The Vector Geometric Approach to Multicollinearity Diagnostics

Ijomah Maxwell .A.[1*] Bazuaye Frank .E.[2]

1. Department of Mathematics/Statistics, University of Port Harcourt, Rivers State, Nigeria.
2. Department of Mathematics/Statistics, University of Port Harcourt, Rivers State, Nigeria.

* E-mail of the corresponding author: maxwell.ijomah@uniport.edu.ng

Abstract

The problems of multicollinearity among the independent variables in least-squares regression are by now well-known and published. In the presence of multi-collinearity problem, the parameter estimation method based on the ordinary least squares' procedure is unsatisfactory. Most of the available multicollineraity diagnostic methods may lead to dramatically different conclusions based on their cutoff points and what might be gained from the different alternatives in any specific empirical situation is often unclear due to inadequate knowledge about what degree of collinearity is "harmful". In this paper, we considered the vector geometry approach which is a very useful but are scarcely used tool for illustrating regression analysis to multicollinearity diagnostics. Our result reveals that angles in the range of 19 to 45 degrees are closer to the orthogonality than collinearity Also, the variables are dependent when the vectors are almost parallel while variables are independent, when the vectors are nearly orthogonal. Thus, independent random variables are orthogonal. The paper therefore proposes practical angles and the corresponding correlation coefficients that determine the presence of collinearity in a regression model.

KEY WORDS: Multicollinearity; Vectors; Dimensional Space; Euclidean norm; cosine of angles; correlation coefficient

1. Introduction

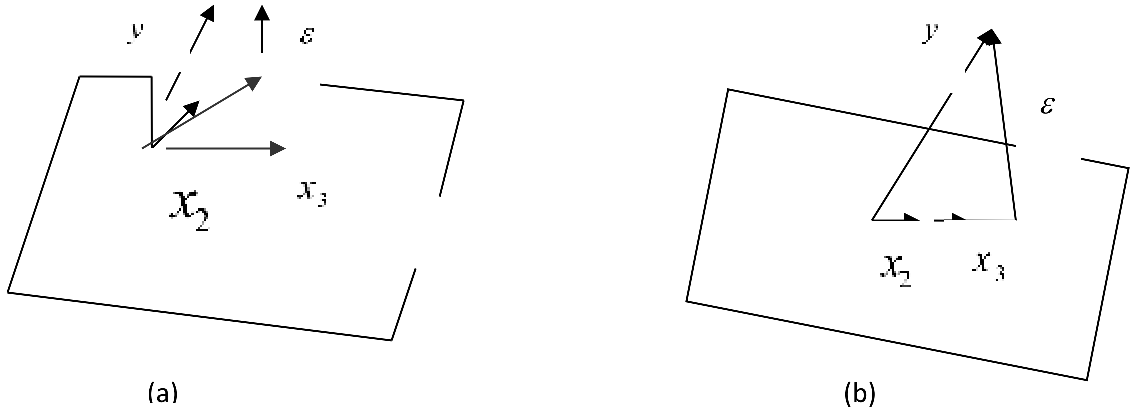
Multicollinearity is a problem with being able to separate the effects of two (or more) variables on an outcome variable. When two X variables are highly correlated, they both convey essentially the same information. In this case, neither may contribute significantly to the model after the other one is included. But together they contribute a lot. If both variables are removed from the model, the fit would be much worse. So the overall model fits the data well, but neither X variable makes a significant contribution when it is added to the model. When this happens, the X variables are collinear and the results show multicollinearity. If two variables are significantly alike, it becomes impossible to determine which of the variables accounts for variance in the dependent variable (Shana. et al.2006).. As a rule of thumb, the problem primarily occurs when x variables are more highly correlated with each other than they are with the dependent variable. It commonly occurs when a large number of independent variables are incorporated in a regression model. This is so because some of them may measure the same concepts or phenomena (Reddy et al. 2003). When a model is not full ranked, that is, the inverse of X cannot be defined, there can be an infinite number of least squares solutions.

The nature of the problem may also be illustrated geometrically as shown in figure below. The x_2, x_3 vectors are not perfectly collinear and they span a two-dimensional subspace in $\Re^n$. Dropping a perpendicular from y to that subspace slits y into

$$\boldsymbol{Y}=\hat{\boldsymbol{Y}}+\boldsymbol{\varepsilon} \qquad (1)$$

where

$$\hat{Y}=X\beta=\hat{\beta}_2x_2+\hat{\beta}_3x_3$$



(a) (b)

The regression vector $\hat{y}$ is a unique linear combination of the column vectors x_2, x_3 only span a one-dimensional subspace (line) in $\Re^n$. The $\hat{y}$ vector is still unambiguously determined by dropping a perpendicular from y to the line, but $\hat{y}$ cannot be expressed uniquely in terms of x_2 and x_3. Mathematically, the problem is that the X matrix is not full rank. When this occurs, the X matrix (and hence the $X'X$ matrix) has determinant zero and cannot be inverted. Given a general 3x3 matrix A as shown below

$$A=\begin{bmatrix}\alpha & \beta & \beta\\ \phi & \gamma & \gamma\\ \theta & \omega & \omega\end{bmatrix}$$

The determinant of this matrix is $\alpha\gamma\omega-\alpha\gamma\omega-\beta\phi\omega+\beta\gamma\theta+\alpha\gamma\omega+\beta\phi\omega-\beta\gamma\theta=0$

Recall from notes on matrix algebra that the inverse can be found using the determinant function: However, when $\det(A)$, all of the elements of the inverse are clearly undefined. This is a case of perfect or exact collinearity. $A^{-1}=\dfrac{1}{\det(A)}adj(A)$

Now consider the following multiple regression models

$$y=X\beta+\varepsilon \qquad (2)$$

where y is an n×1 vector of responses, X is an n×p matrix of the regressor variables, **β** is a p × 1 vector of unknown constants, and ε is an n ×1 vector of random errors, with $\varepsilon_I \sim IIDN(0,\sigma^2)$. It will be convenient to assume that the regressor variables are standardized. Consequently, $X'X$ is a p× p matrix of correlations between the regressors and $X'Y$ is a p × 1 vector of correlation between the regressors and the response. Let the j^{th} column of X matrix be denoted by X_j , so that $X = \lfloor X_1, X_2, ..., X_p \rfloor$. Thus X_j contains the n levels of the regressor variable. Formally multicollinearity can be defined as the linear dependence of the columns of X . The vectors are linearly dependent if there is a set of constants $b_1, b_2, ...b_p$, not all zero such that

$$\sum_{j=1}^{p} b_j X_j = 0 \qquad (3)$$

If Equation (3) holds exactly for a subset of the columns of **X**, then the rank of the $X'X$ matrix is less than p and $(X'X)^{-1}$ does not exist. However, suppose the Equation (3) is approximately true for some subset of the columns of X . Then there will be a near linear dependency in $X'X$ and the problem of multicollinearity is said to exist. It is to be noted that the multicollinearity is a form of ill-conditioning in the $X'X$ matrix. Furthermore, the problem is one of the degrees, that is, every data set will suffer from multicollinearity to some extent unless the columns of X are orthogonal (Jahufer, 2015). Various econometric references have indicated that collinearity increases estimates of parameter variance, yields high R^2 in the face of low parameter significance, and results in parameters with incorrect signs and implausible magnitudes (Besley, et al. 1980; Greene, 1990; & Kmenta, 1986). The presence of multicollinearity can make the usual least-squares analysis of the regression model dramatically inadequate. In some cases, multiple regression results may seem paradoxical. Even though the overall p-value is very low, all of the individual p-values are high. This means that the model fits the data well, even though none of the X variables has a statistically significant impact on predicting Y.

2. Matrix-Geometric Approach on Multicollinearity

Multicollinearity means that there exists (at least) one set of constants $b_0, b_1, ..., b_1, ..., b_p$ not all zero, such that

$$b_1 X_1 + b_2 X_2 + ... + b_p X_P = \sum_{i=1}^{p} b_i X_i$$

Introduce the p × 1 matrix **b**, so we can write $b = \begin{bmatrix} b_1 \\ . \\ . \\ b_p \end{bmatrix}$ so that we can write multicollinearity as

$$b^T X = b_0 \qquad \text{for } b \neq 0$$

If this equation holds, then

$$Var[b^T X] = Var[\sum_{i=1}^{p} b_i X_i] = Var[b_0] = 0$$

Conversely, if $Var[b^T X] = 0$, then $b^T X$ must be equal to some constant, which we can call b_0.

Therefore multicollinearity is equivalent to the existence of a vector $b^T X$ where

$$Var[b^T X] = 0$$

Recall from ( ) $Var[b^T X] = Var\left[\sum_{i=1}^{p} b_i X_i\right]$

$$= \sum_{i=1}^{p}\sum_{j=1}^{p} b_i b_j Cov\left(X_i X_j\right) \qquad (4)$$

$$= b^T Var[X]b \qquad (5)$$

Multicollinearity therefore means the equation $b^T Var[X]b = 0$ has a solution $b \neq 0$.

The presence of multicollinearity has a number of potentially serious effects on the least-squares estimates of the regression coefficients (Joshi & Deshpande, 2012). Some of these effects may be easily demonstrated. Suppose that there are only two regressor variables x_1 and x_2. The model, assuming that x_1, x_2 , and y are scaled unit length,

$$y = \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

and the least-squares normal equations are

$$(X'X)\hat{\beta} = X'y$$

$$\begin{bmatrix} 1 & r_{12} \\ r_{12} & 1 \end{bmatrix}\begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} r_{1y} \\ r_{2y} \end{bmatrix}$$

where r_{12} is the simple correlation between x_1 and x_2 and r_{jy} is the simple correlation between x_j and y , j=1,2. Now the inverse of $(X'X)$

$$C = (X'X)^{-1} = \begin{bmatrix} \dfrac{1}{(1-r_{12})^2} & \dfrac{-r_{12}}{(1-r_{12}{}^2)} \\ \dfrac{-r_{12}}{(1-r_{12}{}^2)} & \dfrac{1}{(1-r_{12})^2} \end{bmatrix} \qquad (6)$$

and the estimates of the regression coefficients are

$$\hat{\beta}_1 = \frac{r_{1y} - r_{12}r_{2y}}{(1-r_{12}{}^2)}, \qquad \hat{\beta}_2 = \frac{r_{2y} - r_{12}r_{1y}}{(1-r_{12}{}^2)} \qquad (7)$$

If there is strong multicollinearity between x_1 and x_2, then the correlation coefficient r_{12} will be large. From Equation (6) we see that as $r_{12} \to 1$, $Var(\hat{\beta}_j) = C_{jj}\sigma^2 \to \infty$ and $Cov(\hat{\beta}_1\hat{\beta}_2) = C_{12}\sigma^2 \to \pm\infty$ depending on whether $r_{12} \to +1$ or $r_{12} \to -1$. Therefore, strong multicollinearity between and result in large variances and covariances for the least-squares estimators of the regression coefficients. When there are more than two regressor variables, multicollinearity produces similar effects.

3. Multicollinearity Diagnostics

To assess whether collinearity is indeed problematic, various diagnostics are frequently employed. If passed, then the results are assumed to be free of `problems.' In a multivariate situation, the literature provides numerous suggestions, ranging from simple rules of thumb to complex indices, for diagnosing the presence of substantive collinearity. Farrar & Glauber (1967), also proposed a procedure for detecting multicollinearity which comprised of three tests (i.e. Chi-square test, F-test and T-test). However, these tests have been greatly criticized. Wichers (1975) claims that the third test, where the authors use the partial-correlation coefficients is ineffective. The use of condition index for collinearity diagnostics alone is not enough. Some other insight is necessary to find which columns of X are involved in the collinearity. Johnston, (1984). insist 10 to 100 as a beginning and serious points that collinearity affects estimates. Condition number suffer from the scaling problems which makes comparison difficult. Besley et al. (1980); Johnston, (1984); Anderson, & Wells, (2008) suggest that condition indices in excess of 20 are problematic. However, this diagnostic does not consider correlations with the dependent variable, which have already been shown to moderate the effects of collinearity. Belsley (1991) has persuasively argued in the general regression context that diagnosing collinearity should be done with a combination of conditioning indices of the data matrix and the variance-decomposition proportions. However, there is no obvious value of the condition index that defines the boundary between degrading and truly harmful collinearity. Belsley (1991) developed a universal procedure to formally test whether there is inadequate signal-to-noise in the data, but this is not easily implemented and therefore does not appear to be used by researchers. Huang, (1970) observed that multicollinearity is said to be "harmful" if $r_{ij} \geq R^2$. Such simple correlation coefficients are sufficient but not necessary condition for multicollinearity. In many cases there are linear dependencies, which involve more than two explanatory variables, that this method cannot detect it. Judge et al. (1985) & Belsley, (1990) pointed out that using correlation matrix is unable to reveal the presence or number of several coexisting collinear relations (Belsley,1990). Furthermore, in using variance inflation factor (VIF) some authors have stated that multicollinearity is problematic if largest VIF exceeds value of 10, or if the mean VIF is much greater than 1. Although VIF greater than 5 or VIF greater than 10 are suggested for detecting multicollinearity, there is no universal agreement as what the cut-off based on values of VIF should be used to detect multicollinearity (Kutner, Nachtsheim & Neter 2004). Caution for misdiagnosis of multicollinearity using low pairwise correlation and low VIF was reported in the literature for collinearity diagnostic as well. O'brien, (2007) demonstrated that VIF rules of thumb should be interpreted with cautions and should be put in context of the effects of other factors that influence the stability of the specific regression coefficient estimate and suggested that any VIF cut-off value should be based on practical consideration. Freund & Wilson (1998), further suggested VIF to be evaluated against the overall fit of the model, using the model R^2 statistics. VIF >1/(1-overall model R^2) indicates that correlation between the predictors is stronger than the regression relationship and multicollinearity can affect their coefficient estimates, while Hair et al.(1995) suggest variance inflation factors (VIF) less than 10 are indicative of inconsequential collinearity. However, low correlations do not automatically imply low collinearity (Belsley,1991). It is possible to have low bivariate correlations between variables in a highly collinear model. As such, correlation-based collinearity metrics such as variance inflation factors (VIFs) are likely to misdiagnose collinearity problems. There is therefore no formal criteria for determining the magnitude of variance inflation factors that cause poorly estimated coefficients. The decision to consider a VIF to be large was essentially arbitrary. Reviewing the literature on ways to diagnosing collinearity reveals several points. First, a variety of alternatives are available and may lead to dramatically different conclusions based on their cutoff points. Second, what might be gained from the different alternatives in any specific empirical situation is often unclear. Part of this ambiguity is likely to be due to inadequate knowledge about what degree of collinearity is "harmful" (Mason & Perreault 1991). In much of the empirical research on collinearity diagnostics, data with extreme levels of collinearity are used to provide rigorous tests of the approach being proposed. Such extreme collinearity is rarely found in actual cross-sectional data.

4. Materials and Methods

Let the explanatory variables, $X(n,m)$, in the model $y = X\beta + \varepsilon$ as in Equation (2) be measured such that each of its columns has a zero mean and unit standard deviation. In that case, $X'X = nR$ where R is the intercorrelation matrix and r_{ij} is the cosine of the angle between x_i and x_j vectors. Ideally, the $X'X$ matrix should be diagonal. That signifies a total absence of multicollinearity. However, this is far from the real-world situation. Since the cosine of an angle must lie between –1 and 1, multicollinearity gets to its highest degree when any one or more off-diagonal element(s) of R is (are) ± 1. The predictors $X_1, X_{2,}......, X_p$ form a p-dimensional random vector X. Ordinarily, we expect this random vector to be scattered throughout p-dimensional space. When we have collinearity (or multicollinearity), the vectors are actually confined to a lower-dimensional subspace (Imdadullah,, Aslam, & Altaf 2016). The column rank of a matrix is the number of linearly independent columns it has. If x has column rank q < p, then the data vectors are confined to a q-dimensional subspace. It looks like we've got p different variables, but really by a change of coordinates we could get away with just q of them. The space with Euclidean norm and scalar product is considered. In n-dimensional space vector $x = (x_1, x_2, ..., x_n)$ is considered. The Euclidean norm $\|x\|$ of this vector is given by the formula [1]:

$$\|x\| = \sqrt{\sum_{i=1}^{n} x^2} \qquad (8)$$

In three-dimensional space or on a plane, Euclidean norm of vector is its length. The scalar product (dot product) of vector $x = [x_1, x_2, ..., x_n]$ and vector $y = [y_1, y_2, ..., y_n]$ is equal to (1):

$$x.y = \sum_{i=1}^{n} x_i y_i \qquad (9)$$

Simultaneously, the dot product of two vectors can be represented as follows:

$$x.y = \|x\|.\|y\|.\cos(x, y) \qquad (10)$$

In the expression (10) $\cos(x, y)$ is the cosine of the angle between two vectors:

$$\cos(x, y) = \frac{x.y}{\|x\|.\|y\|} \tag{11}$$

Hence, the angle between vectors can be calculated using the arccosine function.

Now if we consider a measure of the relationship between two random variables x and y which is the covariance

$$\sigma_{XY} = E(X - \mu_X)(Y - \mu_Y) \tag{12}$$

and the covariance normalized to unity is called the correlation coefficient

$$\rho_{xy} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{E(X - \mu_X)(Y - \mu_Y)}{\sqrt{E[(X - \mu_X)^2]E[(Y - \mu_Y)^2]}} \tag{13}$$

Expression (13) can be further converted to the form:

$$\rho_{xy} = \frac{\sum_{i=1}^{n}[(X_i - \bar{x})(Y_i - \bar{y})]}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{x})^2 \sqrt{\sum_{i=1}^{n}(Y_i - \bar{y})^2}}} \tag{14}$$

The correlation coefficient between two variables is equal to the covariance of variables subject to standardization. Setting equation (14) can be converted to the form:

$$\rho_{xy} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x^2 \sqrt{\sum_{i=1}^{n} y^2}}} \tag{15}$$

The resulting expression is the ratio of two elements. The numerator is the scalar product of two vectors, while the denominator is the product of its lengths:

$$\rho_{xy} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x^2 \sqrt{\sum_{i=1}^{n} y^2}}} = \frac{x.y}{\|x\|.\|x\|} = \cos(x, y) \tag{16}$$

Expression (16) shows the formal identity between the correlation coefficient, and the cosine of the angle between two random vectors.

Consider the angle between the column vectors of X given as

$$\cos(\theta) = \frac{x_k^T x_{k+1}}{\|x_k\|_2 \|x_{k+1}\|_2}$$

where $k(0 \le k < k+1 \le m)$ is a column index. Note that when the angle between the column vectors is $\pi/2$, the vector are orthogonal and when the angle is 0, the vectors are exactly collinear.

5. Numerical Application

Table 1 presents the cosines of different angles (different correlation coefficients) and the corresponding coefficients of determination expressed as a percentage. Two random vectors are (almost) orthogonal, if the cosine of the angle between them (also determination coefficient) is (almost) equal to zero. This means that the random variables represented by these vectors are independent or random vectors are (near) orthogonal. To compare the vector geometric approach with the existing variance inflation factor for dealing with collinearity, we considered angles from 0 to 45 and 90 data sets with the (collinear) predictors X_1X_2. We then explored the predictive performance of the methods on test data sets with five different collinearity structures. The result is as shown in table 1. When the angle between X_1 and X_2 is 0 (i.e. case of perfect correlation), the correlation between X_1 and X_2 is 1 and when it is 90, which means that the correlation between these variables is zero. Hence the importance of X_2 in explaining R^2 is zero according to the product measure. Therefore, X_1 is responsible for all the length of X_2. However, if we exclude X_2, X_1 would not ex plain much at all of the variation in y. Hence with the product measure, x2's contribution to R2 is zero and x1's contribution to R^2 equals R^2. Following the general rule is that the VIF should not exceed 10, Belsley, Kuh, & Welsch, [10], severe collinearity occurs when the angles are between 1 to 18 degrees which accounted for over 90 percent of the variance**.**

Table 1: The cosine of the angle against determination coefficient

Angle (degrees)	**Angle (rad)**	**The cosine of the angle**	**R^2**	**Explained % of the variance**	**Coefficient of Alienation**	**VIF**
0	0.0000	1.0000	1.0000	100	0.0000	-
1	0.0175	0.9998	0.9996	99.96	0.0004	2500.25
2	0.0349	0.9994	0.9988	99.88	0.0012	833.58
3	0.0524	0.9986	0.9972	99.72	0.0028	357.39
4	0.0698	0.9976	0.9952	99.52	0.0048	208.59
5	0.0873	0.9962	0.9924	99.24	0.0076	131.83
6	0.1047	0.9945	0.9890	98.90	0.0110	91.16
7	0.1222	0.9925	0.9851	98.51	0.0149	66.92
8	0.1396	0.9903	0.9807	98.07	0.0193	51.80
9	0.1571	0.9877	0.9756	97.56	0.0244	40.90
10	0.1745	0.9848	0.9698	96.98	0.0302	33.15
11	0.1920	0.9816	0.9635	96.35	0.0365	27.43
12	0.2094	0.9781	0.9567	95.67	0.0433	23.08
13	0.2269	0.9744	0.9495	94.95	0.0505	19.78
14	0.2443	0.9703	0.9415	94.15	0.0585	17.09
15	0.2618	0.9659	0.9330	93.30	0.0670	14.92
16	0.2793	0.9613	0.9241	92.41	0.0759	13.17
17	0.2967	0.9563	0.9145	91.45	0.0855	11.70
18	0.3142	0.9511	0.9046	90.46	0.0954	10.48
19	0.3316	0.9455	0.8940	89.40	0.1060	9.43
20	0.3491	0.9397	0.8830	88.30	0.1170	8.55
21	0.3665	0.9336	0.8716	87.16	0.1284	7.79
22	0.3840	0.9272	0.8597	85.97	0.1403	7.13
23	0.4014	0.9205	0.8473	84.73	0.1527	6.55
24	0.4189	0.9135	0.8345	83.45	0.1655	6.04
25	0.4363	0.9063	0.8214	82.14	0.1786	5.60

26	0.4538	0.8988	0.8078	80.78	0.1922	5.20
27	0.4712	0.8910	0.7939	79.39	0.2061	4.85
28	0.4887	0.8829	0.7795	77.95	0.2205	4.54
29	0.5061	0.8746	0.7649	76.49	0.2351	4.25
30	0.5238	0.8660	0.7500	75.00	0.2500	3.99
31	0.5411	0.8572	0.7348	73.48	0.2652	3.77
32	0.5585	0.8480	0.7191	71.91	0.2809	3.56
33	0.5760	0.8387	0.7034	70.34	0.2966	3.37
34	0.5934	0.8290	0.6872	68.72	0.3128	3.20
35	0.6109	0.8192	0.6711	67.11	0.3289	3.04
36	0.6283	0.8090	0.6545	65.45	0.3455	2.89
37	0.6458	0.7986	0.6378	63.78	0.3622	2.76
38	0.6632	0.7880	0.6209	62.09	0.3791	2.64
39	0.6807	0.7771	0.6039	60.39	0.3961	2.52
40	0.6981	0.7660	0.5868	58.68	0.4132	2.42
41	0.7156	0.7547	0.5696	56.96	0.4304	2.32
42	0.7330	0.7431	0.5522	55.22	0.4478	2.23
43	0.7505	0.7314	0.5350	53.49	0.4651	2.15
44	0.7679	0.7193	0.5174	51.74	0.4826	2.07
45	0.7854	0.7071	0.4999	50.00	0.5000	1.99
90	1.50708	0.0000	0	0.00		1.0000

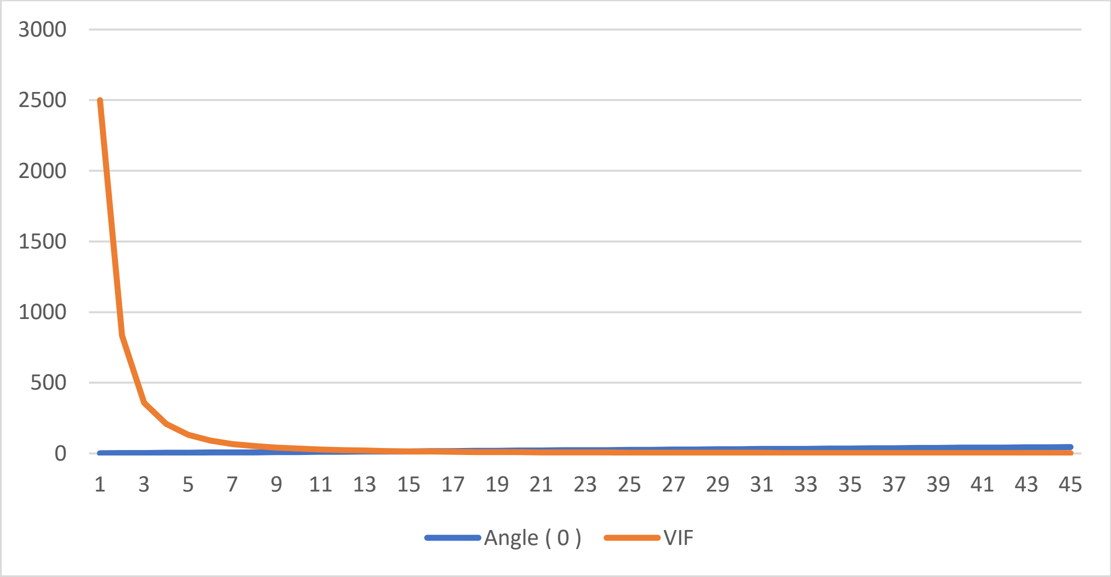


Figure 1: Variance inflation factor (VIF) with the angles

Figure 1 shows the directions of variance inflation factor (VIF) with the angles. The graph shows that the VIF dropped significantly between 1 and 3 degrees of the angle. It maintained a slight difference as the angle progresses. Between 23 degrees to the end of the process, the VIF was lower to the angle. Importantly, after 19 degrees the VIF showed a lower trend than the angle which indicates absence of collinearity following [10],

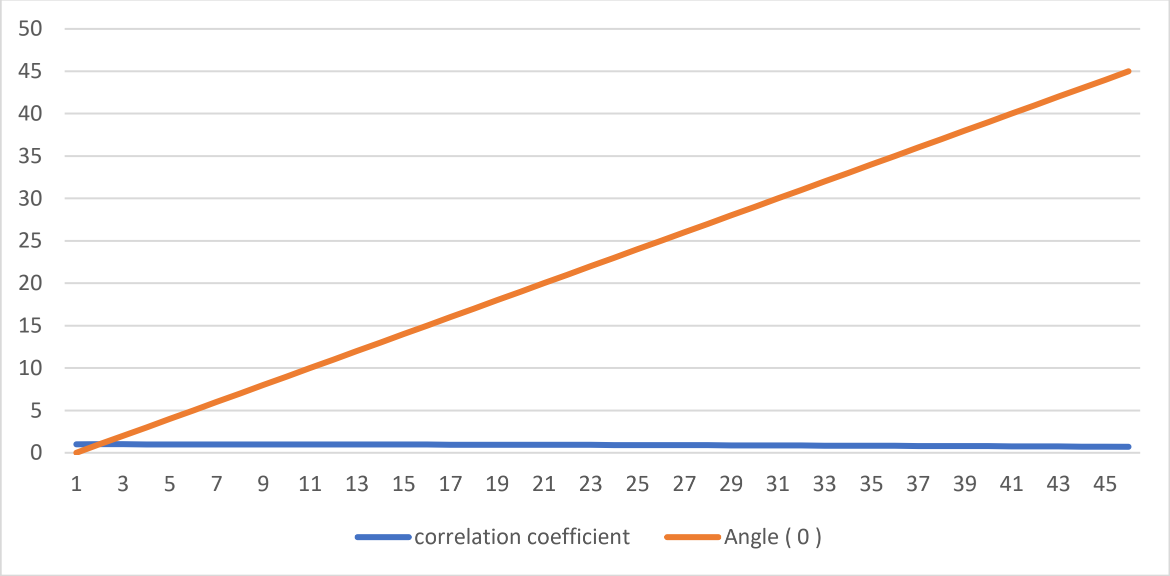


Figure 2. Angles with corresponding correlation coefficient (cosine of the angles)

The above graph of angles with corresponding correlation coefficient, indicates that there is a strong positive relationship between angles and correlation coefficient. The graph shows that the angle between the regression lines is a function of the coefficient of correlation.

Table 2: Intensity of Collinearity based cosine of angles, coefficient of determination and VIF

Level of Collinearity	**Angle (degrees)**	**The cosine of the angle**	**R^2 (%)**	**VIF**
Severe	1 – 9	≥ 0.9870	≥ 97	≥ 40.0
Moderate	10 – 26	0.8988- 0.9046	80.0 – 96.0	5.0 -39.0
Low	27 – 89	0.1000 - 0.8889	10 – 79.99	1.0 – 4.90.
No collinearity	90, 270	0	0	< 1.0

Table 2 gives an overview of the various classification of collinearity based on the VIF, cosine of angle and coefficient of determination. From the table angles between 1 and 9 degrees are indication of severe collinearity as a resulting from very strong correlation between variables, between 10 and 26 are moderate collinearity while between 27 and 89 degrees, vectors are closer to perpendicular and so show low collinearity. Finally, the angles 90 and 270 revealed no collinearity at all. It is clear that at angles 90 and 270 degrees, the correlation between these variables is zero which is an indication of orthogonality.

6. Discussion and Conclusion

We have illustrated how the value of the correlation coefficient, treated as the cosine of the angle between random vectors, contains information about the level of dependence of the variables. The cosine close to zero means that the vectors are (almost) orthogonal, so the random variables are independent. If the cosine is close to one or minus one, the vectors are (almost) parallel and random variables are strongly correlated. In the range of 19 to 45 degrees, the vectors are closer to the orthogonality than collinearity. The angles of 45 degrees and above are the limit angles.

For these angles, the vector is equally far from the orthogonality and parallelism. The coefficient of determination is equal to 50%. Exactly half of the variation in one variable can be explained by the second variable. For angles less than 45 degrees or greater than 135 degrees, vectors are closer to parallel than perpendicular – it can be assumed that the random variables are dependent. Vectors are close to parallel when they lie at an angle less than 30 degrees and greater than 150 degrees, with respect to the reference vector. If the angle is lesser than 15 degrees or greater than 165 degrees, variables are strongly correlated. In summary, the paper presents the possibility of geometrical interpretation of the collinearity. It is noted that the correlation coefficient is formally equivalent to the cosine of angle between random vectors. The variables are dependent when the vectors are almost parallel. The variables are independent, when the vectors are nearly orthogonal. Thus, independent random variables are orthogonal. The paper proposes practical angles and the corresponding correlation coefficients that determine the presence of collinearity in a regression model.

References

Anderson, W., & M. T. Wells (2008) "Numerical Analysis in Least Squares Regression with an Application to the Abortion-Crime Debate," 5 J of Empiical Legal Studies 647.

Besley, D. A., Kuh, E., Welsch, R. E. (1980). Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. Wiley New York.

Belsley,,D..A.,(1990), Conditioning Diagnostics: Collinearity and Weak Data in Regression, John Wiley, New York

Belsley, D.A. (1991). A guide to using the collinearity diagnostics. Computer Science in Economics and Management, 4, 33-50.

Farrar & Glauber (1967), "Multicollinearity in regression analysis" review of economics and statistics, 49, pp. 92-107 [6] Goldberger, A. S. (1991), A Course in Econometrics. Cambridge, MA: Harvard University Press.

Freund, R.J. & Wilson, W.J., (1998). Regression Analysis – Statistical modelling of a response variable. Academic Press, London. 444 pp.

Greene, W. H. (1990). Econometric Analysis, Macmillan Publishing Company, New York.

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

Huang, T. (1970). New Foraminiferida from the Taiwan Proceedings of the Geological Society of China. 13 : 108-114.

Jahufer, A., (2015). Effect of multicollinearity in unbiased regression models. Proceedings of the 1st International Symposium 2011 on Post-War Economic Development through Science, Technology and Management 177.

Johnston, J. (1984). Econometric Methods. New York: McGraw-Hill.

Joshi H., Kulkarni, H., and Deshpande, S., (2012) Multicollinearity Diagnostics in Statistical Modeling & Remedies to deal with it using SAS. PhUSE 2012

Judge, George G., Griffiths, William E., Hill, Robin C., Lutkepohl, Helmut & Lee, T.C., (1985), The Theory and Practice of Econometrics, New York: Wiley.

Kmenta, J. (1986). Elements of Econometrics, 2nd edn, Macmillan Publishing Company, New York.

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

Imdadullah, M., Aslam, M., and Altaf S., (2016) mctest: An R Package for Detection of Collinearity among Regressors. The R Journal Vol. 8 (2.)

Mason, C.H. & W.D. Perreault Jr.,(1991). Collinearity, power and interpretation of multiple regression analysis. J. Market. Res., 28: 268-280.

O'brien, R.M. (2007). A Caution Regarding Rules of Thumb for Variance Inflation Factors. Qual Quant 41, 673–690.

Reddy M.C., P.Balasubramanyam and Subbarayudu M. (2003), (An Effective Approach to Resolve Multicollinearity in Agriculture Data. International Journal of Research in Electronics and Computer Engineering.

Shana, Y. et al. (2006). Machine learning of poorly predictable ecological data. Ecol. Model. 195: 129– 1 38.

Wichers. R.C. (1975). The Detection of Multicollinearity: A Comment. The Review of Economics and Statistics, vol. 57, issue 3, 366-68.