

# Measuring the Impact of Collinearity in Epidemiological Research

Andrew Woolston<sup>1,2</sup>, Yu-Kang Tu<sup>3</sup>, Mark S. Gilthorpe<sup>1</sup> & Paul D. Baxter<sup>1</sup>

<sup>1</sup> Division of Biostatistics, Leeds Institute of Genetics, Health & Therapeutics, University of Leeds, Leeds, UK

<sup>2</sup> Institute of Statistical Science, Academia Sinica, Taipei, Taiwan

<sup>3</sup> Institute of Epidemiology & Preventive Medicine, College of Public Health, National Taiwan University, Taipei, Taiwan

Correspondence: Andrew Woolston, Institute of Statistical Science, Academia Sinica, 128 Academia Road, Section 2, Nankang, Taipei 11529, Taiwan. Tel: 886-2-2783-5611 Ext.182. E-mail: awoolston@webmail.stat.sinica.edu.tw

Received: December 11, 2012 Accepted: January 28, 2013 Online Published: February 4, 2013

doi:10.5539/ijsp.v2n2p1

URL: <http://dx.doi.org/10.5539/ijsp.v2n2p1>

*AW is funded by a UK Medical Research Council PhD studentship award. MSG (E-mail: M.S.Gilthorpe@leeds.ac.uk) and PDB (E-mail: P.D.Baxter@leeds.ac.uk) are funded by the United Kingdom government Higher Education Funding Council for England (HEFCE)*

## Abstract

Collinearity amongst covariates in linear regression models has long been recognised as a potential source of bias. Various ‘solutions’ have been proposed, though one issue almost entirely omitted in the current literature is the importance of the relationship between the outcome and the correlated covariates. Using vector geometry, it can be shown that the impact of collinearity on the model, such as changes in regression coefficients, cannot be judged by the correlation structure of the covariates alone—their relationship with the outcome is crucial. Traditional diagnostics of collinearity are thus insufficient in evaluating adverse effects or model instability. Collinearity diagnostics should play an important role in assessing this impact, both adverse and beneficial, on model parameters. The objective of this study was to build a new index that measures the impact of collinearity in the model environment, rather than providing only a description of the feature. Vector geometry was used to design a measure that accounts for the relationship between the outcome and the correlated covariates—labelled the D-index. The D-index was implemented as part of a regression study to develop a parsimonious model for body fat using easily obtainable body circumference measurements. The covariates were selected based on the degree of collinearity amongst the predictors in the model and the variance explained in the response. Such a model would potentially allow for a reduction in the number of body size measurements required, reducing study length and cost, whilst maintaining measurements that most accurately represent total body fat.

**Keywords:** collinearity, diagnostic, index, vector geometry, variance inflation factor

## 1. Introduction

In epidemiological and clinical research, it is not surprising to find that many covariates are correlated as they often share common physiological mechanisms, or measure different aspects of the same underlying mechanism. The question is not whether collinearity is an issue, but what the impact is on the modelling process. The least squares assumption that covariates are independent implies that all pair-wise covariate associations should be negligible—a most unlikely scenario for biological and epidemiological data. Small, but significant, departures from the assumption of independence can severely distort the interpretation of a model and the role of each covariate, causing increased inaccuracy as expressed through bias within regression coefficients and increased uncertainty as expressed through coefficient standard errors.

The variance inflation factor (VIF) (Marquardt, 1970; Stine, 1995) and condition index (CI) (Belsley, Kuh, & Welsch, 1980) are often labelled collinearity ‘diagnostic’ tools, however this description is perhaps misguided. Collinearity itself is not a ‘disease’. Symptoms such as a change of sign or an adverse change in the variance and point estimates may be considered ‘problematic’. However, they are only problematic based on *prior* biological knowledge. In some circumstances, such as confounding, including a collinear variable in the model may be

beneficial to increasing the precision and accuracy of the assessment of a cause-effect relationship. These statistical measures are not ‘diagnosing’ a disease, but instead providing a description of a feature of the data. This description, along with external biological knowledge, should facilitate the process of deciding whether problematic collinearity exists in the data and whether any remedial action is necessary.

Collinearity indices such as the VIF and CI belong to a class of ‘correlation based’ diagnostics as the assessment rests entirely on the  $X^T X$  matrix (i.e. the matrix of sums of squares and cross products of all predictors). The VIF is calculated as follows,

$$VIF = \frac{1}{1 - R_{x_j}^2} \quad (1)$$

where  $R_{x_j}^2$  is the explained variance of the variable  $x_j$  regressed on the remaining predictors included in the model. Regardless of the chosen response entered into the model, the assessment of collinearity from a correlation based index such as the VIF will not change. The measure is providing a description of the collinearity present amongst the predictors only. This result may be of limited use in application. The researcher will hold an interest in understanding the potential *impact* of collinearity on the parameter estimates from the model and subsequently a potential impact on clinical and biological interpretation of the estimates. An arbitrary ‘rule of thumb’ will often be employed to indicate *serious* collinearity in a dataset. For instance, VIF’s ranging from 4 to 30 have been previously used as an indication that *severe* collinearity is present in the data (O’Brien, 2007). This may encourage the use of remedial action (such as the removal of collinear variables, entering linear combinations as a single predictor or employing alternative, often more complex, methodology) to relieve or resolve the ‘problems’ of collinearity. If other factors had been accounted for in the initial assessment of the data, the need for such action may be much less than first thought.

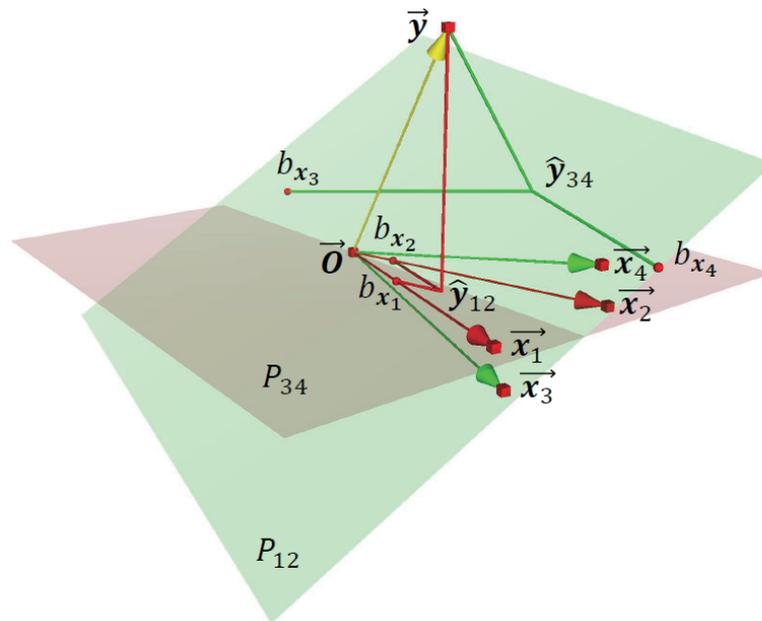


Figure 1. An illustration of the role of the response dictating the impact of collinearity

The impact of collinearity on parameter estimates is governed by factors such as the response, the sample size and sampling variation. These are all features of the ‘model environment’. In Figure 1 there is a rotation of the regression planes (labelled  $P_{12}$  and  $P_{34}$ ) that are spanned by the green and red pairs of predictors respectively (see Wickens (1995) for a description of the vector geometry and Draper and Smith (1998) for matrix approaches to regression analysis). This movement represents a change in the position of the response (e.g. a result of sampling variation) as the predictors are assumed to be measured without error (Freund & Wilson, 1998). When the response is closer to the regression plane in the green example (reflected by an increased coefficient of determination- $R_y^2$ ), a change in the slope of the plane will conceptually have less impact on the deviation of the coefficient estimates. Further to that, an increased correlation between the covariates (i.e. an increase in  $r_{12}$ , reflected by a reduced

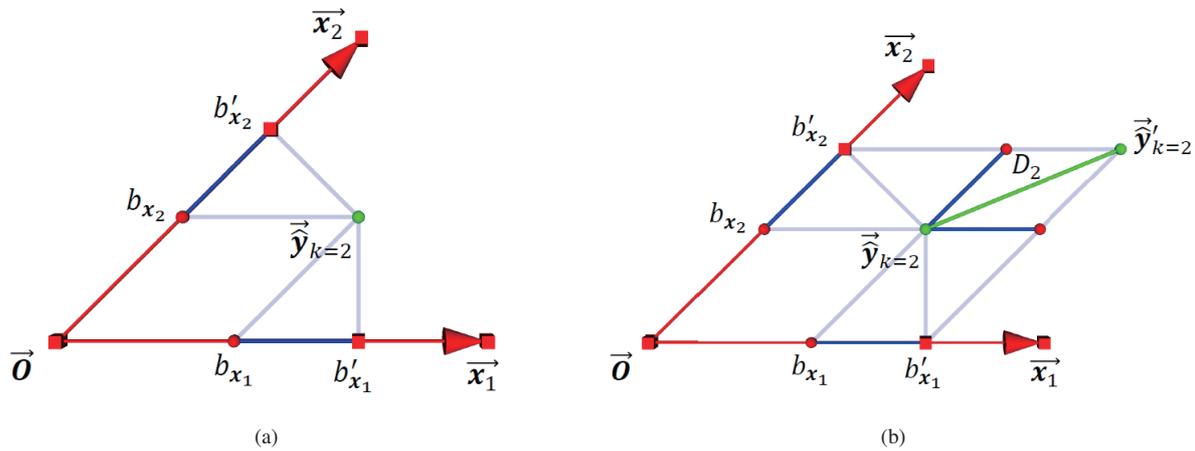


Figure 2. Placing two models on common collinear axes

angle between vectors  $\vec{x}_1$  and  $\vec{x}_2$ ) would demonstrate that small changes in the position of the response would be amplified by the change in the coefficient point estimates. This relationship of the response with the covariates mediates the impact of collinearity on the coefficients and standard errors. O'Brien (2007) demonstrates this on the variance of the estimates using a variance deflation factor (VDF).

$$VDF = 1 - R_y^2 \tag{2}$$

The VIF represents a multiplicative factor of the inflation in variance against a baseline of independence (i.e.  $r_{12} = 0, VIF = 1$ ). If we were to consider a VIF·VDF measure, it would place the impact of collinearity on the variance of a coefficient in the context of the model environment. The VIF (and similarly VIF·VDF) is a measurement on each of the predictors in the model, however such a measure is often difficult to interpret without a 'global' indicator of the collinearity present in a model. It will also not indicate which covariates are involved in linear dependencies (Belsley et al., 1980). In section 2.1 we develop a new index motivated by vector geometry that incorporates the covariance structure between the predictors and the response. In section 2.2 this concept is extended to the general case to provide a measure of the impact in regression models with  $k > 2$  predictors. In section 2.3 we further develop the measure to identify the individual role of each predictor in contributing to the observed 'global' impact. Finally, in section 3 we provide an illustrative regression study with the interpretation of the results discussed and compared to existing correlation based indices.

## 2. Methods

### 2.1 The Development of a Covariance Based Collinearity Index

A researcher should not rely exclusively on study data to assess the validity of a model. External information should be incorporated into the analysis to tailor the assessment to a particular discipline or setting. If we consider the impact of collinearity on an estimate to be a 'problem', then to assess that 'problem' we need an idea of what the population structure is. For instance, suppose  $x_1$  and  $x_2$  are two uncorrelated predictors in a population, but the sample observations are correlated. If we believe the population values to be uncorrelated (i.e. *a priori* assumption), the expectation is that the multivariable regression coefficients on both  $x_1$  and  $x_2$  are unchanged compared to their univariable regression coefficients (i.e. the regression models with only  $x_1$  or  $x_2$  entered). The fact that the sample values are correlated causes the univariable and multivariable estimates to differ. One feature of a 'covariance based' index would be to indicate the 'magnitude' of this deviation in the sample from this (or any other) chosen baseline. Another potential use is in model selection by comparing regression models with different predictors entered. Similarly, this may involve comparing univariable coefficients (as a baseline) to different multivariable regression models. In whichever application the index is required, the motivation remains to measure the deviation of the point estimates between models to illustrate the impact of collinearity on an expectation or a sample estimate.

To measure the disparity between two sets of estimates, we need to put them in a comparable setting. In vector geometry, we could do so by considering both in a common space. The traditional vector geometry representation is to project the response  $y$  orthogonally onto the regression space spanned by  $X$  (i.e. the vectors  $\vec{x}_1$  and  $\vec{x}_2$  in the bivariable example). The fitted response (labelled  $\hat{y}_{k=2}$ ) is then projected orthogonally onto the covariate vectors

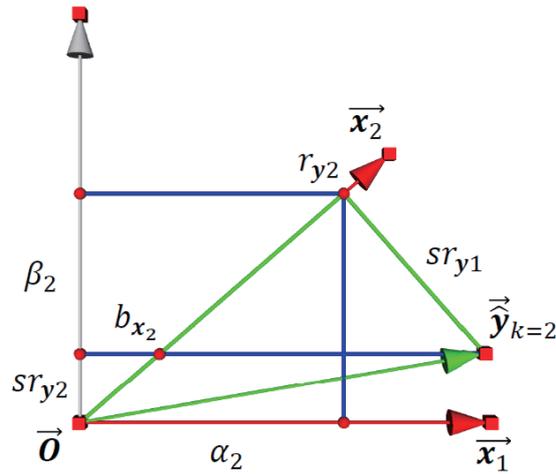


Figure 3. An Illustration of the computation of components  $\alpha_2$  and  $\beta_2$

to find the univariable point estimates  $b'_{x_j}$  and parallel to the complementary covariate vector to find multivariable estimates  $b_{x_j}$  (see Figure 2a). The distance between the projections would indicate the change in point estimates on each predictor moving from the univariable to multivariable model (i.e. shown in blue). To gain a 'global' measure of this impact on the overall model we choose to represent the individual univariable estimates as a single multivariable model involving orthogonal predictors (see Figure 2b). This is achieved by identifying an alternative fitted response  $\hat{y}'_{k=2}$ , that when projected parallel to the covariates (analogous to the construction of  $\hat{y}_{k=2}$ ), would attain the univariable point estimates  $b'_{x_j}$ , rather than the  $b_{x_j}$ . The distance between  $\hat{y}_{k=2}$  and  $\hat{y}'_{k=2}$  is the product of the change in regression coefficients, relative to the collinearity in the model (i.e. shown by the green line in Figure 2b). Using  $\hat{y}'_{k=2}$  removes the effect of a non-orthogonal projection and places the estimates from the univariable models on the collinear axes. This movement from  $\hat{y}_{k=2}$  to  $\hat{y}'_{k=2}$  represents a global measure of coefficient *deviation*, which we label  $D_2$ .

For the bivariable example, the calculation of the index  $D_2$  (with the subscript denoting the two covariates entered into the model) is found to be  $R_y \cdot r_{12}$  (see Appendix for a derivation of this result). The proof divides the index into 2 components. The first labelled  $\alpha_2$  is measured by the deviation parallel to  $\vec{x}_1$  and the second labelled  $\beta_2$  is the deviation orthogonal to  $\vec{x}_1$  (see Figure 3). To explain these components further, first consider a single predictor model including only  $x_1$  (i.e. a simple regression), which naturally assumes a zero impact of collinearity (i.e.  $D_1 = 0$ ). A second predictor  $x_2$  is then added to this model to generate an impact demonstrated by a non-zero  $D_2$  (unless  $x_2$  is uncorrelated with  $x_1$  or neither predictor explains any variance in the response). The unadjusted variance explained by  $x_2$  is  $r_{y2}$  (This quantity is demonstrated as a distance from the origin along the vector  $\vec{x}_2$ ). This variance on  $x_2$  can be divided into a portion that is 'overlapped' with  $x_1$  (i.e.  $\alpha_2$ ) and a portion of the variance explained by  $x_1$  confounded with  $x_2$  (i.e.  $\beta_2$ ). The component  $\alpha_2$  is demonstrated by a simple regression of the geometrical point  $r_{y2}$  (on the vector  $\vec{x}_2$ ) onto  $x_1$  (i.e.  $r_{y2} \cdot r_{12}$ ). The second component  $\beta_2$  is the residual variance of  $r_{y2}$  from this regression, subtracting the semi-partial correlation of  $x_2$  with  $y$  (i.e.  $r_{y2}$  variance attributed to  $x_2$  only). This is found to be  $r_{12} \cdot sr_{y1}$ . Therefore, we are projecting two components of the fitted response  $\hat{y}_{k=2}$  (i.e.  $r_{y2}$  and  $sr_{y1}$ , where  $|\hat{y}_{k=2}| = R_y = \sqrt{r_{y2}^2 + sr_{y1}^2}$ ) onto vectors to which they would have zero correlation at baseline. Any deviation of these components away from zero will represent an *impact* of collinearity demonstrated by a deviation in the point estimates.

The bivariable index  $D_2^2 = (r_{y2} \cdot r_{12})^2 + (sr_{y1} \cdot r_{12})^2 = (R_y \cdot r_{12})^2$  (squared to make the magnitude comparable to variance based diagnostics such as the VIF) represents the impact on the coefficient point estimates associated with the collinearity amongst the predictors and also the covariates relationship with the response. The composite direction vector formed by the two univariable regression coefficients ( $\hat{y}'_{k=2}$ ) is in the covariance maximizing direction on a single dimension. This is equivalent to a one component partial least squares regression (PLS) (Phatak & Dejong, 1997; Wold, Sjostrom, & Eriksson, 2001). The  $\hat{y}_{k=2}$  vector represents the OLS estimation. By definition, this is the covariance maximizing direction in the bivariable model, thus equivalent to a PLS regression with a full complement of components retained. We understand that the D-index is measuring the distance between an uncorrelated composite single dimension vector and the collinear predictors of the bivariable model. From the vector

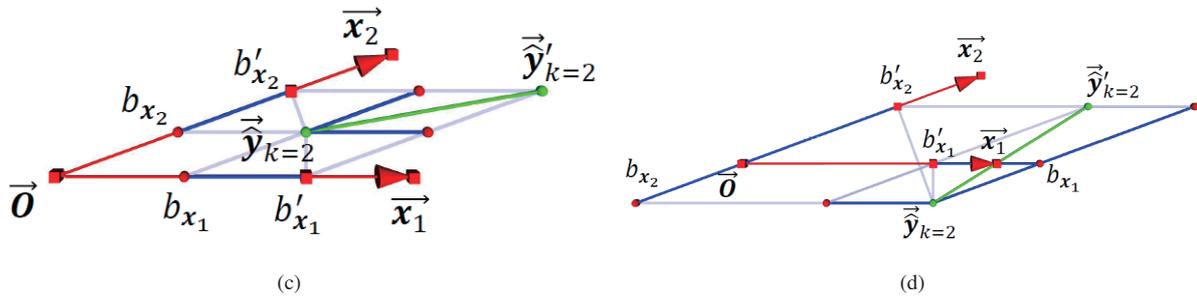


Figure 4. Vector geometry illustrating two examples with equal  $r_{12}$

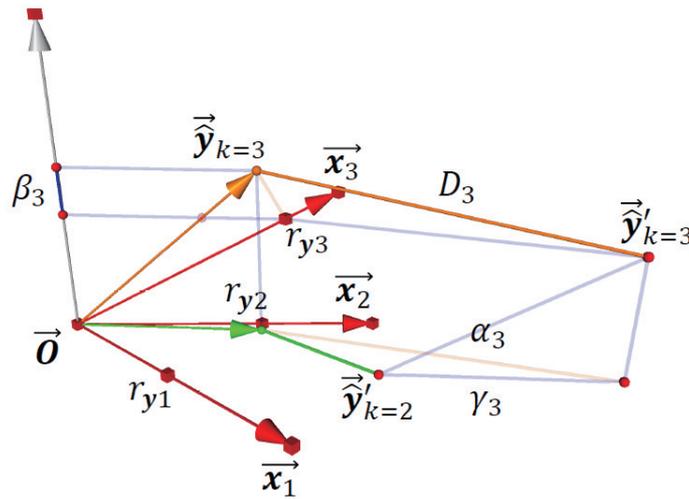


Figure 5. Illustration of the extension to  $D_2$  with  $x_3$  included

geometry, the length of the vector  $\hat{y}'_{k=2}$  is equal to the summation of the two univariable  $r_y^2$  estimates, relative to the collinearity present (i.e.  $|\hat{y}'_{k=2}| = r_{y1}^2 + r_{y2}^2 + 2r_{y1}r_{y2}r_{12}$  by the cosine rule). The length of  $\hat{y}_{k=2}$  is equal to the  $R_y$  found in the sample. The D-index represents the *impact* of collinearity on the point estimates of moving from an uncorrelated *prior* to a correlated sample estimate, or similarly the impact of collinearity in adding a second predictor to a ‘simple’ regression model. ‘Correlation based’ indices would be unable to distinguish between the examples in Figure 4 as the correlation amongst the covariates is identical. As illustrated by the geometry in Figure 4b the movement in the point estimates is far greater than in Figure 4a and a change of sign has occurred on  $\vec{x}_2$ . The change of sign may not be of particular interest statistically, but it could represent a potential change in the clinical interpretation. The D-index does not directly indicate a change of sign, but rather the greater propensity for a change of sign is reflected by an increased D-statistic (i.e. a greater movement). Under sampling variation, these deviations can become inflated or dampened with a potential impact on the conclusions of the study.

2.2 Extension of the Bivariable Case to a General Index

For the index to be of use in application it is important that it can be extended to models for  $k > 2$  predictors. We consider two options for extending this measure. First we look for the additional impact on the existing bivariable model (including  $x_1$  and  $x_2$ ) of adding a third predictor  $x_3$  (labelled  $\hat{D}_3$ ), and second the impact of collinearity on a baseline model that assumes orthogonality amongst *all* of the predictors (labelled  $D_3$ ). Figure 5 illustrates the vector geometry for the three predictor regression model. The fitted response of the trivariable model  $\vec{y}'_{k=3}$  is first projected orthogonally onto the covariate vectors  $\vec{x}_j$  to obtain the individual  $r_{yj}$  (i.e. regression coefficients from each univariable model). The  $r_{yj}$  are then projected along the plane formed by the remaining two predictors to identify  $\vec{y}'_{k=3}$  representing our baseline model of orthogonality amongst the covariates. The distance between  $\vec{y}'_{k=3}$  and  $\vec{y}_{k=3}$  forms the new  $D_3$  (analogous to the  $D_2$  computation). The fitted response in the three predictor model  $\vec{y}_{k=3}$  is an extension of  $\vec{y}_{k=2}$  in the direction orthogonal to the plane spanned by  $\vec{x}_1$  and  $\vec{x}_2$ . The orthogonality with the plane demonstrates that this extension represents a partial correlation between  $y$  and  $x_3$ , whilst holding  $x_1$  and

$\mathbf{x}_2$  constant-we label this correlation  $pr_{y_3|2}$ . The  $\hat{\mathbf{y}}'_{k=3}$  is an extension of the  $\hat{\mathbf{y}}'_{k=2}$  in the direction of  $\vec{\mathbf{x}}_3$  with length (i.e. variance) equal to  $r_{y_3}$ .

We first consider the calculation of the additional impact of adding  $\mathbf{x}_3$  to the already assumed  $D_2$  impact from the bivariable case. First, we project  $\vec{D}_3$  onto the 2-dimensional plane spanned by the vectors  $\vec{\mathbf{x}}_1$  and  $\vec{\mathbf{x}}_2$ . The projected  $\vec{D}_3$  is demonstrating the overlap between  $\mathbf{x}_3$  with variance  $r_{y_3}$  and the existing predictors in the model (i.e.  $\vec{\mathbf{x}}_1$  and  $\vec{\mathbf{x}}_2$ )-analogous to  $\alpha_2$  in the bivariable index. This projection is labelled  $\alpha_3$ , which is composed of  $D_2$  and a further component  $\gamma_3$  (see Figure 5). Following the previous construction of  $D_2$  we compute  $\gamma_3$  as two components. The first is parallel to  $\vec{\mathbf{x}}_1$ , found by an orthogonal projection of  $\vec{\mathbf{x}}_3$  with variance  $r_{y_3}$  onto  $\vec{\mathbf{x}}_1$ ,

$$\dot{\gamma}_3 = r_{y_3} \cdot r_{13} \quad (3)$$

This represents the overlap of  $\mathbf{x}_3$  (of length  $r_{y_3}$ ) and  $\mathbf{x}_1$ . The second (labelled  $\ddot{\gamma}_3$ ) is in the direction orthogonal to  $\vec{\mathbf{x}}_1$  in the plane spanned by  $\vec{\mathbf{x}}_1$  and  $\vec{\mathbf{x}}_2$ . This demonstrates that  $\vec{\mathbf{x}}_1$  is held constant, thus defining a semi-partial correlation between  $\vec{\mathbf{x}}_3$  and  $\vec{\mathbf{x}}_2$ , holding  $\vec{\mathbf{x}}_1$  constant (labelled  $sr_{23}$ ).

$$\ddot{\gamma}_3 = r_{y_3} \cdot sr_{23} \quad (4)$$

Therefore,  $\gamma_3$  is calculated as the squared sum of orthogonal components,

$$\gamma_3 = \sqrt{\dot{\gamma}_3^2 + \ddot{\gamma}_3^2} = \sqrt{(r_{y_3}r_{13})^2 + (r_{y_3}sr_{23})^2} \quad (5)$$

From Equation 5 we have an extension to  $D_2$  in the plane spanned by  $\mathbf{x}_1$  and  $\mathbf{x}_2$  after adding  $\mathbf{x}_3$  to the model. Finally, there is an additional deviation that would represent the new  $\beta$  component (labelled  $\beta_3$ ).  $\beta_3$  represents a deviation of the coefficients in a dimension orthogonal to the computation of  $D_2$ . The vector geometry illustrates that this is a projection of the remaining explained variance of  $\hat{\mathbf{y}}$  (i.e. the component of  $\mathbf{y}$  orthogonal to  $\vec{\mathbf{x}}_3$ ) onto an arbitrary axis orthogonal to the plane spanned by  $\vec{\mathbf{x}}_1$  and  $\vec{\mathbf{x}}_2$ . There is a residual from  $\mathbf{x}_3$  (of length  $r_{y_3}$ ) after regressing on  $\mathbf{x}_1$  and  $\mathbf{x}_2$ . This residual is composed of  $pr_{y_3|2}$  and  $\beta_3$  (analogous to our proof for  $D_2$  with the residual composed of  $sr_{y_2}$  and  $\beta_2$ ). Therefore,  $\beta_3$  is an impact of collinearity representing the explained variance of the original model confounded with  $\mathbf{x}_3$ .

$$\beta_3 = R_{3|2} \sqrt{sr_{y_1|3}^2 + pr_{y_2|2}^2} \quad (6)$$

The index  $\dot{D}_3$  can be calculated as the squared sum of the components  $\gamma_3$  and  $\beta_3$ ,

$$\dot{D}_3^2 = \gamma_3^2 + \beta_3^2 = [r_{y_3}^2(r_{13}^2 + sr_{23}^2)] + [R_{3|2} \sqrt{sr_{y_1|3}^2 + pr_{y_2|2}^2}]^2 \quad (7)$$

Returning to the vector geometry, we can summarise the computation of  $\dot{D}_3$ . The response  $\hat{\mathbf{y}}$  has been split into two components ( $r_{y_3}$  and  $sr_{y_2} + pr_{y_1|2}$ ). We project  $r_{y_3}$  onto the surface spanned by  $\vec{\mathbf{x}}_1$  and  $\vec{\mathbf{x}}_2$  (which would have zero correlation if it were uncorrelated with the baseline model) and project the second component ( $sr_{y_2} + pr_{y_1|2}$ ) onto  $\vec{\mathbf{x}}_3$  (which would similarly be uncorrelated at baseline). However, if a correlation is present it will generate a deviation of the point estimates represented by a non-zero  $\dot{D}_3$ . The advantage of using this measure (i.e. the bivariable model as baseline) is that the interpretation is much the same as the example for  $D_2$ . We again have two components of the response to project, only in this example one component represents a baseline model with the explained variance of two predictors rather than one.

The second index  $D_3$  is an impact of collinearity in moving from uncorrelated covariates at baseline to the three predictor model. In other words, if  $\mathbf{x}_3$  had zero correlation with both  $\mathbf{x}_1$  and  $\mathbf{x}_2$ , the  $\dot{D}_3$  would always be zero. However,  $\mathbf{x}_1$  and  $\mathbf{x}_2$  could still be correlated and so an impact on the point estimates from baseline orthogonality would still be seen, but it would be represented solely in the  $D_2$ . Now we look for an overall impact of collinearity to give  $D_2$  and  $D_3$  a common baseline for comparison. In this computation we place an emphasis on  $\mathbf{x}_3$  by considering the first component of  $\hat{\mathbf{y}}$  to be  $r_{y_3}$  (followed by  $sr_{y_1|3}$  and  $pr_{y_2|3}$ , however any construction of  $\hat{\mathbf{y}}$  would produce the same 'global' result. In the  $D_3$  measure we once again split  $\alpha_3$  into two components. The first deviation component is parallel to  $\vec{\mathbf{x}}_1$ , which is the summation of  $\alpha_2$  and  $\dot{\gamma}_3$ . This represents the portion of explained variance from  $\mathbf{x}_2$  and  $\mathbf{x}_3$  overlapped with  $\mathbf{x}_1$ . There is a second component of this impact in the plane spanned by  $\vec{\mathbf{x}}_1$  and  $\vec{\mathbf{x}}_2$ . This consists of the shared variance of  $r_{y_3}$  with  $\mathbf{x}_2$ , whilst holding  $\mathbf{x}_1$  constant. This is represented by the addition of  $\ddot{\gamma}_3$  and  $\beta_2$ . The final component of  $D_3$  is the deviation orthogonal to the plane spanned by  $\vec{\mathbf{x}}_1$  and  $\vec{\mathbf{x}}_2$ -this is the

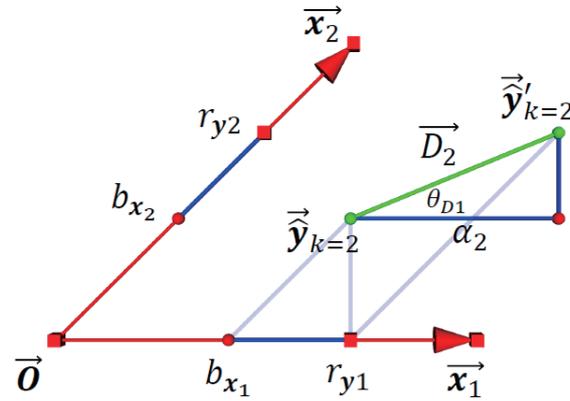


Figure 6. Computation of the angle between  $\vec{D}_2$  and  $\vec{x}_1$

component  $\beta_3$  identical to that computed for  $\vec{D}_3$ . The  $D_3^2$  demonstrating the overall impact from the baseline of orthogonality can now be expressed as follows,

$$D_3^2 = [r_{y2}r_{12} + r_{y3}r_{13}]^2 + [r_{y3}sr_{231}]^2 + [r_{12}sr_{y12} + R_{312} \sqrt{sr_{y13}^2 + pr_{y213}^2}]^2 \tag{8}$$

In both measures, the components of the response are projected onto vectors which would have zero correlation at the specified baseline. This deviation contributes to the global measure. We can extend these measures generated for the trivariable model to the general case with  $k$  predictors.

$$\dot{D}_k^2 = [r_{yk}^2(r_{k1}^2 + pr_{k21}^2 + \dots + pr_{k(k-1)1,2,\dots,(k-2)}^2)] + [R_{k1,2,\dots,(k-1)} \sqrt{pr_{y1k}^2 + pr_{y2k}^2 + \dots + pr_{y(k-1)k}^2}]^2 \tag{9}$$

$$D_k^2 = [r_{y2}r_{12} + r_{y3}r_{13} + \dots + r_{yk}r_{1k}]^2 + [r_{y3}sr_{231} + r_{y4}sr_{241} + \dots + r_{yk}sr_{2k1}]^2 + \dots + [r_{yk}pr_{(k-1)k1,\dots,(k-2)}]^2 + [R_{k1,2,\dots,(k-1)} \sqrt{pr_{y1k}^2 + pr_{y2k}^2 + \dots + pr_{y(k-1)k}^2}]^2 \tag{10}$$

We suggest that our measure in higher dimensions represents a generalized form of  $R_x \cdot R_y$ . The first index  $\dot{D}_k^2$  measures the impact of adding a single predictor to a baseline model (assumed as the model including  $k - 1$  predictors). The second index  $D_k^2$  assumes the predictors to be uncorrelated at baseline and incorporates the previous  $D_{k-1}^2$  impact as part of an overall measure.

### 2.3 Measurement of Impact on Individual Predictors

The global D-index only partly achieves our original goal in creating a regression tool for applied research. It is useful to highlight when there exists a high impact of collinearity on the point estimates, however it will not indicate which covariate contributes a greater impact to the deviation (a similar limitation to the VIF). This is the strength of an index such as the CI on the correlation matrix of the covariates. A feature of the D-index that we have ignored to this point is the direction of the deviation. In coordinate free vector geometry, the direction is relative to the collinear axes of the covariates. Therefore, we choose to focus on the angle between each covariate and the deviation  $D_k^2$  (which we now consider in vector form- $\vec{D}_k$ ). The angles (that in turn provide correlations) can perform a similar role to variance decomposition proportions alongside the CI in identifying which predictors are involved in a near dependency (Belsley, 1991). The vector geometry in Figure 6 demonstrates that each correlation (i.e. cosine of  $\theta_{D1}$ ) can be calculated in the bivariable model as the ratio of  $r_{yj}$  and  $R_y$ . For example,

$$\cos \theta_{D1} = r_{D1} = \frac{\alpha_2}{D_2} = \frac{r_{12} \cdot r_{y2}}{r_{12} \cdot R_y} = \frac{r_{y2}}{R_y} \tag{11}$$

The component  $\alpha_2$  is redefined for the target variable with which we wish to identify its contribution to the impact of collinearity. The correlation with  $\vec{D}_2$  is computed by setting arbitrary axes parallel and orthogonal to the target covariate. Therefore, if we are adding  $x_1$  to the simple regression model consisting of the predictor  $x_2$ , the arbitrary axis would be formed parallel to  $\vec{x}_2$  and represent the degree to which  $r_{y1}$  is explained by  $x_2$ . Scaling by  $D_2$  removes the inflation effect of collinearity, thus normalizing the quantity to place the estimate on a scale of 0 – 1. If the explained variance on each predictor in the univariable models is equal, then the correlations with  $\vec{D}_2$  (in the

bivariable case) will be equally split. However, if the ratio is larger on one covariate, then the covariate with the weaker correlation to the response will have a greater association with  $\vec{D}_2$ . This dictates the direction of global change (i.e.  $\vec{D}_2$ ) to be greater in the direction of the covariate with the weaker correlation to the response. Extending to the general case, the calculation remains similar with the correlation calculated as the ratio of  $\alpha_k$  to  $D_k$ .

### 3. Example

The index was applied to data from a study by Penrose et al. (1985). The study recorded percentage body fat and several body circumference measures of 252 men. We used the data to explore the inter-relationship between body composition using external measurements of different body circumference variables and how these highly correlated variables can then be explored to create the optimal model to explain percentage body fat.

Table 1. Pearson correlations for the body fat study

	y = body fat	x <sub>1</sub> = neck	x <sub>2</sub> = abdomen	x <sub>3</sub> = biceps
x <sub>1</sub> = neck	0.49			
x <sub>2</sub> = abdomen	0.81	0.75		
x <sub>3</sub> = biceps	0.49	0.73	0.68	
x <sub>4</sub> = ribs	0.35	0.74	0.62	0.63

The aim was to discover which subset of easily measurable body circumference measurements (x<sub>1</sub> = neck, x<sub>2</sub> = abdomen, x<sub>3</sub> = biceps, x<sub>4</sub> = ribs) could be used to represent body fat (see Table 1 for correlations between the predictors and the response). This would allow a reduction in the number of measurements required reducing study length, cost and participant burden whilst maintaining the measurements that most accurately represent total body fat.

Table 2. Results from the D-index for the four predictor body fat study

Model	$R_y^2$	$D_k^2$ (95% CI)	$r_{D1}$	$r_{D2}$	$r_{D3}$	$r_{D4}$
x <sub>1</sub> , x <sub>2</sub>	0.24	0.13 (0.08 to 0.18)	0.99	0.71	-	-
x <sub>1</sub> , x <sub>3</sub>	0.70	0.27 (0.25 to 0.29)	0.97	-	0.41	-
x <sub>1</sub> , x <sub>4</sub>	0.25	0.10 (0.06 to 0.13)	0.99	-	-	0.70
x <sub>2</sub> , x <sub>3</sub>	0.70	0.40 (0.36 to 0.43)	-	0.97	0.59	-
x <sub>2</sub> , x <sub>4</sub>	0.28	0.15 (0.10 to 0.20)	-	0.93	-	0.93
x <sub>3</sub> , x <sub>4</sub>	0.67	0.31 (0.28 to 0.34)	-	-	0.60	0.99
x <sub>1</sub> , x <sub>2</sub> , x <sub>3</sub>	0.71	0.89 (0.65 to 1.16)	0.92	0.92	0.62	-
x <sub>1</sub> , x <sub>2</sub> , x <sub>4</sub>	0.28	0.50 (0.32 to 0.74)	0.95	0.87	-	0.81
x <sub>1</sub> , x <sub>3</sub> , x <sub>4</sub>	0.70	0.79 (0.54 to 1.05)	0.92	-	0.62	0.87
x <sub>2</sub> , x <sub>3</sub> , x <sub>4</sub>	0.70	1.06 (0.80 to 1.32)	-	0.95	0.69	0.89
x <sub>1</sub> , x <sub>2</sub> , x <sub>3</sub> , x <sub>4</sub>	0.71	1.77 (1.28 to 2.33)	0.89	0.93	0.70	0.85

Consider the  $D_k^2$  produced by each model (see Table 2), the greatest impact of collinearity is highlighted for the model involving x<sub>2</sub> and x<sub>3</sub>. This follows with the maximal correlation and subsequently the VIF ( $r_{23} = 0.75$ ,  $VIF_{23} = 2.29$ ). The  $r_{D2}$  demonstrates that the covariate with the greater correlation to the response is x<sub>3</sub>, highlighting that x<sub>2</sub> provides the greater contribution to the impact of collinearity in the model. Studying the correlations between covariates indicates that the model involving x<sub>2</sub> and x<sub>4</sub> has a similarly high correlation ( $r_{24} = 0.73$ ). For this example the variance in y explained by both predictors is low ( $r_{y2} = 0.49$ ,  $r_{y4} = 0.49$ ) and so the impact of collinearity on the model has been limited by the low model  $R_y^2$ . However, both  $r_{Dj}$  are large indicating that the collinearity is high. Therefore, the individual predictors perform an important role in indicating a potential ‘problem’ even when the global inflation indicated by  $D_2$  is low.

In each bivariable model the covariate x<sub>1</sub> had the strongest correlation with  $\vec{D}_2$ . This is demonstrated by the low correlation with y (i.e.  $r_{y1} = 0.35$ ). We also observe that x<sub>3</sub> consistently had the lowest correlation with  $\vec{D}_2$  (for any model) suggesting that it would be a useful predictor to include in the model due to its high explanatory power. A confidence interval for the two predictor models (shown in parentheses in Table 2) was generated using the standard error of  $R_y^2$  (Cohen, 2003), whilst  $r_{12}$  is fixed (due to predictors assumed to be measured without error). A confidence interval for the three and four predictor models was bootstrapped using a “leave one out” approach

(Tukey, 1958). We notice that  $R_y^2$  does not increase greatly beyond the two predictor model that included  $x_1$  and  $x_3$ . The correlations  $r_{Dj}$  indicate that  $x_1$  was the main contributor to this impact of collinearity in the bivariable model. We observe a very moderate increase in  $R_y^2$  after including  $x_2$  in this model. However, with this inclusion our D-index has increased from 0.27 to 0.89. We can calculate the additional impact of adding the predictor  $x_2$  to this model as  $\bar{D}_3^2 = 0.49$ . This would appear high when viewed alongside other bivariable measures to attain a small increase in  $R_y^2$ .

We notice that when  $x_3$  is entered into the model along with  $x_1$  and  $x_2$  the  $r_{Dj}$  are equal for both  $x_1$  and  $x_2$ . In comparison, when  $x_4$  is added to the model with  $x_1$  and  $x_2$ ,  $r_{D1}$  is greater than  $r_{D2}$ . This demonstrates how the role of each predictor changes dependent on others entered into the model. In the full four predictor model the  $R_y^2$  reaches 0.71, however the deviation peaks at 1.77. The collinearity structure of the four predictors and the variance explained suggests that  $x_2$  is the greatest contributor to this impact of collinearity. This is a change to  $x_1$  being consistently high in previous models. Observing model parsimony would seem to discount the four predictor model. Removing  $x_2$  produces a model with a high  $R_y^2$  (0.70) and moderately low  $D_3^2$  (0.79). Excluding  $x_2$  from the model would not seem obvious from only observing the three predictor models (due to the consistently high  $r_{D1}$ ), however noticing the impact in the full model has highlighted the statistical dependency of this predictor with others in the study.

#### 4. Discussion

From a model building perspective the bivariable model with  $x_1$  and  $x_3$  included as predictors would appear optimal. This model explained a high variance of the response and had a relatively low  $D_2$ . We can demonstrate that adding the predictor  $x_2$  to this model, whilst moderately increasing the  $R_y^2$ , would generate a high deviation in the point estimates of the existing model. Also, when considering the full set of predictors,  $x_2$  would seem to have the greatest impact. Therefore, if any predictor would be added to the bivariable model,  $x_4$  would seem the better option from a collinearity perspective. However, adding  $x_4$  does not increase the explanatory power of the model and so this may not be wise. This example has been much simplified as we have not considered the nature of any causal relationships amongst the covariates. This would raise the complexity of the problem and our understanding of incorporating collinearity in the model. We are instead focussing on the purely statistical aspect of what our measure indicates.

The greatest change in impact from  $D_2$  to  $D_3$  is after adding  $x_3$  to the model including  $x_2$  and  $x_4$ . However,  $x_3$  contributes the greatest explained variance individually ( $r_{y3} = 0.81$ ) and so including this predictor would appear a sensible decision. It is labelled the most beneficial of the predictors by our correlations with  $\bar{D}$ , suggesting  $x_1$  and  $x_2$  contribute greatly to the impact of collinearity in this model. The high change in global impact could be misleading if it were interpreted as a measure of some collinearity ‘problem’. This is why it would seem beneficial that any change in D-index between models be interpreted alongside the  $R_y^2$ . If little explained variance is gained by including an additional predictor in the model, but the deviation is high, then this should perhaps be viewed as a potential warning (based on the conceptual model employed) of the impact of collinearity on the model estimates. If the global inflation is small, but the correlations between predictor and  $\bar{D}$  are large, this would suggest a high degree of collinearity that is being moderated by a low  $R_y^2$ .

#### 5. Concluding Remarks

In this study we have demonstrated the important role that the response plays in mediating the *impact* of collinearity in an applied regression study. This has demonstrated the need for a collinearity index, not simply to describe the degree of collinearity amongst the covariates, but to identify the potential impact on the variance and point estimates in relation to the response entered into the model. We have developed a novel index based on vector geometry and regression theory that assesses a global deviation in the point estimates and analyses the role of each predictor in contributing to this effect. When interpreting the D-index for use in model building it may appear conceptually appealing to assume a greater  $R_y^2$  to be beneficial to the estimation. This is how variance based inflation (such as the  $VIF \cdot VDF$ ) would be interpreted and to some would seem a more natural metric. However, it is important to stress that the D-index is not necessarily measuring a ‘problem’. A high  $R_y^2$  will inflate the point estimates under collinearity and this is subsequently reflected in our index. If we were comparing to a baseline *prior*, whether that be a zero correlation or some ‘guesstimate’ of a population correlation, then we would wish to know the deviation of the estimates away from our expectation. This is not representing a biased or ‘wrong’ estimate, but a population *prior* that is not reflected in the single sample case. Therefore, if a greater  $R_y^2$  inflates the change in coefficients, then we would wish to know the degree of inflation in the sample data.

The D-index could be developed in the future to produce a more natural interpretation for model building. Replacing explained variance with some reciprocal estimate could deflate the impact. However, collinearity remains a complex feature in application and the development of a statistical index still requires a very careful conceptual understanding to be of benefit in application. The work in this paper should only be viewed as a starting point for future methodological development and simulation studies. An achievement in the development of this index is in the use of vector geometry to create the measure and to interpret it. One of the reasons for proposing the geometric alternative to the  $VIF \cdot VDF$  is that it allows flexibility to incorporate different *a priori* assumptions. This would be achieved by varying angles of projection to reflect different correlations.

## References

- Belsley, D. A. (1991). *Conditioning Diagnostics: Collinearity and Weak Data in Regression*. New York: John Wiley & Sons.
- Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: Wiley. <http://dx.doi.org/10.1002/0471725153>
- Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied Multiple Regression/Correlation Analysis for Behavioral Sciences* (3rd ed.). Mahwah, N.J.: Lawrence Erlbaum Associates.
- Draper, N. R., & Smith, H. (1998). *Applied Regression Analysis* (3rd ed.). New York: Wiley-Blackwell.
- Freund, R. J., & Wilson, W. J. (1998). *Regression analysis : statistical modeling of a response variable*. San Diego: Academic Press.
- Marquardt, D. W. (1970). Generalized Inverses, Ridge Regression, Biased Linear Estimation, and Nonlinear Estimation. *Technometrics*, 12(3), 591-612. <http://dx.doi.org/10.2307/1267205>
- O'Brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity*, 41(5), 673-690. <http://dx.doi.org/10.1007/s11135-006-9018-6>
- Penrose, K. W., Nelson, A. G., & Fisher, A. G. (1985). Generalized Body-Composition Prediction Equation for Men Using Simple Measurement Techniques. *Medicine and Science in Sports and Exercise*, 17(2), 189. <http://dx.doi.org/10.1249/00005768-198504000-00037>
- Phatak, A., & Dejong, S. (1997). The geometry of partial least squares. *Journal of Chemometrics*, 11(4), 311-338. [http://dx.doi.org/10.1002/\(SICI\)1099-128X\(199707\)11:4<311::AID-CEM478>3.3.CO;2-W](http://dx.doi.org/10.1002/(SICI)1099-128X(199707)11:4<311::AID-CEM478>3.3.CO;2-W)
- Stine, R. A. (1995). Graphical Interpretation of Variance Inflation Factors. *American Statistician*, 49(1), 53-56. <http://dx.doi.org/10.2307/2684812>
- Wickens, T. D. (1995). *The Geometry of Multivariate Statistics*. London: Psychology Press.
- Wold, S., Sjostrom, M., & Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58(2), 109-130. [http://dx.doi.org/10.1016/S0169-7439\(01\)00155-1](http://dx.doi.org/10.1016/S0169-7439(01)00155-1)

**Appendix**

The derivation of the  $D_2$  index can be demonstrated by considering two components of the  $D_2$  vector (labelled  $\alpha_2$  and  $\beta_2$ )-see Figure 7.

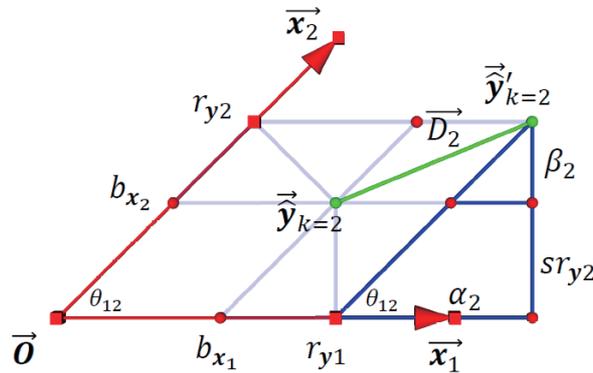


Figure 7. Construction of  $D_2$  as two components  $\alpha_2$  and  $\beta_2$

The first component  $\alpha_2$  is parallel to  $\vec{x}_1$  and the second component  $\beta_2$  is orthogonal to  $\vec{x}_1$ . The proof of the index (using the triangle highlighted in blue) can be shown as follows,

$$\alpha_2 = r_{y2} \cos \theta_{12} = r_{12}r_{y2} \tag{12}$$

$$\begin{aligned} \beta_2 &= r_{y2} \sin \theta_{12} - sr_{y2} \\ &= r_{y2} \sqrt{1 - r_{12}^2} - \frac{r_{y2} - r_{y1}r_{12}}{\sqrt{1 - r_{12}^2}} \\ &= \frac{r_{y2}(1 - r_{12}^2) - r_{y2} + r_{y1}r_{12}}{\sqrt{1 - r_{12}^2}} \\ &= r_{12} \cdot \frac{r_{y1} - r_{y2}r_{12}}{\sqrt{1 - r_{12}^2}} \\ &= r_{12}sr_{y1} \end{aligned} \tag{13}$$

The index is then calculated as the squared sum of orthogonal components,

$$D_2^2 = \alpha_2^2 + \beta_2^2 = (r_{12}r_{y2})^2 + (r_{12}sr_{y1})^2 = R_y^2 r_{12}^2$$

$$D_2 = R_y^2 r_{12} \tag{14}$$

Therefore, the index is the product of the correlation between  $x_1$  and  $x_2$  (i.e.  $r_{12}$ ) and the root of the variance explained in the response  $R_y$ .