Scientia Africana, Vol. 18 (No. 3), December, 2019. Pp 97-110
 ISSN 1118 – 1931

A REVIEW OF STANDARDIZATION AND CENTERING TECHNIQUES IN A MULTICOLLINEAR REGRESSION MODEL

[1]Ijomah Maxwell Azubuike and [2]Nwakuya Murren Tobe

[1,2]Department of Mathematics/Statistics
University of Port Harcourt, Choba, Rivers State.
maxwell.ijomah@uniport.edu.ng
murren.nwakuya@uniport.edu.ng

Received: 23-10-19
Accepted: 19-11-19

ABSTRACT

The issue of multicollinearity is well published but the available methods for multicollinearity correction is still debated. In this paper, we review the strength or performance of standardization and centering techniques in reducing multicollinearity problem in both linear and non-linear regression model using the variance inflation factor (VIF). A Monte Carlo simulation was carried out to show the precise effects of mean centering and standardization on both individual correlation coefficients as well as overall linear and non-linear model indices. Our findings reveal that use of centering and standardization are not very effective under severe collinearity. It is therefore hoped that practicing researchers will cautiously incorporate these diagnostics into their analyses.

Keywords: multicollinearity; variance inflation factor; standardization; centering, uncentered

INTRODUCTION

The efficiency of the prediction results in regression model always depends on the effective identification of multicollinearity in the data before actual prediction. The literature on linear models with special focus on multicollinearity spans several decades already. However, no conclusive solution has been achieved so far due to some deficiencies in its diagnostics, so that there is still a continuing active interest on the problem. Briefly, multicollinearity is a high degree of correlation (linear dependency) among several independent variables. It commonly occurs when a large number of independent variables are incorporated in a regression model. The presence of multicollinearity has some destructive effects on regression analysis such as prediction inferences and estimations. Consequently, the validity of parameter estimation becomes questionable (Montgomery et al., 2001; Kutner et al., 2004; Chatterjee and Hadi, 2006; Midi et al., 2010). As multicollinearity increases, the least squares estimates of the regression coefficients remain unbiased, but the determinants of the independent variables' covariance and correlation matrices approach zero, and the standard errors of the coefficients increase. Also, the expected distance between the vector of least squares coefficients and the vector of true regression coefficients increase with some estimates frequently having either unreasonably large values or unreasonable

signs. Moreover, slight sampling fluctuations in the estimates of the zero-order covariances can cause large swings in the values and signs of least-squares estimates of the coefficients in the presence of multicollinearity- a phenomenon someone once called the problem of the "bouncing betas" (Smith & Sasaki, 1979). As the determinant of the covariance matrix decreases, the rounding error in computing the inverse of the matrix, which is needed for the least-squares estimates. (Blalock, 1963; Gordon, 1968; Althauser, 1971; and Rockwell, 1975.)

The notion of centering and standardizing variables in regression as solution to multicollinearity is the source of constant debate and questioning. There is a big controversy in literature about the standardization and Centering of data in regression. Several conflicting views appear on the question of whether data in the X-matrix should be mean-centered or standardized before collinearity is assessed. Belsley (1984) contrasts with authors like Stewart (1987), Schall and Dunne (1987), Gunst (1983), Marquardt (1980) and Marquardt and Snee (1975) who advocate mean centering. There is less argument on the question on whether X should be standardized although the question of 'how the standardizing must be done could be vague. Stewart (1987) pointed out that any combination of three elements could be standardized: the matrix X, the vector β (its elements should be close together). The effects of predictor scaling on coefficients of regression equations (centered versus uncentered solutions) and higher order interaction effects has thoughtfully been covered by Aiken and West (1991). Their example illustrates that considerable multicollinearity is introduced into a regression equation with an interaction term when the variables are not centered. The variance inflation factor should detect the degree of multicollinearity when variables are uncentered (Freund, Littell, & Creighton, 2003). The problem of whether the observations should be centered around their mean or not before applying the diagnostic tools for multicollinearity is an issue which is still not completely resolved (Besley, 1984). In opposition to Smith and Campbell (1980), Marquardt (1980) states that the centering of observations removes nonessential ill conditioning. If the uncentered data is ill conditioned, then the small errors in inputs have large impact on the estimates of parameters. Kim (1987) pointed out that standardization can reduce multicollinearity among the linear, quadratic, and cubic terms is substantially reduced, while the correlation coefficients with other variables are not affected by this transformation. According to Kim (1993), without standardization, we may lose accuracy because of rounding errors in the course of calculating the variance or covariance. This is especially true when a variable with large values, such as income, is included as an independent variable in the regression equation, involving many variables and many cases. Furthermore, Marquardt and Snee (1975) argued, that "the ill conditioning that results from failure to standardize is all the more insidious because it is not due to any real defect in the data, but only the arbitrary origins of the scales on which the predictor variables are expressed". That is why they recommend standardizing whenever a constant term is present in the model. Belsley, Kuh and Welsch (1980), by contrast, indicated that "mean centering typically masks the role of

Scientia Africana, Vol. 18 (No. 3), December, 2019. Pp 97-110
© Faculty of Science, University of Port Harcourt, Printed in Nigeria *ISSN 1118 – 1931*

the constant term in any underlying near dependencies and produces misleadingly favorable conditioning diagnostics.

The purpose of this paper is to highlight a basic data analytic solution in fitting linear and non-linear collinear equations to an observation matrix when the independent variables are centered or standardized. Attention is restricted to collinear models in two independent variables (second order polynomial) but the approach may be applied in more complex situations. The variance inflation factor as a measure of the degree of multicollinearity however has not been examined in context with standardized, centered versus uncentered variables in both linear and non-linear regression equations.

The plan of the paper is as follows. In Section 2, various methods of dealing with multicollinearity were highlighted with theoretical framework on centering and standardization techniques. The relationship between centering and standardization in a regression model are considered in this section. The material and methods which involves simulation and results are presented in Section 3. In the last Section 4, findings and some conclusions are reported.

Dealing with Multicollinearity

A number of suggestions have been offered regarding how one should deal with data exhibiting high levels of multicollinearity. Virtually all of the approaches offered are aimed at increasing the precision of coefficient estimates. One suggestion that has been frequently made in trying to overcome the problem of multicollinearity is to collect new data (Ryan, 1997). Sometimes, the problem of multicollinearity occurs due to inadequate or erroneous data. Unfortunately, this is not always possible since some analysis must be based on the available data. Furthermore, this solution is not possible when the presence of multicollinearity is the result of internal constraints of the system being studied (Rawlings et al., 1998).

Hoerl (1962) and Hoerl and Kennard (1970a, 1970b) have proposed ridge regression as one way of overcoming the problems of multicollinearity using larger or more efficient samples and by increasing the numerical accuracy in one's data in order to reduce the size of the standard errors ; however, reasonable applications of these tactics frequently do not overcome the problems posed by the multicollinearity inherent in the models with cross product terms (see Marquardt and Snee, 1975; and Deegan, 1975, Henry, 1976.) The improvements suggested recently by Guilkey and Murphy (1975) and Kasarda and Shih (1977) make ridge regression even more attractive; however, ridge regression does not provide a minimum mean-square error for the model. The complication occurs in the choice of how the shrinkage is to be induced to achieve some 'optimality'. Oftentimes, there is no assurance that a particular choice of shrinkage would indeed result to desirable properties of the estimators.

Dropping of variables that duplicate the role of other 'more' important variables has been proposed as a natural solution to the multicollinearity problem. However, for models that strictly adhere to some theoretical framework, this is equivalent to massive loss of information. Dropping predictor variables is not only intellectually dishonest (Philippi 1993), it also

contaminates the remaining predictors (Box 1966). Suppose we believe that the true causal relationship involves all three predictors, but we drop X_3 to limit the effects of collinearity. In spite of this, the estimates of the regression coefficients remain contaminated by the eliminated predictor. Carnes and Slade, (1988) simulated apparent competition known to exist at least experimentally. They also tried to delete some variables from the model to resolve the multicollinearity problem, but realized that many important features of community organization/dynamics were lost. Thus, the true measure of competition cannot be assessed from a model that missed some important indexes in the competition framework because of the 'dropped' variables.

Then principal component regression is also proposed, hoping that the linear combination of the x's as a regressor will be able to keep all the variables into the model. Depending on the structure of the relationship among the regressors, component loadings may affect parameter estimates, e.g., loadings are similar, resulting to regression coefficients that are similar for all regressors.

This will cause problem in interpretation because the relative importance of predictors is being masked by the way the linear combination is formed.

Centering

Centering is defined as subtracting the mean (a constant) from each score, X, yielding a centered score. Aiken and West (1991) demonstrated that using other transformations, additive constant or uncentered scores can have a profound effect on interaction results. The X-scaled variables are assumed such that X'X has the form of a correlation matrix. Although centering in ordinary linear regression has been a subject of considerable debate recently (Hocking (1984), Snee (1983), Be1s1ey (1984b), it is generally recognized that centering reduces the condition number of the incidence matrix X in the (ordinary) linear regression model. As pointed out by Bradley and Srivastava (1979), centering in polynomial regression models is even more critical since the "intercorrelation" of the variables (X_1, X_2, X_3, etc.) becomes higher as the degree of the polynomial increases. Regression with higher order terms has covariance between interaction terms (XZ) and each component (X and Z) depends in part upon the means of the individual predictors. Rescaling, changes the means, thus changes the predictor covariance, yielding different regression weights for the predictors in the higher order function. Centering is therefore an important step in testing for interaction effects in multiple regression to obtain a meaningful interpretation of results.

Consider a regression model with two independent variables. When two independent variables are included in a multiple linear regression model, the model can be defined as

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i ,$$
$$i = 1,2,.....n, \qquad (1)$$

where the y is the score on the dependent variable of i the i^{th} subject, x_{1i} and x_{2i} are the values of the independent variables for the i^{th} subject, , and $\beta_0, \beta_1, \beta_2$ are population regression coefficients and ε_i is the error term assumed to be normally distributed with mean of zero and constant variance.

Scientia Africana, Vol. 18 (No. 3), December, 2019. Pp 97-110
© Faculty of Science, University of Port Harcourt, Printed in Nigeria *ISSN 1118 – 1931*

Often collinearity is assured by considering the explanatory variables only and ignoring the intercept β_0. This is accomplished by centering the response and predictor variables, which corresponds to fitting the model

$$\frac{1}{n}\sum_{i=1}^{n} Y_i = \beta_0 + \beta_1 \frac{1}{n}\sum_{i=1}^{1} X_{1i} + \beta_2 \frac{1}{n}\sum_{i=1}^{n} X_{2i} + \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i$$

this can be put in another notation as

$$\bar{Y} = \beta_0 + \beta_1 \bar{X}_1 + \beta_2 \bar{X}_2 \qquad (2)$$

Subtracting (2) from (1) implies a centered regression model without intercept

$$\begin{aligned} Y_i - \bar{Y} &= \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon \\ &-(\beta_0 + \beta_1 \bar{X}_1 + \beta_2 \bar{X}_2) \\ &= \beta_1 (X_{i1} - \bar{X}_1) + \beta_2 (X_{i2} - \bar{X}_2) + \varepsilon \end{aligned} \qquad (3)$$

which is given by

$$Y_{ic} = \beta_1 X_{1i}^{c} + \beta_2 X_{2i}^{c}$$

$$i = 1,....n$$

or in vector notation

$$Y^{c} = X^{c} \beta^{T}{}_{(\beta_0)} \qquad (4)$$

Noting that $\hat{\beta}_0 = \bar{Y} - \sum_{j=1}^{k} \beta_j \bar{X}_j$ in the original model, we observe that centering the predictors and the response forces the estimated intercept in that centered model to be 0. Hence, it can be dropped from that model.

Standardization

To obtain standardized regression coefficients, we can transform the variables of the mean as shown in (3) above by dividing with the standard deviation of y on both sides, and multiply each term on the right side by 1 in the form of $\frac{S_{x_j}}{S_{x_j}}$ to obtain

$$\frac{Y_i - \bar{Y}}{S_y} = \left(\beta_1 \frac{S_{x_1}}{S_Y}\right) \frac{(X_{1i} - \bar{X}_1)}{S_{x_1}} + \left(\beta_2 \frac{S_{x_2}}{S_Y}\right) \frac{(X_{2i} - \bar{X}_2)}{S_{x2}} + \frac{\varepsilon_i}{S_Y}$$

$$Y_i^{*} = \beta_0 + \beta_1 Z_{1i} + \beta_2 Z_{2i} + \varepsilon_i \qquad (5)$$

where $Z_{ij} = \frac{X_{i,j} - \bar{X}_j}{\sqrt{S_{jj}}}$

where $S_{jj} = \sum_{i=1}^{n} \left(X_{i,j} - \bar{X}_j\right)^2$

This leads to the standardized model

$$\frac{Y_i - \bar{Y}}{S_Y} = \hat{\beta}_1 \frac{(X_{i1} - \bar{X}_1)}{S_{x1}} + \hat{\beta}_2 \frac{(X_{i2} - \bar{X}_2)}{S_{x2}} + \varepsilon_i \qquad (6)$$

putting (6) in vector notation, we have

$$Y^{*} = Z\beta + \varepsilon^{*} \qquad (7)$$

The coefficients for the standardized model is given by

$$\hat{\beta}_j = \beta_j \left(\frac{S_j}{S_y}\right) \quad j = 1,2,....., k \qquad (8)$$

and thus the relationship between the estimates of the original and standardized regression coefficients is given by

$$\hat{\beta}_{jj} = \beta_j \left(\frac{1}{S_y}\right)^{\frac{1}{2}} \qquad j = 1,2,....., k$$

MATERIALS AND METHODS

The purpose of this paper is to review the impact of standardization and centering techniques in reducing multicollinear regression models using simulated incidence matrices and the COLLIN option in SAS 9.0 version PROC REG. To compare the performance of standardization and centering methods for dealing with collinearity, we simulated datasets with various range of predictor collinearity. Let X be a matrix of two independent variables: X_1, and X_2; and, let Σ be a variance-covariance matrix of a vector X. Our purpose is to generate the vector X from multivariate normal distribution with mean zero vector and variance-covariance matrix Σ, where Σ is a symmetric 2 by 2 matrix and a positive definite. We can obtain correlated normal variables X = C ′ Z where Z are standard normal random variables and C is an upper triangular matrix such that C ′ C = Σ for any positive definite matrix Σ. We begin by multiplying each of the selected variable with 0.5 in order to make the variables uniform. U = ranuni (start), X_1=U+rannor (start)*0.5 X_2=U+rannor (start)*0.5 with Y=1+X_1+X_2+rannor (start). For our simulation experiment, we created datasets that had sample size n (n = 1000) and two explanatory normally distributed variables (predictors). Sample sizes of 100, 200, 250 and 500 were considered; however, in order to obtain a more complete assessment, a small sample (50) and very large sample (1000) were included as well. The simulation study is performed to examine how collinearity between two independent variables in multiple linear and non-linear regression models could be reduced using both standardization technique and centering method.

RESULTS

The results of the simulation are presented in Tables I and 2. In order to save space only the results for correlations 0.5, 0.7 and 0.9 are reported here. Table 1 displays the mean variance inflation factors (VIF) of each method in a linear and non-linear component observed in the 1000 sample. In this table, the result showed that the centered and standardized techniques outperformed the OLS (uncentered). As expected, the VIF increases with increase in correlation between the independent variables. Both the centered and standardized techniques exhibited absence of collinearity (with minimum VIF) in the linear model following Hair et al. (1995) which suggest variance inflation factors (VIF) less than 10 is indicative of inconsequential collinearity. VIF values calculated using centering and standardization methods remained relatively constant over sample sizes and simulation runs for both linear and non-linear models. This agrees with the findings that mean centering and standardization help reduce potentially bad effects of interrelated variables (Irwin & McClelland, 2001; Jaccard, Wan, & Turrisi, 1990; Smith & Sasaki, 1979). However, as the collinearity gets more severe $(\rho_{x1x2} \geq 0.9)$, the presence of multicollinearity becomes pronounced. With sample size of 50, the centered model performed better than standardized technique but as the sample size increases, both performed equally. That is to say that centered technique performs better for small samples. For the non-linear model, the centered method performed better than the standardized technique irrespective of the sample size. There was a pronounced difference in the mean VIF

Scientia Africana, Vol. 18 (No. 3), December, 2019. Pp 97-110
© Faculty of Science, University of Port Harcourt, Printed in Nigeria *ISSN 1118 – 1931*

between centered and standardized method. We therefore recommended that use of centering should be preferred to standardizing technique when dealing with multicollinearity problems for non-linear models.

Table 1: **Mean VIF for both linear and non-linear models**

n		Linear model			Non-linear model		
	ρ_{x1x2}	**Uncentered**	**Centered**	**Standardized**	**Uncentered**	**Centered**	**Standardized**
	0.5	32.38785	4.7799	6.7062	194.8376	1.7416	49.3982
	0.7	58.837	6.2940	9.6415	345.7295	3.5393	76.4091
50	0.9	124.2699	14.3211	19.8592	1002.247	23.9121	281.6419
	0.5	10.6838	3.8116	4.3751	49.3685	1.5942	13.7337
	0.7	27.5349	5.8824	6.1137	150.9621	2.5328	31.6597
100	0.9	89.7225	13.62775	13.6056	632.2307	10.7475	131.789
	0.5	8.9708	4.04375	3.8669	55.1101	1.5679	17.8093
	0.7	21.1189	5.92958	5.6631	141.0235	2.8102	37.2936
200	0.9	70.68965	13.61145	14.3445	560.7861	15.10735	153.3723
	0.5	9.31275	3.89265	3.76735	55.7062	1.53835	17.9999
	0.7	20.8460	5.5562	5.3882	136.4729	2.7364	35.9106
250	0.9	65.4489	12.8681	12.875	526.3433	14.4871	140.9292
	0.5	8.70695	3.7781	3.8147	51.6663	1.69025	17.44865
	0.7	19.1807	5.47725	5.6217	126.5204	3.13035	35.2614
500	0.9	63.4686	13.5421	13.8173	542.9985	16.3107	152.2818
	0.5	8.42425	3.72435	3.8946	49.8798	1.7922	17.3781
	0.7	18.2917	5.48925	5.8273	122.0067	3.39	35.3736
1000	0.9	60.2517	13.9542	14.5036	539.849	18.082	157.6132

The graphs in figures 1a and 1b show the mean VIF for both linear and non-linear models at different levels of collinearity. The effect of increasing sample size is clear from the graphs showing that at the large sample sizes both variables have a low VIF indicating stable behaviour over the simulation runs. This suggests that very large sample sizes negate the effect of multicollinearity in that coefficient estimation of highly collinear variables becomes relatively stable. Figure 1a shows that centering exhibited slight lower VIF when the sample is 50 but approximates the standardized method as the sample increases. For the non-linear graph in fig. 1b, centering again maintained a stable lower VIF, this time all through the samples. As expected, the VIF values drop sharply if the correlated variables are removed from the fitted model.

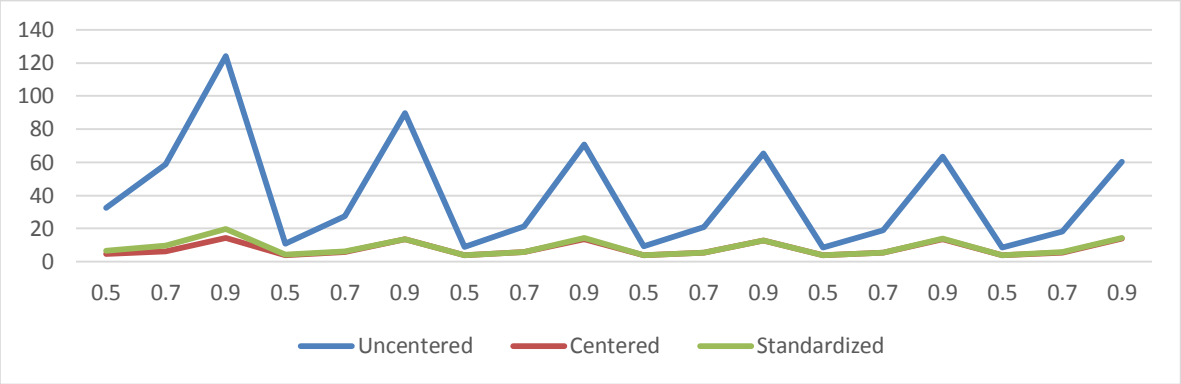

Fig. 1a: Comparison of mean VIF with a linear model for different scaling methods.

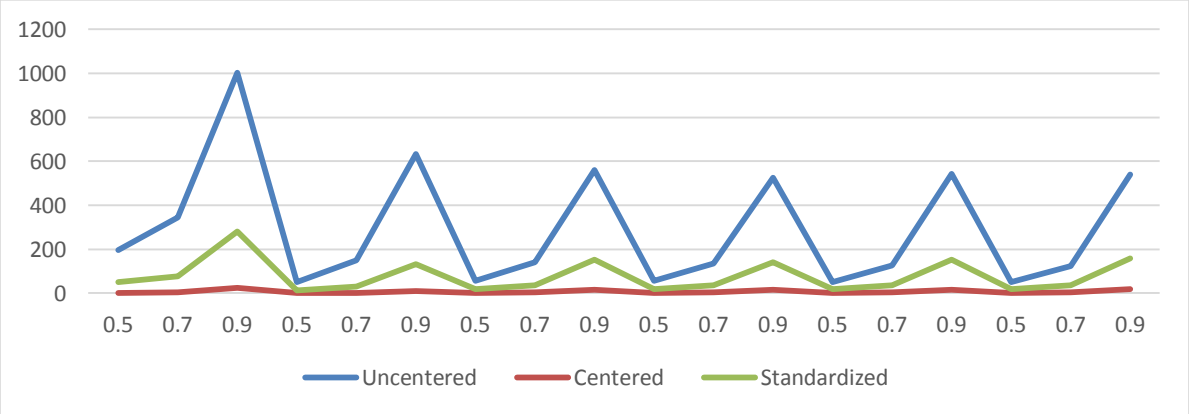

Fig. 1b: Comparison of mean VIF with non-linear model for different scaling methods.

To assess the effect of centering and standardization techniques on the stability of coefficient estimation when multicollinearity is present, the mean Standard Error (SE) of the coefficients of each of the variables was calculated under linear and non-linear models. Table 2 shows that centering as well as standardized model were more reliable estimates than the uncentered for linear component since the standard error of the coefficient indicates the precision of the coefficient estimates. The importance of centering and standardization techniques become even more clear when one considers the graphs of these techniques as indicated in figures 2a and 2b below. However, figure 2b again confirmed the superiority of centering over standardization in solving multicollinear problems for non-linear model since it maintains a minimum standard error for the samples under study.

Table 2: Mean Standard Error of regression coefficients on both linear and non-linear models

n		Linear model			Non-linear model		
	ρ_{x1x2}	**Uncentered**	**Centered**	**Standardized**	**Uncentered**	**Centered**	**Standardized**
50	0.5	0.2542	0.1016	0.1197	0.0655	0.0184	0.0469
	0.7	0.3944	0.1324	0.1635	0.1029	0.0349	0.0718
	0.9	0.6259	0.2130	0.2509	0.1861	0.1087	0.1429
100	0.5	0.0973	0.0581	0.0622	0.0204	0.0090	0.0144
	0.7	0.1772	0.0823	0.0839	0.0418	0.0167	0.0277

Scientia Africana, Vol. 18 (No. 3), December, 2019. Pp 97-110
© Faculty of Science, University of Port Harcourt, Printed in Nigeria *ISSN 1118 – 1931*

	0.9	0.3492	0.1373	0.1334	0.0952	0.0476	0.0672
	0.5	0.0652	0.0438	0.0428	0.0166	0.0067	0.0123
	0.7	0.1135	0.0601	0.0587	0.0309	0.0126	0.0221
200	0.9	0.2273	0.0997	0.1024	0.0683	0.0392	0.0527
	0.5	0.0604	0.0390	0.0384	0.0151	0.0060	0.0112
	0.7	0.1030	0.0532	0.0524	0.0277	0.0112	0.0198
250	0.9	0.2009	0.0891	0.0891	0.0607	0.0346	0.0465
	0.5	0.0403	0.0266	0.0267	0.0099	0.0042	0.0075
	0.7	0.0681	0.0364	0.0369	0.0181	0.0082	0.0134
500	0.9	0.1361	0.0629	0.0635	0.0422	0.0259	0.0333
	0.5	0.0279	0.0186	0.0190	0.0068	0.0030	0.0052
	0.7	0.0466	0.0256	0.0263	0.0125	0.0059	0.0094
1000	0.9	0.0929	0.0447	0.0456	0.0296	0.0189	0.0238

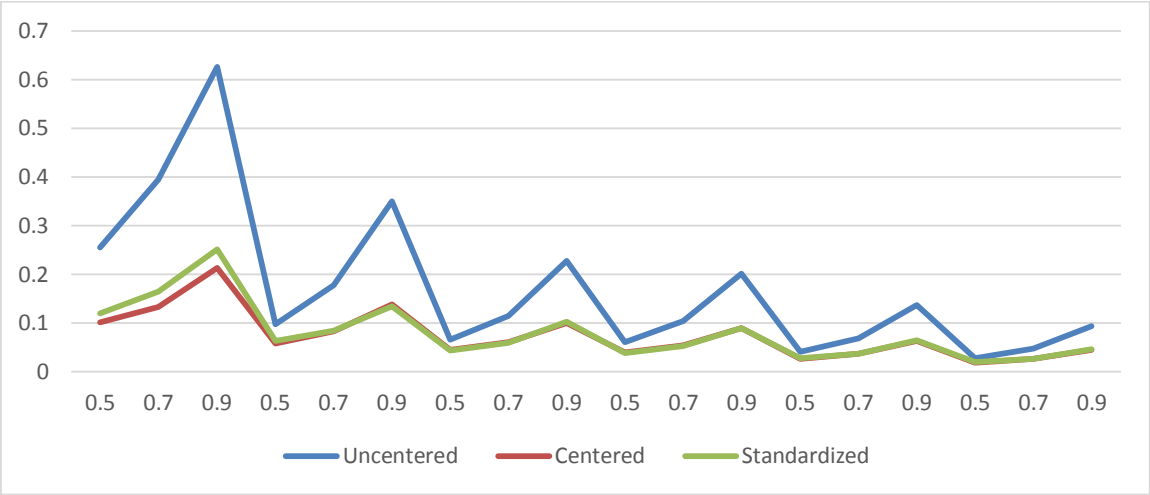

Fig. 2a: *Comparison of mean standard error of regression coefficients for linear model in Centering and Standardized methods.*

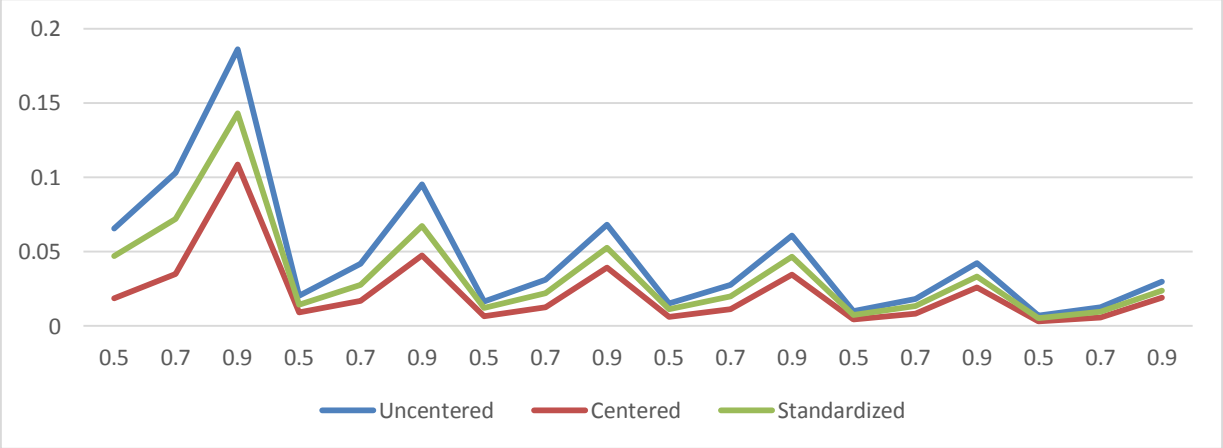

Fig. 2b: *Comparison of mean standard error of regression coefficients for non-linear model in Centering and Standardized methods.*

VIF at n = 200

$\rho_{x_1x_2}$	OLS		Centered		Standardized	
	Linear	Non-Linear	Linear	Non-Linear	Linear	Non-Linear
	9.5212	59.7188	4.28	1.4980	4.0551	20.5896
0.50	8.4204	50.5014	3.80754	1.6378	3.6787	15.0290
	13.4625	88.2977	4.9888	1.8873	4.7097	27.2204
0.60	13.2247	85.5925	4.5834	2.0556	4.3031	22.8969
	20.1678	135.5114	6.12365	2.6875	5.8207	38.0424
0.70	22.0699	146.5356	5.7355	2.9328	5.5055	36.5448
	30.9466	212.4456	7.9418	4.4209	7.6758	56.6317
0.80	36.1635	247.4167	7.4775	4.8753	7.6260	60.7676
	64.7614	492.4905	13.9964	14.2672	14.0010	137.1997
0.90	76.6179	629.0817	13.2265	15.9475	14.6880	169.5448
	152.2208	1626.5412	30.5533	74.4346	31.6656	548.1982
0.96	172.6957	2238.3277	29.3080	82.0857	33.4013	719.5539

VIF at n = 250

$\rho_{x_1x_2}$	OLS		Centered		Standardized	
	Linear	Non-Linear	Linear	Non-Linear	Linear	Non-Linear
	9.7717	58.9661	4.0524	1.5102	3.9881	20.2632
0.50	8.8538	52.4462	3.7329	1.5665	3.5466	15.7365
	13.5271	85.3520	4.6449	1.91166	4.5658	26.2091
0.60	13.6213	86.3870	4.4209	1.9527	4.1488	23.2229
	19.8026	129.3417	5.6647	2.7244	5.5403	36.0936
0.70	21.8893	143.6041	5.4476	2.7484	5.2361	35.7275
	29.8465	202.5697	7.3865	4.4572	7.1582	53.4972
0.80	34.5188	236.1196	7.0286	4.4701	7.1004	57.2919
	61.1124	474.6687	13.3162	14.0268	12.6201	130.0257
0.90	69.7854	578.0179	12.4199	14.9473	13.1299	151.8327
	139.7615	1567.3052	29.6897	70.3390	27.6343	514.9495
0.96	152.7096	1989.7931	27.9396	70.4712	28.9476	622.5118

VIF at n = 500

$\rho_{x_1x_2}$	OLS		Centered		Standardized	
	Linear	Non-Linear	Linear	Non-Linear	Linear	Non-Linear
	9.5017	60.7947	4.1830	1.6859	3.9978	20.3068
0.50	7.9122	42.5379	3.3732	1.6946	3.6316	14.5905
	13.6209	89.5747	4.8043	2.1684	4.6634	26.9523
0.60	11.4973	68.3127	4.0116	2.1699	4.3072	20.9741
	20.4800	137.2792	5.8588	3.1389	5.7891	38.0033
0.70	17.8814	115.7615	5.0957	3.1218	5.4543	32.5195

Scientia Africana, Vol. 18 (No. 3), December, 2019. Pp 97-110
© Faculty of Science, University of Port Harcourt, Printed in Nigeria *ISSN 1118 – 1931*

0.80	31.2945	217.1005	7.6439	5.1979	7.6490	57.7720
	28.5769	202.1484	6.9099	5.1345	7.3648	54.2658
0.90	64.3305	528.1081	13.9202	16.4712	13.8657	148.7279
	62.6067	557.8889	13.1639	16.1501	13.7689	155.8357
0.96	147.7285	1872.7307	31.7235	82.2972	30.8281	622.8872
	148.1251	2093.3441	30.7316	80.7850	31.0774	675.0853

VIF at n = 1000

$\rho_{x_1x_2}$	OLS		Centered		Standardized	
	Linear	Non-Linear	Linear	Non-Linear	Linear	Non-Linear
0.50	9.4159	59.1879	3.9701	1.8054	3.9831	20.3650
	7.4326	40.5717	3.4786	1.7790	3.8061	14.3912
0.60	13.6055	88.4174	4.6219	2.23414	4.7478	27.4625
	10.5443	63.7005	4.1264	2.3048	4.4861	20.3854
0.70	20.5432	138.1499	5.7319	3.4157	6.0155	39.5322
	16.0402	105.8635	5.2466	3.3643	5.6391	31.2150
0.80	31.4488	223.8855	7.60542	5.6892	8.0755	61.5499
	25.3215	183.1277	7.1517	5.6199	7.5653	51.8542
0.90	64.60149	569.2332	14.1279	18.1211	14.8781	164.2748
	55.9019	510.4648	13.7805	18.0429	14.1291	150.9516
0.96	146.7917	2093.8852	32.4831	90.8335	33.2649	701.4855
	135.3616	1985.3914	32.2982	91.0454	32.2181	675.8134

CONCLUSION

In this paper, we reviewed the importance of centering and standardization technique in dealing with multicollinearity problem using variance inflation factor from a hypothetical regression model. This procedure allowed us to identify how VIF as a collinearity diagnostic responds to centered or standardized technique in linear and non-linear models. Our findings revealed that use of centering and standardization play an important role in regression model especially when collinearity is not very severe. Although the performance of the fitted models using both techniques seem not to be adversely affected by the presence of multicollinearity, on close inspection the values of the VIFs vary substantially, especially at small sample sizes, sometimes may result to misleading interpretation of the regression coefficients. It is important to note here that centering and standardization are not very effective under severe collinearity as shown in the results (Ijomah 2019, Dalal & Zickar, 2012; Echambadi & Hess, 2007). This is an indication that the use of centering and standardization is restricted to the degree of severity of multicollinearity. Based on our analysis, we conclude that centering and standardization techniques as solution to multicollinearity are only necessary for moderate collinearity.

REFERENCE

Aiken, L.S. & West, S.G. (1991). Multiple Regression: Testing and Interpreting Interactions. Sage Publications: Thousand Oaks, CA.

Althauser, R. P. (1971) "Multicollinearity and non-additive regression models," pp. 453-472 in H. M Blalock, Jr. (ed.) Causal Models in the Social Sciences. Chicago: Aldine-Atherton.

Be1s1ey, D. A. (1984b). "Demeaning Conditioning Diagnostics Through Centering," The American Statistician 38, 73 – 93

Belsley, D.A. (1984): Demeaning conditioning diagnostics through centering, and Reply. The American Statistician, 38, 73-77 and 90-93.

Belsley, D.A., Kuh, E. & Welsch, R.H. (1980). Regression diagnostics: Identifying influential data and sources of collinearity. New York: John Wiley & Sons.

Blalock, H. M., Jr. (1963) "Correlated independent variables: the problem of multicollinearity." Social Forces 62: 233-238.

Box, G.E.P. (1966) Use and abuse of regression. Technometrics 8,625-629.

Bradley, R. A.; Srivastava, S. S. (1979). Correlation in Polynomial Regression. Am. Stat. 33, 11-14.

Carnes, B. A., and N. A. Slade (1988). The use of regression for detecting competition with multicollinear data. Ecology 69:1266-1274.

Chatterjee, S. and A.S. Hadi, 2006. Regression Analysis by Example. 4th Edn., Wiley, New York, ISBN-10: 0471746966.

Dalal, D. K., and Zickar, M. J. (2012). Some common myths about centering predictor variables in moderated multiple regression and polynomial regression. *Organizational Research Methods, 15*(3), 339–362

Deegan, J., Jr. (1975) .The process of political development: an illustrative use of a strategy for regression in the presence of multicollinearity. Soc. Methods & Research 3: 384-415.

Echambadi, Raj and James D Hess (2007), "Mean-Centering Does Not Alleviate Collinearity Problems in Moderated Multiple Regression Models," *Marketing Science*, 26 (3), 438-48.

Freund, R.J., Littell, R.C., and Creighton, L. (2003). Regression Using JMP. Cary, NC: SAS Institute, Inc.

Gordon, R. A. (1968) "Issues in multiple regression." Amer. J: of Sociology 73: 592616.

Guilkey, D. K. and Murphy J.L. (1975). Directed ridge regression techniques in cases of multicollinearity. Journal of the Amer. Statistical Assn. 70: 769-775.

Gunst, R.F. (1983), "Regression Analysis with Multicollinear Predictor Variables: Definition, Detection and Effects," Communications in Statistics, A12 (Special Issue on Statistical Reviews), 2217-2260.

Hair, J. F., Jr., Anderson, R. E., Tatham, R. L. and Black, W. C. (1995) Multivariate Data Analysis, 3rd ed, Macmillan Publishing Company, New York.

Henry, N. W. (1976). A note on ridge regression. Soc. Methods & Research 4: 495-500.

Scientia Africana, Vol. 18 (No. 3), December, 2019. Pp 97-110
© Faculty of Science, University of Port Harcourt, Printed in Nigeria *ISSN 1118 – 1931*

Hoerl, A.E., (1962). Application of ridge analysis to regression problems. Chem. Eng. Prog. 58: 54-59.

Hoerl A.E. and Kennard R.V. (1970a): Ridge regression: Applications to non-orthogonal problems. Technometrics 12, 69-82

Hoerl, A.E. and Kennard R.V. (1970b): Ridge regression: Baised estimation for non-orthogonal problems. Technometrics 12, 55-69.

Hocking, R.R. (1983), "Developments in Linear Regression Methodology: 1959-1982" (with discussion), Technometrics, 25, 219-249.

Ijomah, M.A. (2019), "Standardization Technique and Centering on the Variance Inflation Factor of a Structured Collinear Model" Probability Statistics and Econometric Journal. 1(1): 13-25,

Irwin, J. R., & McClelland, G. H. (2001). Misleading heuristics and moderated multiple regression models. Journal of Marketing Research, 38(February), 100–109.

Jaccard, J., Wan, C. K., & Turrisi, R. (1990). The detection and interpretation of interaction effects between continuous variables in multiple regression. Multivariate Behavioral Research, 25(4), 467–478.

Kasarda, J. D. and Shih W.F.P (1977). Optimal bias in ridge regression approaches to multicollinearity. Soc. Methods & Research 5: 461-470.

Kim, Doo-Sub, (1987). Socioeconomic Status, Inequality and Ferti1ity. The Population and Development Studies Center, Seoul Nationa1 University. Seoul, Korea.

Kutner, M.H., C.J. Nachtsheim and J. Neter, (2004). Applied Linear Regression Models. 4th Edn., McGraw Hill, New York.

Marquardt, D.W. (1980). Comment on "A critique on some ridge regression methods" by G. Smith and F. Campbell: "You should standardize the predictor variables in your regression models". Journal of the American Statistical Association, 75(369), 87-91.

Marquardt, D. W. and Snee R.D. (1975). Ridge regression in practice. Amer. Statistician 29 (February): 3-20.

Midi, H., A. Bagheri and A.H.M.R. Imon, (2010). The application of robust multicollinearity diagnostic method based on robust coefficient determination to a non-collinear data. J. Applied Sci., 10: 611-619

Montgomery, D.C., E.A. Peck and G.G. Vining, 2001. Introduction to Linear Regression Analysis. 3rd Edn., Jon Wiley and Sons, New York, USA., 672.

Philippi, T.E. (1993). Multiple Regression: Herbivory. Design and Analysis of Ecological Experiments (eds S.M. Scheiner and J. Gurevitch), pp. 183-210. Chapman and Hall, New York.

Rawlings. J.O, Pantula SG, Dickey DA. (1998). Applied Regression Analysis: Research Tool. 2. New York, NY: Springer-Verlag New York, Inc;.

Rockwell, R. C. (1975) "Assessment of multicollinearity: the Haitovsky test

of the determinant." Soc. Methods & Research 3 (February): 308-320.

Ryan, T., (1997). Modern Regression Methods. Har/Dis Edn., Wiley, New York, USA.,

Schall .R and Dunne T.T. (1987): A note on the relationship between parameter collinearity and local influence, Technical Report.1/88 Institute for Biostatistics of the SA Medical Research Journal Tygerberg 7505.

Snee R.D. (1977): Validation of regression models: methods and examples. Technometrics, 19, 415-428.

Snee R.D. (1983): Discussion of "Developments rn Linear Regression Methodology: 1959-1982" by R.R. Hocking. Technometrics, 25, 230-237.

Smith K.W. and Sasaki M.S. (1979). Decreasing Multicollinearity: A Method for Models with Multiplicative Functions. Sociological Methods & Research, 8 no. I, 35-56 e 1979 Sage Publications, Inc.

Smith G. and Campbell F.(1980): A critique of some ridge regression methods. Vith comments by Ronald A. Thisted, Donald V. Marquardt, R. Craig Van Nostrand,D. V. Lindley, Robert L. Obenchain, Lawrence C. Peele, Thomas P. Ryan, H. D. Vinod and Richard F. Gunst, and with a reply by the authors. Journal of the American Statistical Association, 75, no. 369, 74-103.

Stewart, G.W. (1987). Collinearity and least squares regression. Statistical Science, 2(1), 68-84.