Academic Journal of Statistics and Mathematics (AJSM)
Vol.6, No.4; April - 2020;
ISSN (5730 – 7151);
p –ISSN 4052 – 392X
Impact factor: 3.92

ANOTHER CAUTIONARY NOTE ABOUT MULTICOLLINEARITY PROBLEM IN A LINEAR REGRESSION: A NEW LOOK AT AN OLD ISSUE

Ijomah Maxwell Azubuike and Nduka Ethelbert Chinaka
Department of Mathematics/Statistics, University of Port Harcourt, Choba, Rivers State.

ABSTRACT: The problem of Multicollinearity in a regression model is often considered as the correlation between two or more explanatory variables (regressors). Though the literature on ways of coping with collinearity is extensive, relatively little effort has been made to clarify the conditions under which collinearity affect estimates developed with multiple regression analysis or how pronounced those effects are. In this paper, effort is made to capture the nature of collinearity within the regressors as well as between the regressand and regressors in both near perfect positive and negative correlation. Furthermore, most literature on how number of correlated regressors affect the degree of collinearity in a regression model are very scanty. Our results provide critical insight that both helps avoid misleading interpretations and yields better understanding for the impact of intercorrelation among predictor variables in linear regression analyses.

Key words: multicollinearity; collinearity; regressors; regressand; correlation;

1. INTRODUCTION

Collinearity is often viewed as a situation in which two or more predictor variables are highly correlated. This statement seems to be inadequate since correlation is a statistical property of random variables, and the regressor variables need not be stochastic, since they could represent preselected variable values in a designed experiment. When there are no preselected variable values the 'correlation' may be merely a characteristic of a period, say, and might not be same type at other times [1]. Metaphorically, multicollinearity is like a red herring in a mystery novel: seemingly guilty, but actually innocent, wasting the time of the detective in the search for the culprit [2]. As literature indicates, collinearity misleadingly inflates the standard error in an excessive amount, leading to unstable p-values and causing wider confidence intervals thereby increasing the chance to reject the significant test statistic. In such case, the coefficient may provide high estimates of changes in the multiple regressions when only low changes can be seen in the model or the data [[3-6].This causes the regression estimates to be inaccurate showing wrong signs and doubtful extent for some predictor variables since the effects of these variables are all assorted. Moreover, it indicates that any little change in the data may result to sizeable variation in regression coefficients thereby making interpretation more difficult as a result of lot of

Academic Journal of Statistics and Mathematics (AJSM)
Vol.6, No.4; April - 2020;
ISSN (5730 – 7151);
p –ISSN 4052 – 392X
Impact factor: 3.92

frequent deviation in the variables. [7-11]. The problem of multicollinearity in regression is well known and published but what constitute collinearity has not been well established. Many researchers have examined one or more aspects of the multicollinearity problem, each having different views. As Cook [12] rightly pointed out that collinearity evidently implies different things to different people. Some associate collinearity primarily with numerical problems and sensitivity, while others concentrate on variance inflation and related statistical concerns. Few have attempted to incorporate, or even to distinguish between either multicollinearity's nature and effects, or its diagnosis and cure. A considerable number of researchers think of multicollinearity problem as a discontinuous state which either exist or does not exist instead of a continuous process whose strength may be measured. And this has led to a great deal of confusion and some inconsistency [12]. Cohesion requires, first of all, a clear distinction between multicollinearity's nature and effects, and, second, a definition in terms of the former on which diagnosis, and subsequent correction can be based. Kendall tries to achieve a solution neglecting the problem's nature while Klein was of the opinion that the problem involves both nature and effects. Possibly, a further confusing view of the practice is that they propose no collinearity problems are likely to occur in either cases since the confidence interval (CI) is below the limit of 30 and the VIF is substantially below 10.These problems arise because the diagnostics do not accommodate differences between negative and positive correlations, nor do they consider relationships with the dependent variable but are seen as a feature of the independent variables set alone[5].

According to Farrar and Glauber [13], collinearities are due to 'weak' or 'deficient' data. 'Deficient' implies an aberration in the data collection process; however, collinear ties among predictor variables are sometimes an inherent property of the phenomenon under study, in which case 'deficient' data would actually be a misnomer. In the foregoing statements, there are associations which imply more about the predictor variables than simply the existence of a collinearity. Collinearity can cause not only parameter variance estimates to increase but also decrease and this is why collinearity problem is very insidious [5, 14, 15].

From the above premises, no account whatever is taken of the extent, or even the existence, of dependence between the regressand (dependent variable) and the regressors since the effects of collinearity are moderated by the correlations between the regressors and the dependent variable [13]. It is true, of course, that the effect on estimation and specification of interdependence in the explanatory variables is reflected by variances of estimated regression coefficients which also depends partly on the strength of dependence between the regressand and the regressors. The predictor variables or regressors are not slighter multicollinear if correlated to one regressand than correlated to another regressor despite the influence may be more severe in one case than the other. Again, literature on multicollinearity from the point of view of negative and positive relationship among the regressors and the regressand are very scanty and are accompanied with conflicting results. In order to treat the problem, however, it is important to distinguish between negative and positive relationship among the regressors, and to develop diagnostics based on the former. This paper considers misconceptions about collinearity and how empirical challenges associated with it can be addressed. The work therefore aims to perform a simulation study with various scenarios of different collinearity structures to investigate the effects of collinearity under positive and negative correlation amongst regressors and regressand and to compare these results with existing guidelines in deciding the degree of collinearity.

In the sections to follow, we will describe the material and methods that will be used in this study, analyze some characteristics of positive and negative correlations within

Academic Journal of Statistics and Mathematics (AJSM)
Vol.6, No.4; April - 2020;
ISSN (5730 – 7151);
p –ISSN 4052 – 392X
Impact factor: 3.92

regressors and between regressors and the regressand. We will discuss influence negative and positive correlation among the variables to determine the model's accuracy, and also assess the predict power of the estimated model. Finally, we will discuss the potential ramifications of how the data might be used to change and/weaken collinearity.

2. Materials and Methods

To examine the effect of bias due to the presence of collinearity among the regressors in multivariable linear models, a simulation study is was carried out. Let X be a matrix of three independent variables: X_1, X_2, and X_3; and, let Σ be a variance-covariance matrix of a vector X. Our aim is to obtain the vector X with mean zero and a symmetric 3 by 3 variance-covariance matrix Σ from multivariate normal distribution and a positive definite. Consider the linear regression model of the form:

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \beta_3 X_{3t} + \mu_t \quad (1)$$

Where X = 1, 2, 3, and $X_t \sim N(0,1)$ are stochastic and correlated.

For Monte-Carlo simulation study, we generate data by RANNOR (START) using SAS 9.0 version, the parameters of equation (1) were specified and initially fixed as $\beta_0 = 1$, $\beta_1 = 0.5$, $\beta_2 = 0.5$ and $\beta_3 = 0.5$. In order to obtain varying collinearity, we also considered when, $\beta_0 = 1$, $\beta_1 = 1$, $\beta_2 = 1$ and $\beta_3 = 1$. Thereafter, we also subjected the coefficients to $\beta_0 = 1$, $\beta_1 = 1.5$, $\beta_2 = 1.5$ and $\beta_3 = 1.5$. The levels of intercorrelation (multicollinearity) among the independent variables were determined by adjusting (u = 5, 10, 30). By doing so, we obtain severe, moderate and low/no collinearity. With the simulated data, we showed how the regressors at various degrees of multicollinearity influences the parameter estimates and standard errors in the model. Four different scenarios are considered, which have different correlation structures and association among the variables. First, we begin with the case which assumes that three explanatory variables are positively correlated with each other and with the dependent variable in a multivariable model. In this scenario, we investigate the effect of bias in estimates of related regressor coefficients due to the variation in the degree of collinearity between the three variables in the model. Second, we extend the first scenario but this time the positively correlated explanatory variables were correlated with negative dependent variable. Thus, there exist only positive correlation within the explanatory variables. Third, we also consider case where the three explanatory variables are strongly positively related but were not correlated with the regressand. Finally, we consider the case were two of the explanatory variables are negatively correlated but positively correlated with the third explanatory variable.

3. Results and Discussion

The full summary of the simulated results of each estimator for small (n =20) and large (n =100) sample sizes, correlation and nature of relationships are shown in the tables below. Tables 1 to 4 provides the averaged estimates of regression coefficient, standard errors, t-test statistics, p-values, VIF and CN under each correlation scenario. The comparisons, measured in percent change, of estimates from models with higher, moderate and low degree of multicollinearity to estimates from the model at different sample size.

Case 1: positive correlation among three predictor variables (X_1, X_2 and X_3) with dependent variable Y.

This scenario assumes that three (X_1, $X_2$and X_3) are positively correlated with each other and have positive relationship with the regressand (Y). We consider positive correlation among variables for severe, moderate and low collinearity for small (n=20) and large (n = 100) samples. We then examined how and which of the parameter estimates (i.e. β_1, β_2, and β_3) are affected in the bias extent at varying degrees of correlation and sample size. Our main interest in this scenario is to examine the influence of positive correlation among the explanatory variables to the distortion in estimates of three correlated parameter coefficients, β_1, β_2 and β_3. As shown in Table 1, both

Academic Journal of Statistics and Mathematics (AJSM)
Vol.6, No.4; April - 2020;
ISSN (5730 – 7151);
p –ISSN 4052 – 392X
Impact factor: 3.92

estimates of β_1 and β_2 have very similar variation of effects in magnitude under positive correlation structure for small sample. Their parameter estimates decreased smoothly unlike β_3 which increased as collinearity gets more severe in both small and large samples. The estimate β_1 is -0.0872 and -0.0721 respectively for small and large samples when it is uncorrelated, but show a bias almost three times as great under severe collinearity (β_1 = -0.2508) for small sample while for large sample,β_1 show a bias of almost twice decrease (β_1 = -0.1497). However, as the sample increases, β_2 and β_3 showed an increase as the collinearity gets stronger. The estimate of β_3 is similarly biased to the estimate of β_2. It also showed a dramatic increase when collinearity becomes strong. The coefficients β_2, and β_3 appeared less significant especially for small samples unlike when the sample is large both coefficients remained significant. For β_1, it remained insignificant as the sample and degree of collinearity increased.

Table 1: Collinearity ofpositive correlation among three predictor variables(X_1, X_2, X_3) and with the dependent variable Y

Degree of Collinearity	RMSE	R^2	Coefficient	SE coeff.	t-value	Prob	VIF	CI
Low/No Collinearity								
n =20	0.9050	0.9240	B_0 = 2.2517 β_1=-0.0872 β_2 = 0.9692 β_3 = 1.0778	0.4129 0.1616 0.1874 0.1479	5.45 -0.54 5.17 7.29	0.0000 0.5968 0.0000 0.0000	- 1.90 1.53 1.88	1.0000 4.0312 4.8567 5.7199
n =100	1.1084	0.8963	B_0 = 2.1450 β_1 = -0.0721 β_2 = 0.9762 β_3 = 1.0277	0.2101 0.0650 0.0688 0.0634	10.20 -1.11 14.19 16.22	0.0000 0.2700 0.0001 0.0001	- 1.41 1.26 1.42	1.0000 3.4521 3.9521 4.2255
Moderate Collinearity								
n =20	0.9039	0.9728	B_0 = 2.2880 β_1=-0.1283 β_2 = 0.9649 β_3 = 1.1344	0.4454 0.2393 0.2441 0.2246	5.14 -0.54 3.95 5.05	0.0001 0.5992 0.0011 0.0001	- 8.61 7.66 8.79	1.0000 4.6231 13.5917 14.7712
n =100	1.1009	0.9636	B_0 = 2.2497 β_1 = -0.0909 β_2 = 0.9752 β_3 = 1.0575	0.2155 0.0923 0.9752 0.0911	10.44 -0.99 10.63 11.61	0.0001 0.3268 0.0000 0.0000	- 6.33 5.61 6.28	1.0000 4.1839 10.6060 11.0452
Very strong Collinearity								
n =20	0.9036	0.9968	B_0 = 2.2941 β_1= -0.2508 β_2 = 0.9528 β_3 = 1.2878	0.4445 0.4765 0.4611 0.4614	5.16 -0.53 2.07 2.79	0.0000 0.6058 0.0554 0.0131	- 280.3 257.8 271.3	1.0000 4.6216 80.1373 84.4736
			B_0 = 2.2783	0.2118	10.76	0.0000	-	1.0000

Academic Journal of Statistics and Mathematics (AJSM)
Vol.6, No.4; April - 2020;
ISSN (5730 – 7151);
p –ISSN 4052 – 392X
Impact factor: 3.92

n =100	1.0976	0.9960	β_1 = -0.1497 β_2 = 0.9840 β_3 = 1.1439	0.1834 0.1772 0.1815	-0.82 5.56 6.30	0.4165 0.0000 0.0000	204.9 187.9 201.0	4.1665 62.3840 65.3059

Figure 1 gives a clearer picture of the three coefficients β_1, β_2 and β_3, at various degrees of collinearity between the regressand and the regressors. A closer look at the graph shows the estimateβ_3 drifted upward in with positive collinearity among the three regressors and the regressand for small and large samples. Though in the case of large sample, the rate of increase in β_3 was minimized. The estimate β_2 dropped slightly when the sample was small but again slightly increased as the sample becomes large with increase in collinearity. β_1 maintained a decrease irrespective of the sample size as the collinearity increased.

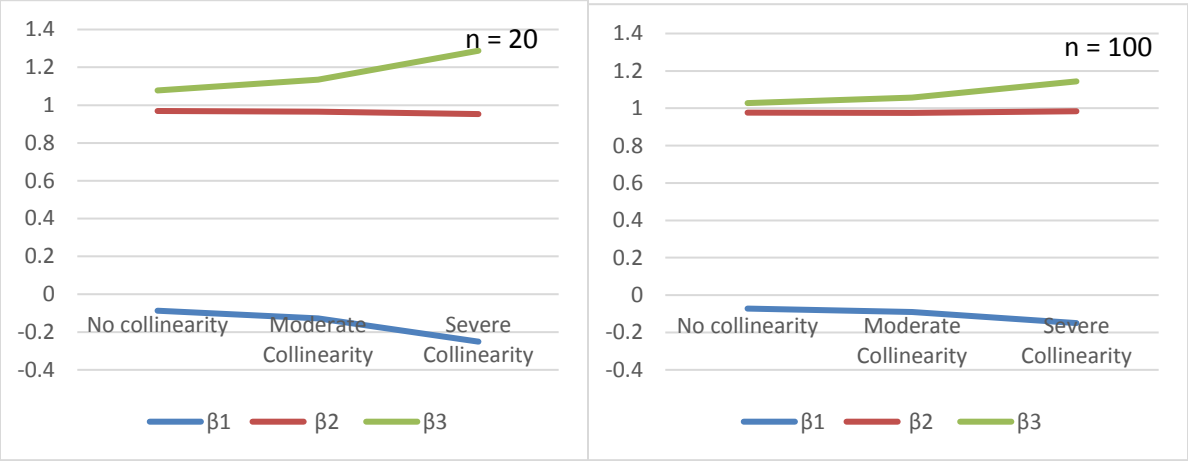


Fig.1: Positive relationship among three predictor variables(Y,X_1, X_2, X_3) with n = 20 and n = 100.

Case 2: Multicollinearity with positive correlation among regressors (X_1, X_2 and X_3) but negative relation with regressand Y.

Considering collinearity scenario with positive correlation among regressors but inverse relationship with the regressand, Table 2 shows that the signs of the coefficients changed when compared with the first scenario. As the degree of collinearity increased, the results showed slight changes in estimates and values of test statistics for all variables when compared to the model in scenario 1. The coefficient β_1 and $\beta_2$here increased in particular for small sample while β_3 decreased which is a reversal of the behavior in case 1 above. Larger changes were observed for the estimates of β_1 and $\beta_2$which altered the statistical significance of the regression coefficient estimates when the correlation was severe judging from the baseline condition of no/low collinearity. However, a large sample size appears to be a good remedy for collinearity with positive correlation among regressors but negative relation with regressand. The result reveals that even with severe collinearity, all the coefficients of the large sample are still significant.

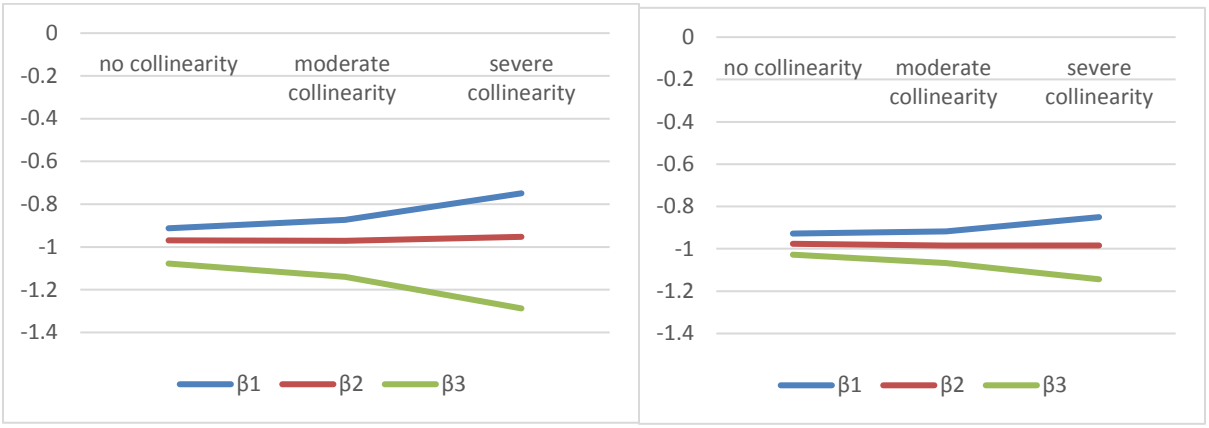
Academic Journal of Statistics and Mathematics (AJSM)
Vol.6, No.4; April - 2020;
ISSN (5730 – 7151);
p –ISSN 4052 – 392X
Impact factor: 3.92

Table 2: Collinearity with positive correlation among three predictor (X_1, X_2, X_3) variables but negatively correlated with Y

Degree of Collinearity	RMSE	R^2	Coefficient	SE coeff.	t-value	Prob	VIF	CI
Low/No Collinearity								
n =20	0.9050	0.9636	B_0 = 0.7483 β_1= -0.9128 β_2 = -0.9693 β_3 = -1.0778	0.4129 0.1616 0.1874 0.1479	1.81 -5.65 -5.17 -7.29	0.0889 0.0000 0.0000 0.0000	- 1.89 1.53 1.88	1.0000 4.0312 4.8567 5.7199
n =100	1.1084	0.9464	B_0 = 0.8580 β_1 = -0.9279 β_2 = -0.9762 β_3 = -1.0277	0.2101 0.0650 0.0688 0.0634	4.08 -14.28 -14.19 -16.22	0.0000 0.0000 0.0000 0.0000	- 1.41 1.26 1.42	1.0000 3.4521 3.9521 4.2255
n =20	0.9037	0.9968	B_0 = 0.7074 β_1= -0.8733 β_2 = -0.9714 β_3 = -1.1401	0.4461 0.2386 0.2354 0.2280	1.59 -3.66 -4.13 -5.00	0.1324 0.0021 0.0008 0.0001	- 31.72 28.92 31.51	1.0000 4.6437 26.8892 28.6550
n =100	1.0989	0.9959	B_0 = 0.7313 β_1 = -0.9183 β_2 = -0.9845 β_3 = -1.0668	0.2137 0.0918 0.0896 0.0908	3.42 -10.0 -10.98 -11.74	0.0009 0.0000 0.0000 0.0000	- 23.27 21.02 22.92	1.0000 4.1896 20.9313 21.8697
n =20	0.9036	0.9990	B_0 = 0.7060 β_1= -0.7492 β_2 = -0.9528 β_3 = -1.2878	0.4445 0.4765 0.4611 0.4614	1.56 -1.57 -2.07 -2.79	0.1318 0.1355 0.0554 0.0131	- 280.3 257.8 271.3	1.0000 4.6216 80.1373 84.4736
n =100	1.0975	0.9982	B_0 = 0.7217 β_1 = -0.8503 β_2 = -0.9840 β_3 = -1.1439	0.2118 0.1834 0.1772 0.1815	3.41 -4.64 -5.55 -6.30	0.0010 0.0000 0.0000 0.0000	- 204.9 187.9 201.0	1.0000 4.1665 62.3840 65.3059

In figure 2, with the change of sign in the dependent variable, all the signs of the coefficients β_1, β_2 and β_3 also changed. As can be observed from the graph, the coefficients in small sample exhibited a significant change as the collinearity becomes severe when compared with the large sample case. In other words, negative relationship of the dependent variable with the explanatory variables affect small sample more than the large sample in a collinearity scenario. The spread of values in β_2 is relatively consistent across the full range of fitted values.

Academic Journal of Statistics and Mathematics (AJSM)
Vol.6, No.4; April - 2020;
ISSN (5730 – 7151);
p –ISSN 4052 – 392X
Impact factor: 3.92



Fig, 2: Negative correlation of the dependent variable with three predictor variables (X_1, X_2, X_3) for n = 20 and n = 100.

Case 3: Positive correlation among the explanatory variables but not correlated with the dependent variable Y.

In this scenario, we consider a case where the explanatory variables are strongly correlated but no./poor correlation with the dependent variable. Our interest here is to examine the contribution of poor correlation between dependent variable and the predictors at various degrees of collinearity among the explanatory variables and the outcome is as seen in the table 3 below. As expected, the coefficient of determination (R^2) is very low since there is a poor relationship between the dependent and explanatory variables. Both estimates β_1 and β_2 have very similar variation in direction but varied in magnitude. The estimates $\beta_2$increased smoothly while β_1 increased significantly with increase in collinearity. The estimate β_1 is 0.0872 and that of β_2 is 0.0307 for small sample when it is uncorrelated, but show a bias increase of β_1 = 0.2508 andβ_2 = 0.0472 for severe collinearity. The insignificance of t-values at low/no collinearity as expected resulted from the fact that the regressand has no correlation with the explanatory variables. When there is poor correlation between regressors and the regressand, the t-values remains insignificant in all cases whether severe, moderate nor low collinearity irrespective of the sample size. The result shows the significance contribution of the regressand to the problem of multicollinearity.

Academic Journal of Statistics and Mathematics (AJSM)
Vol.6, No.4; April - 2020;
ISSN (5730 – 7151);
p –ISSN 4052 – 392X
Impact factor: 3.92

Table 3:**Collnearity among the X_1, X_2, X_3 but poorly correlated with Y.**

Degree of Collinearity	RMSE	R^2	Coefficient	SE coeff.	t-value	Prob	VIF	CI
Low/ No collinearity								
n =20	0.9050	0.0259	B_0 = -0.2517 β_1= 0.0872 β_2 = 0.0307 β_3 = -0.0778	0.4129 0.1616 0.1874 0.1479	-0.61 0.54 0.16 -0.53	0.5507 0.5968 0.8719 0.6062	- 1.90 1.52 1.88	1.0000 4.0312 4.8567 5.7199
n =100	1.1084	0.0175	B_0 = -0.1420 β_1= 0.0721 β_2 = 0.0238 β_3 = -0.0277	0.2101 0.0650 0.0688 0.0634	-0.68 1.11 0.35 -0.44	0.5008 0.2700 0.7304 0.6633	- 1.41 1.26 1.42	1.0000 3.4521 3.9521 4.2255
Moderate collinearity								
n =20	0.9037	0.0283	B_0 = -0.2880 β_1= 0.1283 β_2 = 0.0351 β_3 = -0.1344	0.4454 0.2393 0.2441 0.2246	-0.66 0.54 0.14 -0.60	0.5271 0.5992 0.8874 0.5580	- 8.61 7.66 8.79	1.0000 4.6231 13.5917 14.7712
n =100	1.1009	0.0308	B_0 = -0.2497 β_1= 0.0909 β_2 = 0.0248 β_3 = -0.0575	0.2155 0.0923 0.0913 0.0911	-1.16 0.99 0.27 -0.63	0.2494 0.3268 0.7873 0.5292	- 6.33 5.61 6.28	1.0000 4.1840 10.6060 11.0452
Very strong collinearity								
n =20	0.9036	0.0290	B_0 = -0.2941 β_1 = 0.2508 β_2 = 0.0472 β_3 = -0.2878	0.4445 0.4765 0.4611 0.4614	-0.66 0.53 0.10 -0.62	0.5177 0.6058 0.9197 0.5415	- 280.3 257.8 271.3	1.0000 4.6216 80.1373 84.4736
n =100	1.0975	0.0367	B_0 = -0.2783 β_1 = 0.1499 β_2 = 0.0160 β_3 = -0.1439	0.2118 0.1834 0.1772 0.1815	-1.31 0.82 0.09 -0.70	0.1920 0.4165 0.9282 0.4299	- 204.9 187.9 201.0	1.0000 4.1665 62.3940 65.3059

Similar response was noticed in the graph below for both samples (n =20 and n = 100). With no correlation of the explanatory variables with the dependent variable, β_1 in both samples increased evenly as the collinearity among the explanatory variables increased. For β_2, it was relatively steady while that of β_3 skewed negatively in both small and large samples. Interestingly, both $\beta_1$andβ_3 moved in opposite direction as the degree of collinearity becomes severe.

Academic Journal of Statistics and Mathematics (AJSM)
Vol.6, No.4; April - 2020;
ISSN (5730 – 7151);
p –ISSN 4052 – 392X
Impact factor: 3.92

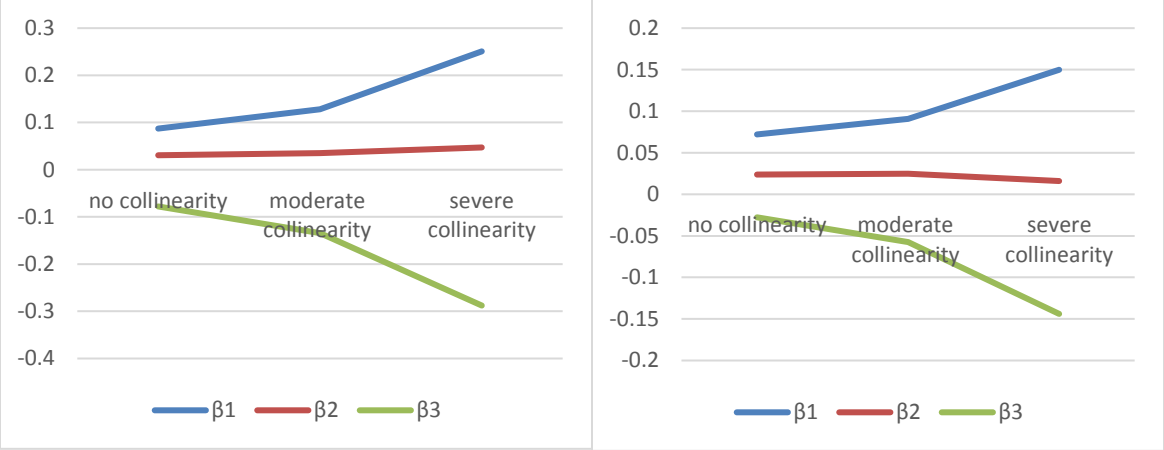


Fig. 3: Correlation among explanatory variables only without the dependent variable for n = 20 and n =100

Case 4: Collinearity with one negative explanatory variable correlated with the other two regressors and regressand

Finally, we also considered where one of the explanatory variables is negatively correlated with the other two variables and the dependent variable in case 4. From the result in table 4, the coefficient of determination increased initially then dropped and increased again as the sample increases. Again, the VIF showed a high degree of collinearity which reduces gradually with increase in sample size. The significance of the coefficients was alsovery high.

Interesting were the results in correlation scenario where one of the explanatory variables is negatively correlated with the other variables. For the small sample, the estimates of β_2 and β_3 increased with increase in collinearity while the coefficient of the variable that is negatively correlated with others, reduced to 0.7492 from 0.9128. Also, for the large sample, the results were similar to the small sample case. β_1 dropped slightly from 0.9279 to 0.8503 as the collinearity becomes severe. Again, the estimates of β_2 and β_3 increased with increase in collinearity while β_1 in small sample case also dropped from 0.9279 to 0.8503.In this scenario, the significance of t-value in both coefficients of X_1 and X_3 is unaffected by the both at moderate and severe collinearity when the sample size is large. However, the only difference was that with the increase of the pairwise correlation between all variables there was larger changes in standard errors, values of test statistics, and switch in the statistical significance to non-significance were observed for sample group (n = 20).

Academic Journal of Statistics and Mathematics (AJSM)
Vol.6, No.4; April - 2020;
ISSN (5730 – 7151);
p –ISSN 4052 – 392X
Impact factor: 3.92

Table 4: Collinearity for X_2 negativelycorrelated with X_1 and X_3 and the regressand Y

Degree of Collinearity	RMSE	R^2	Coefficient	SE coeff.	t-value	Prob	VIF	CI
Low/ No Collinearity								
n =20	0.9050	0.9174	B_0 = 1.2517 β_1= 0.9128 β_2 = 1.0286 β_3 = 1.0778	0.4129 0.1616 0.1874 0.1479	3.03 5.65 5.50 7.29	0.0079 0.0000 0.0000 0.0000	- 1.89 1.53 1.88	1.0000 4.0312 4.8667 5.7199
n =100	1.1084	0.8923	B_0 = 1.1420 β_1 = 0.9279 β_2 = 1.0138 β_3 = 1.0277	0.2101 0.0650 0.0688 0.0634	5.44 14.28 14.88 16.22	0.0000 0.0000 0.0000 0.0000	- 1.41 1.26 1.42	1.0000 3.4521 3.9521 4.2255
Moderate Collinearity								
n =20	0.9037	0.9765	B_0 = 1.2926 β_1= 0.8733 β_2 = 1.0307 β_3 = 1.1401	0.4461 0.2386 0.2354 0.2280	2.90 3.66 4.37 5.00	0.0105 0.0021 0.0005 0.0001	- 31.72 28.92 31.59	1.0000 4.6437 26.8892 28.6550
n =100	1.0988	0.9684	B_0 = 1.2667 β_1 = 0.9183 β_2 = 1.0155 β_3 = 1.0658	0.2137 0.0918 0.0896 0.0908	5.94 10.00 11.33 11.74	0.0000 0.0000 0.0000 0.0000	- 23.27 21.92 22.91	1.0000 4.1896 20.9313 21.8697
Very strong Collinearity								
n =20	0.9036	0.9878	B_0 = 1.2941 β_1= 0.7492 β_2 = 1.0472 β_3 = 1.2878	0.4445 0.4765 0.4611 0.4614	2.91 1.57 2.27 2.79	0.0102 0.1355 0.0373 0.0131	- 280.3 257.8 271.3	1.0000 4.6216 80.1373 84.4736
n =100	1.0975	0.9842	B_0 = 1.2783 β_1 = 0.8503 β_2 = 1.0160 β_3 = 1.1448	0.2118 0.1834 0.1772 0.1815	6.04 4.64 5.73 6.30	0.0000 0.0000 0.0000 0.0000	- 204.9 187.9 201.0	1.0000 4.1665 62.3840 65.3059

Figure shows how the estimates of β_1, β_2, and β_3 vary according to different conditions of two correlation coefficients between positive correlation between X_1 and X_3.As depicted in Figure, β_1 and β_3 show a monotonic relationship. That is, as the value of β_1 increases, the value of β_2 decreases as the collinearity becomes more severe in both samples. However, the rate of divergence is more pronounced with small sample

.

Academic Journal of Statistics and Mathematics (AJSM)
Vol.6, No.4; April - 2020;
ISSN (5730 – 7151);
p –ISSN 4052 – 392X
Impact factor: 3.92

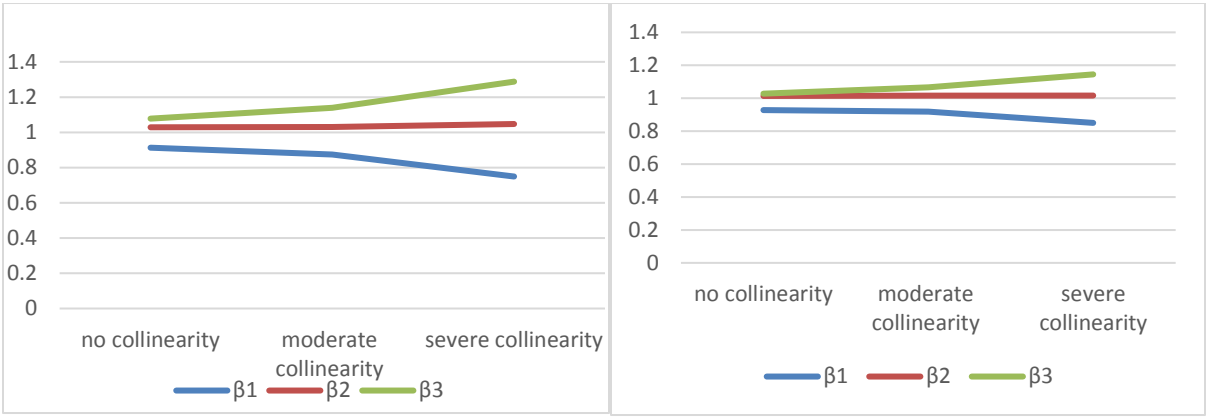


Fig. 4: when one of the explanatory variables is negatively correlated with the other variables for n = 20 and n =100

Conclusion

Collinearity depends on the relationship among the explanatory variables with the dependent variable. This study has exposed some cautionary steps about multicollinearity problem in a linear regression model which has led to conflicting results of multicollinearity in literature. Using simulated data, we illustrated how the regressors at various degrees of multicollinearity influenced the parameter estimates and standard errors in the model. It was observed that small samples are sensitive to negative correlation of any the explanatory variables with other explanatory variables leading to larger changes in standard errors, deviation in the statistical significance to non-significance resulting from changes in the values of test statistics. This enables a clear demonstration of the effect of collinearity on regression and portray some conditions under which the various collinearity diagnostics are informative and some conditions where they are misleading. As with estimation of regression coefficients, the effect of positive and negative collinearity is not symmetric. In other words, they exhibited the non-monotonicity of the relationship. Furthermore, caution should be taken when considering collinearity among explanatory variables that are poorly correlated with the dependent variable as these can lead to a misleading result even when VIF is indicating absence of collinearity (VIF < 2) but the t- values were still insignificant. Importantly, multicollinearity increases the variance of the coefficient estimates but it does not increase the variance of the entire model, which is why it doesn't affect the goodness-of-fit statistics and RMSE. In all, a large sample size appears to be a good remedy for collinearity with positive correlation among regressors that are negatively related with regressand.

Reference

1. Thiart, C. (1990). Collinearity and consequences for estimation: a study and simulation. University of Cape Town.
2. McClelland G.H., Irwin J.R., Disatnik D., & Sivan L (2017). 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 (2016). Behav Res 49:394–402.
3. Mason G. (1987). Coping with multicollinearity. The Canadian Journal of program evaluation. 2:87–93.
4. Tu YK, Clerehugh V, Gilthorpe MS. (2004). Collinearity in linear regression is a serious

problem in oral health research. Eur J Oral Sci.;112:389–397.

5. Mela C.F, Kopalle P. K. (2002). The impact of collinearity on analysis: the asymmetric effect of negative and positive correlations. Applied Economics, 34:667–677.
6. Hoffmann J.P., and Shafer K. (2015). Linear Regression Analysis: Applications and Assumptions. 2nd. NASW Press; Washington, D.C.
7. Tu Y.K., Clerehugh V, Gilthorpe M. S. (2004). Collinearity in linear regression is a serious problem in oral health research. Eur J Oral Sci. 112:389–397.
8. 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.
9. 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, 221-239.
10. Belsley D.A., (1976). Multicollinearity: Diagnosing its presence and assessing the potential damage it causes least square estimation. NBER Working Paper, No. W0154.
11. Stewart GW. (1987). Collinearity and Least Square Regression. Statistical Science, 2(1), 68-94.
12. Cook R.D. (1984). Comment on demeaning conditioning diagnostics through centering. The American Statistician 38: 78–79.
13. Farrar D.E., Glauber R.R., (1967). Multicollinearity 1n regression analysis: The problem revisited. Review of Economics and Statistics, 49, 92- 107.
14. Wittink, D. R. (1988). The Application of Regression Analysis, Simon & Schuster, Needham Heights, Massachusetts.
15. Lehmann D. R., Gupta, S. and Steckel, J. (1988). Marketing Research, Addison-Wesley Educational Publishers, Inc., Reading, Massachussetts.