is to compare the statistical estimates of a centered mode in a second order regression model at various degrees of collinearity for two predictor variables (X_1 & X_2) in a linear component, quadratic component and interaction (cross product).

The rest of the paper is mapped out as follows: Section 2 presents related literature and theoretical perspective of polynomial regression. Section 3 is the materials and

methods of the paper, following interpretation of the empirical result, section 4 is the conclusion of the paper.

2. Polynomial Regression

In Statistics polynomial regression is a form of linear regression in which the relationship between the independent variable x and the dependent variable y is modeled as an n^{th} degree polynomial. Polynomial regression fits a nonlinear relationship between the value of x and the corresponding conditional mean of y , denoted $E(y/x)$. In general, we can model the expected value of y as an n^{th} degree polynomial. The polynomial regression model may contain one, two, or more than two predictor variables. Each predictor variable may be present in various powers.

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \dots + \beta_n x^n + \varepsilon \quad (5)$$

Further, each predictor variable may be present in various powers. We begin by considering a polynomial regression model with one predictor variable raised to the first and second power

$$y = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon \quad (6)$$

where $x_i = X_i - \bar{X}$

This polynomial model is called a second-order with one predictor variable because the single predictor variable is expressed in the model to the first and second powers. The predictor variable is centered – that is, expressed as a deviation around its mean and that the i^{th} centered observation is denoted by.

The two predictor variables of second order is given by

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{11} x_{i1}^2 + \beta_{22} x_{i2}^2 + \beta_{12} x_{i1} x_{i2} + \varepsilon_i \quad (7)$$

where $\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$ represents the linear component, $\beta_{11} x_{i1}^2 + \beta_{22} x_{i2}^2$ = quadratic component and $\beta_{12} x_{i1} x_{i2}$ which is the cross product or interaction component.

The above is a second-order model with two predictor variables. The response function is:

$$E\{Y\} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 \quad (8)$$

Centering is defined as subtracting the mean (a constant) from each score, X , yielding a centered score. It is therefore an important step when testing interaction effects in multiple regression to obtain a meaningful interpretation of results. Centering the variables places the intercept at the means of all the variables. A regression equation with an intercept is often misunderstood in the context of multicollinearity. The intercept is an estimate of the response at the origin where all independent variables are zero, thus inclusion of the intercept in the study of collinearity is not of much interest. When variables have been centered, the intercept has no effect on the collinearity of the other variables (Belsley, Kuh, and Welsch, 1980). Centering is also consistent with the computation of the variance inflation factor (VIF) and

therefore it is suggested that VIF be computed only after first centering variables (Freund, Littell, and Creighton, 2003).

Ostertagová, (2012) describe how polynomial regression model is useful when there is reason to believe that the relationship between two variables is curvilinear, and illustrated using data from a drilling-hole in the engineering field. Michael et al (2005) describes the reason for centering predictor variable in the polynomial regression model is that X and X^2 will often be highly correlated, and recommended centering as a means of reducing the multicollinearity. They observed that after the regression model has been fitted, the fitted values and residuals for the regression function in terms of X are exactly the same as the regression function in terms of the centered values of x . Also they stated that the estimated standard deviations of the regression coefficients in terms of the centered variables x do not apply to the regression coefficient in terms of the original variables of X . A number of scholars have considered issues related to mean centering with regard to the inclusion of product terms in a multiple regression model to test for moderators (Iacobucci et al., 2016). They noted that mean centering variables does not change the nature of the relationships between any variable in the set that does not include the product term.

Aitken and West (1991) encourages centering only in the presence of interactions while David and Richard (2011) in their work describes the term centering on mean as subtracting the mean value from an independent variable, in their work they noted that the interaction coefficient does not change from the non centered model to the centered model. Kramer and Bassay (2005), in a study on “centering in regression analysis: a strategy to prevent errors in statistical inference”, noted that non centered data in regression

analysis, often leads to inconsistent and misleading results. Also that centering does not change the predicted values when predictor values are perfectly correlated except in the case of multicollinearity which are usually associated with models with interaction or a higher order term such as X^2 . McClelland et. al. (2016) in their critical research verified the irrelevance of multicollinearity in the model with moderator variables. While Kaur (2017) observed that centering only helps multicollinearity disappear and doesn't quite improve the regression model as such.

3. Materials and Methods

To show the effect of mean centering on an ill-conditioned regression model, we ran a Monte Carlo simulation study using SAS 9.0 version. We begin by multiplying each of the selected variable with 0.5 in order to make the variables uniform.

```
U= ranuni (start)
X1=U+rannor (start)*0.5
X2=U+rannor (start)*0.5
Y=1+X1+X2+rannor (start)
```

That is, we set ρX_1 , $Y=\rho X_2$, $Y=0.5$ to represent modest sized effects of two predictors on the dependent variable, and varied the extent of multicollinearity, ρX_1 , X_2 from 0.1 to 0.9. To achieve the objective of varying collinearity, we perturb the random error μ by multiplying it with values (i.e. 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 15, 30). This generates different values of the correlation coefficient between the two explanatory variables as shown in table 1.

Table 1. Table of values of correlation

u	0	1	2	3	4	5	6	7	8	9	15	30
r	0.155	0.300	0.493	0.699	0.810	0.871	0.910	0.931	0.946	0.957	0.984	0.995

Three different scenarios are considered from the above table 1, which have different correlation structures (i.e. low, moderate and severe) among the variables. Starting with the simplest case which assumes that two variables are uncorrelated or have minimal correlation with each other in a multivariable model. In this bivariate collinearity scenario, we investigate the effect of bias in estimates of related regressor coefficients due to the variation in the degree of collinearity between the centered and uncentered models. For the next group, we extend the first scenario by allowing some degree of collinearity among the variables (moderate collinearity). And in the last group, which is the case of severe collinearity where the degree of correlation between the variables is very strong. In each case, three components were captured and they include: (1) linear component, (2) quadratic form of a variable and (3) interaction of two variables. To evaluate the sample size variation on precision of estimation, for each level of $\rho X_1 X_2$, we generated a

random sample of size $N = 50, 100, 200, 500$ and 1000 size. We computed a product score, $X_1 X_2$. We ran a regression using the main effects and interaction, X_1 , X_2 , and $X_1 X_2$ to predict Y . We obtained the β estimates, their standard errors, and the p-values representing their significance tests. We also took the same generated variables and mean centered X_1 and X_2 . We computed the new product score, ran another regression, and obtained the new β s, standard errors, and p-values.

Case 1: Minimal Collinearity among the explanatory variables

This scenario assumes little or no correlation between the explanatory variables. Here we noticed that for linear component, the regression coefficients were higher in the centered model than the uncentered model. This is an indication that greater multicollinearity dampens the estimate of β . Since the standard error of the coefficient (SE

Coef) indicates the precision of the coefficient estimates, the table shows that the centered model has more reliable estimates than the uncentered for linear component. By comparison, mean centering reduces standard errors and thus benefits p-values and the likelihood of finding β_1 or β_2 significant. This is unlike the case of quadratic and interaction term, both the mean-centered and the uncentered models for quadratic and interaction term provided an identical fit to the data. As expected, the coefficients of the interaction, the standard errors, and the t-statistics obtained from both the models are identical. A point that may confuse some researchers in this regard is that t-statistics for individual regressors may change when data are mean-centered. This does not occur for the quadratic and interaction term. As noted by Aiken and West (1991) and shown here, the coefficient and the standard error for the interaction term, and hence the significance of this term, will be identical with or without mean-centering. However, t-statistics may change for the linear terms as a result of shifting the interpretation of the effect. In a regression without mean-centering, the coefficients represent simple effects of the exogenous variables, i.e., the effects of each variable when the other variables are at zero. When data are mean-centered, the coefficients represent main effects of these variables, i.e., the effects of each variable when the other variables are at their mean values.

Case 2: Moderate Collinearity among the variables

In this case, we considered moderate collinearity between X_1 and X_2 variables. As shown in the table 3, here again the same scenario repeated itself for linear component in terms of regression coefficient and standard error but for the quadratic and interaction components were quite different. An examination of the variance inflation factor under moderate collinearity from Table 3 reveals that the Variance inflation factor (VIF) of the centered model indicated absence of collinearity ($VIF < 10$) in all the three components considered while for uncentered model, collinearity was present in all the components (i.e $VIF > 10$). This may be an indication that mean centering helps to ameliorate collinearity problems in quadratic and interaction components judging by $VIF < 10$.

Case 3: Severe Collinearity among variables

This is a scenario where there is a very strong correlation between X_1 and X_2 variables. Table 4 indicates again that the standard error of centered model were lower than that of uncentered model for linear components but for quadratic and interaction components, the standard errors in both

models were relatively the same. The VIF shows absence of collinearity in linear component judging with the rule of thumb ($VIF < 10$) but there was presence of collinearity quadratic and interaction components in the centered model. The uncentered models indicated that there is presence of collinearity in all the three components. It is interesting to note that the VIF of uncentered model is more than ten times VIF in centered model especially for quadratic and interaction components. Our findings here show that centering model helps to reduce collinearity problem. It should be pointed out that in using centered care must be taken on the degree of collinearity.

The response of the centered and uncentered models was graphically represented by 3D response surface generated by the three cases (i.e. minimal collinearity, moderate collinearity and severe collinearity). Figures 1 and 2 depict the nature of how mean centering enhances a regression analysis. Figure 1 described the response of centered and uncentered models to multicollinearity using variance inflation factor (VIF). It may be also observed that with low collinearity, the centered model experienced VIF between 0 and 1.16 while the uncentered model responded to VIF between 0 and 4.20. However, with moderate collinearity VIF slightly increased further but was still less than 3 ($VIF < 3$) an indication of absence of collinearity. For the uncentered model, it was a different scenario entirely. The increase in VIF was very pronounced. Little wonder while the bumps on the graph with the uncentered model. When the collinearity was severe, the VIF of both the centered and the uncentered models were on the increase. As can be seen on the graph, a slight bump was experienced by the centered model unlike the case of the uncentered model which showed number of bumps.

The effect of standard error on both centered and uncentered model was also represented in figure 2. With minimal collinearity, both models showed a negligible increase in standard error of estimates. For moderate collinearity, again, for centered model, the standard error becomes inflated with severe collinearity while for uncentered model; the standard error is inflated in all the three cases (minimal collinearity, moderate collinearity and severe collinearity). Surprisingly the direction of the peak among the standard errors varied in both models. This may be as a result of the lower standard errors of linear component in the centered model than that of uncentered model. While the standard error of the quadratic and interaction terms of the uncentered model were lower than that of the centered model.

Table 2. Centered and Uncentered models with low collinearity

Centered model					Uncentered model			
n=50	Estimate	Std Error	t value	VIF	Estimate	Std Error	t value	VIF
Linear	$\beta_0 = 2.03684$	0.22371	9.10	0	$\beta_0 = 0.95716$	0.25799	3.71	0
	$\beta_1 = 0.93032$	0.25406	3.66	1.04721	$\beta_1 = 0.62470$	0.47544	1.31	0.36724
	$\beta_2 = 1.42923$	0.27841	4.06	1.07047	$\beta_2 = 1.38789$	0.65277	2.13	5.88479
Quadratic	$\beta_{11} = 0.23607$	0.33826	0.70	1.11547	$\beta_{11} = 0.23607$	0.33826	0.70	3.04310
	$\beta_{22} = -0.29310$	0.48212	-0.61	1.04756	$\beta_{22} = -0.29310$	0.48212	-0.61	4.88221
Cross product	$\beta_{12} = 0.12777$	0.58736	0.22	1.15145	$\beta_{12} = 0.12777$	0.58736	0.22	4.07360
n=100	Estimate	Std Error	t value	VIF	Estimate	Std Error	t value	VIF
Linear	$\beta_0 = 2.24310$	0.13560	16.54	0	$\beta_0 = 0.94584$	0.19106	4.95	0
	$\beta_1 = 0.98783$	0.17136	5.76	1.05694	$\beta_1 = 0.85891$	0.34441	2.49	4.26944
	$\beta_2 = 1.11496$	0.17785	6.27	1.06610	$\beta_2 = 1.08887$	0.32516	4.89	3.56355
Quadratic	$\beta_{11} = 0.14448$	0.19726	0.73	1.02179	$\beta_{11} = 0.14448$	0.19726	0.73	2.66924
	$\beta_{22} = -0.38538$	0.19644	-1.96	1.03480	$\beta_{22} = -0.38538$	0.19644	-1.96	2.22477
Cross product	$\beta_{12} = -0.07130$	0.32870	-0.22	1.02069	$\beta_{12} = -0.07130$	0.32870	-0.22	4.20287
n=200	Estimate	Std Error	t value	VIF	Estimate	Std Error	t value	VIF
Linear	$\beta_0 = 1.95920$	0.10499	18.66	0	$\beta_0 = 1.00999$	0.12463	8.10	0
	$\beta_1 = 0.99555$	0.12795	7.78	1.02752	$\beta_1 = 0.58166$	0.23752	2.45	3.54077
	$\beta_2 = 1.03691$	0.13367	7.76	1.03433	$\beta_2 = 1.02887$	0.16240	5.36	3.04566
Quadratic	$\beta_{11} = 0.34783$	0.16240	2.14	1.02251	$\beta_{11} = 0.34783$	0.16240	2.14	2.77758
	$\beta_{22} = -0.26597$	0.17125	-1.55	1.02420	$\beta_{22} = -0.26597$	0.17125	-1.55	2.41685
Cross product	$\beta_{12} = 0.15252$	0.24894	0.61	1.02186	$\beta_{12} = 0.15252$	0.24894	0.61	3.15796
n=500	Estimate	Std Error	t value	VIF	Estimate	Std Error	t value	VIF
Linear	$\beta_0 = 1.95920$	0.06484	30.66	0	$\beta_0 = 1.07993$	0.07621	14.17	0
	$\beta_1 = 0.99555$	0.08021	12.59	1.05121	$\beta_1 = 0.76635$	0.14326	5.35	3.35393
	$\beta_2 = 1.03691$	0.08374	11.75	1.05404	$\beta_2 = 0.94197$	0.13879	6.79	2.89505
Quadratic	$\beta_{11} = 0.34783$	0.10564	1.05	1.05307	$\beta_{11} = 0.11043$	0.10564	1.05	2.96494
	$\beta_{22} = -0.26597$	0.10366	-0.97	1.05330	$\beta_{22} = -1.0098$	0.10366	-0.97	2.41468
Cross product	$\beta_{12} = 0.15252$	0.15739	1.75	1.10005	$\beta_{12} = 0.27595$	0.15739	1.75	3.57830
n=1000	Estimate	Std Error	t value	VIF	Estimate	Std Error	t value	VIF
Linear	$\beta_0 = 1.99873$	0.04620	43.26	0	$\beta_0 = 1.04567$	0.05322	19.65	0
	$\beta_1 = 1.02066$	0.05804	17.59	1.06217	$\beta_1 = 0.95072$	0.09986	9.52	3.14497
	$\beta_2 = 0.99507$	0.05855	17.00	1.06162	$\beta_2 = 0.91453$	0.09702	9.43	2.91495
Quadratic	$\beta_{11} = 0.02847$	0.07519	0.38	1.08828	$\beta_{11} = 0.02847$	0.07519	0.38	2.99585
	$\beta_{22} = 0.03880$	0.07352	0.53	1.07787	$\beta_{22} = 0.03880$	0.07352	0.53	2.55888
Cross product	$\beta_{12} = 0.08624$	0.11306	0.76	1.15957	$\beta_{12} = 0.08624$	0.11306	0.76	3.59693

Table 3. Centered and Uncentered models with moderate collinearity

Centered model					Uncentered model			
n=50	Estimate	Std Error	t value	VIF	Estimate	Std Error	t value	VIF
Linear	$\beta_0 = 3.02740$	0.24181	12.52	0	$\beta_0 = 0.95064$	0.43559	2.18	0
	$\beta_1 = 0.91967$	0.22091	4.16	1.36037	$\beta_1 = 0.70736$	0.51648	1.37	12.61213
	$\beta_2 = 1.4413$	0.25210	4.42	1.37017	$\beta_2 = 1.32802$	0.68132	1.95	17.22152
Quadratic	$\beta_{11} = 0.07698$	0.27672	0.28	1.73703	$\beta_{11} = 0.01656$	0.25093	0.07	31.52905
	$\beta_{22} = -0.21470$	0.40097	-0.54	1.33077	$\beta_{22} = -0.12518$	0.33502	-0.37	40.18248
Cross product	$\beta_{12} = 0.10815$	0.47292	0.23	2.13932	$\beta_{12} = 0.10647$	0.46454	0.23	77.57886
n=100	Estimate	Std Error	t value	VIF	Estimate	Std Error	t value	VIF
Linear	$\beta_0 = 3.28414$	0.14407	22.79	0	$\beta_0 = 0.86607$	0.32465	2.67	0
	$\beta_1 = 0.99057$	0.15097	6.46	1.38941	$\beta_1 = 0.58341$	0.31701	1.84	10.25779
	$\beta_2 = 1.71019$	0.15871	7.00	1.38730	$\beta_2 = 1.62986$	0.34220	4.76	10.84868
Quadratic	$\beta_{11} = 0.09226$	0.14366	0.64	1.22902	$\beta_{11} = 0.04677$	0.12242	0.38	17.24170
	$\beta_{22} = -0.30591$	0.16488	-1.86	1.25681	$\beta_{22} = -0.24324$	0.14796	-1.64	21.24170
Cross product	$\beta_{12} = 0.07100$	0.25052	0.28	1.44087	$\beta_{12} = 0.15178$	0.22595	0.67	48.86517
n=200	Estimate	Std Error	t value	VIF	Estimate	Std Error	t value	VIF
Linear	$\beta_0 = 2.88786$	0.10918	26.45	0	$\beta_0 = 1.03546$	0.18964	5.46	0
	$\beta_1 = 1.00097$	0.11087	9.03	1.35228	$\beta_1 = 0.40536$	0.22981	1.76	9.93943
	$\beta_2 = 1.56253$	0.11555	9.30	1.339067	$\beta_2 = 1.44310$	0.23890	6.04	9.91256
Quadratic	$\beta_{11} = 0.22230$	0.11870	1.87	1.43146	$\beta_{11} = 0.13143$	0.10282	1.28	20.78607
	$\beta_{22} = -0.23806$	0.13688	-1.74	1.47109	$\beta_{22} = -0.20649$	0.12113	-1.70	24.78607
Cross product	$\beta_{12} = 0.16100$	0.20045	0.80	1.93807	$\beta_{12} = 0.15001$	0.19003	0.79	54.97491
n=500	Estimate	Std Error	t value	VIF	Estimate	Std Error	t value	VIF
Linear	$\beta_0 = 2.96722$	0.06709	44.22	0	$\beta_0 = 1.07724$	0.12219	8.82	0
	$\beta_1 = 1.01744$	0.07161	14.21	1.42694	$\beta_1 = 0.82367$	0.14544	5.66	9.99823
	$\beta_2 = 0.98964$	0.07411	13.35	1.42552	$\beta_2 = 1.11827$	0.14622	7.65	9.95839
Quadratic	$\beta_{11} = 0.03174$	0.08167	0.39	1.62306	$\beta_{11} = -0.00523$	0.07214	-0.07	26.07775
	$\beta_{22} = -0.12338$	0.08285	-1.49	1.61557	$\beta_{22} = -0.11856$	0.07402	-1.60	24.68597
Cross product	$\beta_{12} = 0.17838$	0.13456	1.33	2.35733	$\beta_{12} = 0.14710$	0.12820	1.15	67.26752
n=1000	Estimate	Std Error	t value	VIF	Estimate	Std Error	t value	VIF
Linear	$\beta_0 = 2.99026$	0.04785	62.49	0	$\beta_0 = 1.03242$	0.08335	12.39	0
	$\beta_1 = 1.02490$	0.05193	19.74	1.46428	$\beta_1 = 0.97717$	0.10439	9.36	10.15819
	$\beta_2 = 0.99828$	0.05195	19.22	1.46250	$\beta_2 = 1.00683$	0.10244	9.83	9.81776
Quadratic	$\beta_{11} = 0.01590$	0.05968	0.27	1.75125	$\beta_{11} = 0.00973$	0.05333	0.18	27.86125
	$\beta_{22} = -0.0001$	0.05846	-0.00001	1.72015	$\beta_{22} = -0.00972$	0.05246	-0.19	25.97650
Cross product	$\beta_{12} = 0.03120$	0.09872	0.32	2.61393	$\beta_{12} = 0.01243$	0.09460	0.13	72.90034

Table 4. Centered and Uncentered models with severe collinearity

Centered model					Uncentered model			
n=50	Estimate	Std Error	t value	VIF	Estimate	Std Error	t value	VIF
Linear	$\beta_0 = 6.86655$	0.24923	27.55	0	$\beta_0 = 0.99421$	0.45870	2.17	0
	$\beta_1 = 0.91567$	0.20512	4.46	6.40087	$\beta_1 = 0.82185$	0.51451	1.60	40.27258
	$\beta_2 = 1.09420$	0.22034	4.97	6.32756	$\beta_2 = 1.15881$	0.56612	2.05	41.77011
Quadratic	$\beta_{11} = -0.02002$	0.23421	-0.09	24.80891	$\beta_{11} = -0.02002$	0.23421	-0.09	341.21672
	$\beta_{22} = -0.04577$	0.27588	-0.17	26.52366	$\beta_{22} = -0.04577$	0.27588	-0.17	381.21672
Cross product	$\beta_{12} = 0.07099$	0.47598	0.15	83.22079	$\beta_{12} = 0.07099$	0.47598	0.15	1213.7180
n=100	Estimate	Std Error	t value	VIF	Estimate	Std Error	t value	VIF
Linear	$\beta_0 = 7.35172$	0.15650	46.98	0	$\beta_0 = 0.97870$	0.34353	2.85	0
	$\beta_1 = 0.95248$	0.13685	6.96	6.06082	$\beta_1 = 0.48763$	0.28618	1.70	26.50252
	$\beta_2 = 1.08796$	0.14113	7.71	6.04702	$\beta_2 = 1.52110$	0.30591	4.97	28.41229
Quadratic	$\beta_{11} = -0.02480$	0.10801	-0.23	11.38516	$\beta_{11} = -0.02480$	0.10801	-0.23	159.61277
	$\beta_{22} = -0.16872$	0.12938	-1.30	13.72916	$\beta_{22} = -0.16872$	0.12938	-1.30	213.00036
Cross product	$\beta_{12} = 0.19802$	0.22014	0.90	36.38865	$\beta_{12} = 0.19802$	0.22014	0.90	619.21382
n=200	Estimate	Std Error	t value	VIF	Estimate	Std Error	t value	VIF
Linear	$\beta_0 = 6.65116$	0.11851	56.12	0	$\beta_0 = 1.00565$	0.21214	4.74	0
	$\beta_1 = 0.98066$	0.09928	9.88	6.00869	$\beta_1 = 0.43594$	0.43594	2.07	27.05757
	$\beta_2 = 1.05697$	0.10122	10.44	5.95121	$\beta_2 = 1.51300$	1.51300	6.86	28.24253
Quadratic	$\beta_{11} = 0.02360$	0.09325	0.25	17.58821	$\beta_{11} = 0.02360$	0.02360	0.25	209.76580
	$\beta_{22} = -0.15249$	0.10623	-1.44	19.26952	$\beta_{22} = -0.15249$	0.15249	-1.44	251.42894
Cross product	$\beta_{12} = 0.14502$	0.18857	0.77	56.74538	$\beta_{12} = 0.14502$	0.14302	0.77	783.32981
n=500	Estimate	Std Error	t value	VIF	Estimate	Std Error	t value	VIF
Linear	$\beta_0 = 6.85880$	0.07304	93.91	0	$\beta_0 = 1.01176$	0.13724	7.37	0
	$\beta_1 = 1.01996$	0.06481	15.74	6.44311	$\beta_1 = 0.86065$	0.13298	6.47	27.12422
	$\beta_2 = 0.99165$	0.06567	15.10	6.43738	$\beta_2 = 1.17144$	0.13565	8.64	27.46456
Quadratic	$\beta_{11} = -0.04023$	0.06485	-0.62	20.21784	$\beta_{11} = -0.04023$	0.06485	-0.62	257.39208
	$\beta_{22} = -0.09992$	0.06643	-1.50	19.62362	$\beta_{22} = -0.09992$	0.06643	-1.50	255.95244
Cross product	$\beta_{12} = 0.13638$	0.12542	1.09	62.04372	$\beta_{12} = 0.13638$	0.12542	1.09	897.22934
n=1000	Estimate	Std Error	t value	VIF	Estimate	Std Error	t value	VIF
Linear	$\beta_0 = 6.94492$	0.05175	134.19	0	$\beta_0 = 0.99396$	0.09451	10.52	0
	$\beta_1 = 1.02081$	0.04687	21.78	6.74052	$\beta_1 = 0.98382$	0.09684	10.16	28.77247
	$\beta_2 = 0.99291$	0.04673	21.25	6.73819	$\beta_2 = 1.04292$	0.09557	10.91	28.18179
Quadratic	$\beta_{11} = 0.00425$	0.04824	0.09	22.26323	$\beta_{11} = 0.00425$	0.04824	0.09	282.32426
	$\beta_{22} = -0.01055$	0.04781	-0.22	22.21799	$\beta_{22} = -0.01955$	0.04781	-0.22	274.29796
Cross product	$\beta_{12} = 0.00403$	0.09256	0.04	69.81976	$\beta_{12} = 0.00403$	0.09256	0.04	985.30779

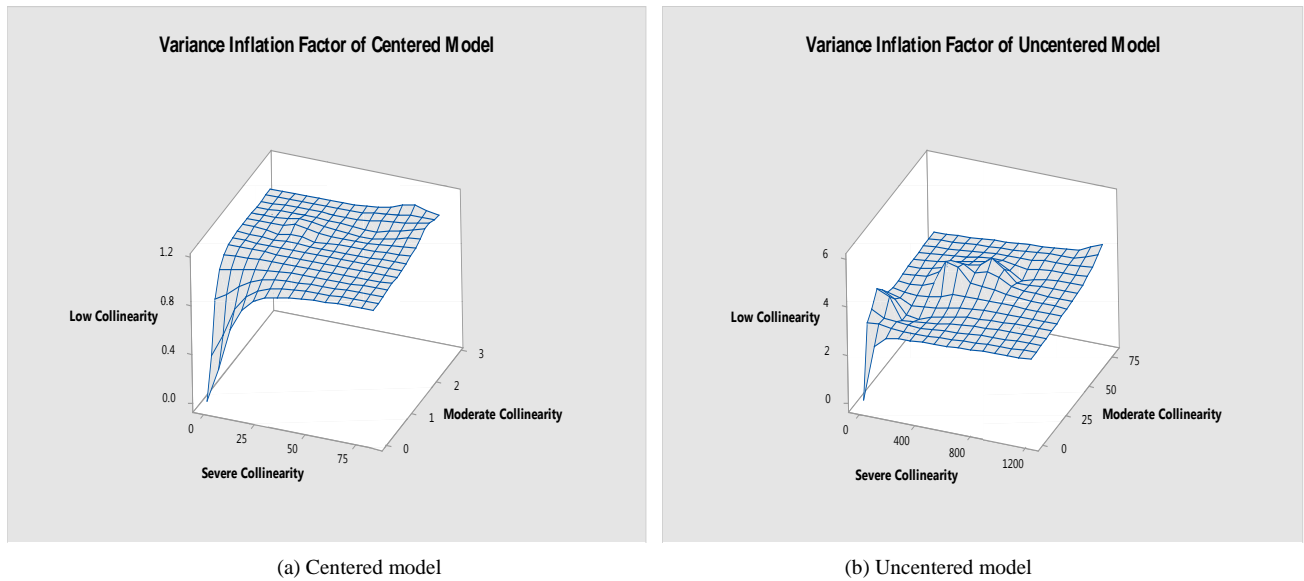


Figure 1. Three-dimensional Relationship of linear, quadratic and interaction response to VIF in low-collinear, moderately collinear and severely collinear variables between centered and uncentered models

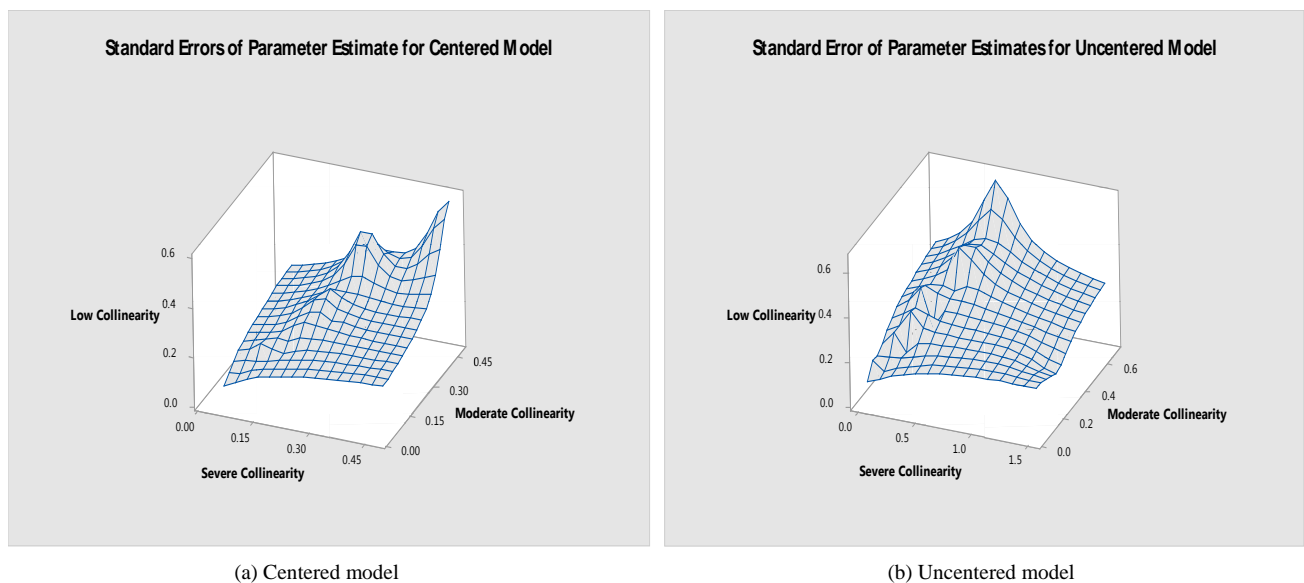


Figure 2. Three-dimensional Relationship of linear, quadratic and interaction response to standard error in low-collinear, moderately collinear and severely collinear variables between centered and uncentered models

4. Conclusions

In this analysis, we focused on how three components (linear, quadratic and interaction) behave in centered and uncentered models to an ill conditioned regression using different collinearity structure (low, moderate and severe). A simulation study considering regression model with linear, quadratic and interaction components for centered and uncentered models was established. The result showed that the linear effect is significant in both the uncentered and mean-centered models for all the three considered collinearity structures, whereas quadratic and interaction effect were insignificant. Also, when there is a meaningful

interaction among the variables, the linear effect will not equal the quadratic and interaction effect. It is worth to note that even the improvement offered by mean centering has its limits. When correlations is very strong, results begin to be affected even if the variables have been mean centered. The author believes that centering alleviates multicollinearity problem but it necessary to consider the degree of collinearity among the explanatory variables before using mean centering. For severe collinearity, use of centering should be discouraged especially for quadratic and interaction components.

REFERENCES

- [1] Aiken, L. S., & West, S. G. (1991). Multiple regression: Testing and interpreting interactions. Newbury Park: Sage.
- [2] Belsley DA. Multicollinearity: (1976). Diagnosing its presence and assessing the potential damage it causes least square estimation. NBER Working Paper, No W0154.
- [3] Belsley. D.A., E. Knu, and R.-. Welsh,. (1980). Regression Diagnostics: Identifying Influential Obserwations and Sources of Collinearity, Wiley, NY.
- [4] Butler, N. and Denham, M., (2000). The Peculiar Shrinkage Properties of Partial Least Squares Regression, *Journal of the Royal Statistical Society Ser. B* 62(3): pp.585-593.
- [5] Carnes BA, Slade NA. (1988). The Use of Regression for Detecting Competition with Multicollinear Data. *Ecology*, 69 (4): pp.1266–1274.
- [6] Cronbach, L. J. (1987). Statistical tests for moderator variables: Flaws in analyses recently proposed. *Psychological Bulletin*, 102, pp.414-41.
- [7] Dohoo I.R., Ducrot C, Fourichon C., (1996). An overview of techniques for dealing with large numbers of independent variables in epidemiologic studies. *Preventive Veterinary Medicine*. 29:pp.221–239.
- [8] Echambadi, R., & Hess, J. D. (2007). Mean centering does not alleviate collinearity problems in moderated multiple regression models. *Marketing Science*, 26(3), 438–445.
- [9] George, E. and Oman, S., (1996). Multiple-Shrinkage Principal Component Regression, *The Statistician* 45(1): pp.111-124.
- [10] Glantz, S.A., Slinker, B.K., (2001). *Primer of Applied Regression and Analysis of Variance*. New York: McGraw-Hill.
- [11] Harleen Kaur., (2017). Efficacy of Centering Techniques for Creating Interaction Terms in Multiple Regression for Modeling Brand Extension Evaluation. *International Journal of Research*, 4 (7), pp. 1422 -1436.
- [12] Hoerl, A. E. and Kennard, R. W. (1970). Ridge Regression: Application to non-orthogonal problems. *Technometrics*, 12, pp. 69-82.
- [13] Iacobucci, D., Schneider, M.J., Popovich, D.L. and Bakamitsos, G.A., (2017). Mean centering, multicollinearity, and moderators in multiple regression: The reconciliation redux. *Behavior research methods*, 49(1), pp.403-404.
- [14] Iacobucci, D., Schneider, M.J., Popovich, D.L. and Bakamitsos, G.A., (2016). Mean centering helps alleviate “micro” but not “macro” multicollinearity. *Behavior research methods*, 48 (4), pp.1308-1317.”
- [15] Irwin, J. R., & McClelland, G. H. (2001). Misleading heuristics and moderated multiple regression models. *Journal of Marketing Research*. 38(February), 100–109.
- [16] Jaccard, J., Wan, C. K., & Turrissi, R. (1990). The detection and interpretation of interaction effects between continuous variables in multiple regression. *Multivariate Behavioral Research*, 25(4), pp.467–478.
- [17] Kromrey, J. D., & Foster-Johnson, L. (1998). Mean centering in moderated multiple regression: Much ado about nothing. *Educational and Psychological Measurement*, 58(1), pp.42–67.
- [18] kraemer, C. H. and Blasey, C.M., (2005), “Centring in regression analyses: a strategy to prevent errors in statistical inference”, *International Journal of Methods Psychiatric Research*, 13(3).
- [19] Marquardt, D. W. (1980). You should standardize the predictor variables in your regression models. *Journal of the American Statistical Association*, 75 (369), 87–91.
- [20] Marquardt, D. W. and Snee, R. D. (1975). Ridge regression in practice. *Amer. Statist.*, 29, 3-19.
- [21] McClelland GH, Irwin JR, Disatnik D, Sivan L (2016). Multicollinearity is a red herring in the search for moderator variables: A guide to interpreting moderated multiple regression models and a critique of Iacobucci, Schneider, Popovich, and Bakamitsos. *Behav Res Methods*. Aug 16. [Epub ahead of print] PubMed PMID: 27531361.
- [22] Mc Donald, G and Galarneau, D., (1975). A Monte Carlo Evaluation of Some Ridge-Type Estimators, *Journal of the American Statistical Association* 70(350): 407-416.
- [23] Mc Donald, G., (1980). Some Algebraic Properties of Ridge Coefficient, *Journal of the Royal Statistical Society Ser. B* 42(1): 31-34.
- [24] Micheal, H.K., Christopher, J. N., John, N., and William L. (2005). *Applied Linear Statistical Models* (5thed.). McGraw Hill. pp 294-331.
- [25] Ostertagova, E., (2012). Modelling Using Polynomial Regression, *Procedia Engineering* 48:500- 506.
- [26] Smith, Kent W., and M. S. Sasaki. (1979). Decreasing multicollinearity: A method for models with multiplicative functions. *Sociological Methods and Research*, 8 (August 1979): 35-56.
- [27] Stewart GW. Collinearity and Least Square Regression. *Statistical Science*. (1987), 2(1): 68–94.
- [28] Shieh, G. (2009). Detecting interaction effects in moderated multiple regression with continuous variables: Power and sample size considerations. *Organizational Research Methods*, 12(3), 510–528.
- [29] Shieh, G. (2010). On the misperception of multicollinearity in detection of moderating effects: Multicollinearity is not always detrimental. *Multivariate Behavioral Research*, 45(3), 483–507.
- [30] Shieh, G. (2011). Clarifying the role of mean centring in multicollinearity of interaction effects. *British Journal of Mathematical and Statistical Psychology*, 64, 462–477.
- [31] Smith, K. W., & Sasaki, M. S. (1979). Decreasing multicollinearity: A method for models with multiplicative functions. *Sociological Methods & Research*, 8(1), 35–56.
- [32] Stone, E. F., & Hollenbeck, J. R. (1984). Some issues associated with the use of moderated regression. *Organizational Behavior and Human Performance*, 34, 195–213.
- [33] Tu YK, Kellett M, Clerehugh V. (2005). Problems of correlations between explanatory variables in multiple

regression analyses in the dental literature. *British Dental Journal*; 199 (7):457–461.

- [34] Vasu E.S., Elmore P.B., (1975). The Effect of Multicollinearity and the Violation of the Assumption of Normality on the Testing of Hypotheses in Regression Analysis. Presented at the Annual Meeting of the American Educational Research Association; Washington, D.C. March 30–April 3.