OnLine Journal of Biological Sciences

Original Research Paper

Comparison of Collinearity Indices for Linear Models in Agricultural Trials

[1]Danny Villegas Rivas, [2]José M. Palacios Sánchez, [2]Cristina A. Alzamora Rivero, [2]Carlos M. Franco Del Carpio, [3]César Osorio Carrera, [1]Martin Grados Vasquez, [1]Luis Ramírez Calderón, [1]Karin Ponce Rojas, [1]José Jorge Rodríguez Figueroa, [2]Cáceres Narrea, [2]Felicia L., [2]Saravia Pachas, [2]Delia A., [2]Arrieta Benoutt, [2]Felipe, [2]Arturo N. Neyra Flores, [2]Pedro E. Zata Pupuche, [1]Carlos Fabián Falcón, [1]Yolanda Maribel Mercedes Chipana Fernández, [1]Víctor Hugo Fernández Rosas, [1]Francisco Alejandro Espinoza Polo, [1]Gaby Esther Chunga Pingo, [4]Mercy Carolina Merejildo Vera, [5]Carlos Alfredo Cerna Muñoz, [5]Luis Orlando Miranda Diaz, [6]Miguel Ángel Hernández López, [7]Martín Desiderio Vejarano Campos, [8]Erick Delgado Bazán, [8]Zadith Garrido Campaña, [9]José Paredes Carranza, [10]Leyli J. Aguilar Ventura, [2]Graciela M. Monroy Correa, [11]Ruth A. Chicana Becerra, [12]Jhonny Richard Rodriguez Barboza, [13]Mariella M. Quipas Bellizza, [14]Fernando Emilio Escudero Vilchez and [14]Silvia Liliana Salazar Llerena

[1]*Graduate School, Universidad César Vallejo, Peru*
[2]*Faculty of Education, Universidad Nacional Federico Villarreal, Peru*
[3]*Graduate School, Universidad Continental, Peru*
[4]*Graduate School, Universidad Privada Norbert Wiener, Lima, Perú*
[5]*Graduate School, Universidad Católica de Trujillo Benedicto XVI, Trujillo, Perú*
[6]*Graduate School, Universidad Nacional San Luis Gonzaga, Ica, Perú*
[7]*Graduate School, Universidad Señor de Sipán, Chiclayo, Perú*
[8]*Faculty of Engineering. Universidad Nacional de Jaén, Cajamarca, Perú*
[9]*School of Medical Technology, Universidad Nacional de Jaén, Cajamarca, Perú*
[10]*Graduate School, Universidad Nacional de Trujillo, Perú*
[11]*Graduate School, Universidad del Pacífico, Perú*
[12]*Center for Language Studies, Universidad San Ignacio de Loyola, Perú*
[13]*Graduate School, Universidad Femenina del Sagrado Corazón, Peru*
[14]*Faculty of Engineering, Sciences and Administration, Universidad Autónoma de Ica, Peru*

Article history
Received: 14-06-2023
Revised: 01-09-2023
Accepted: 09-09-2023

Corresponding Author:
Danny Villegas Rivas
Graduate School, Universidad César Vallejo, Peru
Email: danny_villegas1@yahoo.com

Abstract: The deleterious consequences of collinearity in linear regression on the precision of estimators of regression coefficients and the interpretability of the fitted model are widely recognized. In this study, we compare several methodologies for assessing collinearity in linear models and explore the effect of outliers on collinearity. The robustness of collinearity measures (individual and overall) is validated through two detailed Monte Carlo simulation study which also considers the effect of outliers on collinearity indices. The methods are illustrated with two real-world agricultural and fish morphology l data sets to show potential applications. The results do not provide any evidence for an effect from outliers on collinearity identification using the collinearity indices (individual and overall). The FG and F_j collinearity indices more robust as both sample size and collinearity degree increase. The VIF (individual measure) had a better performance on the fitted model with a greater number of parameters.

Keywords: Multicollinearity, Overall Some Individual Indices, Monte Carlo Simulation, Mctest Package

© 2024 Danny Villegas Rivas, José M. Palacios Sánchez, Cristina A. Alzamora Rivero, Carlos M. Franco Del Carpio, César Osorio Carrera, Martin Grados Vasquez, Luis Ramírez Calderón, Karin Ponce Rojas, José Jorge Rodríguez Figueroa, Cáceres Narrea, Felicia L., Saravia Pachas, Delia A., Arrieta Benoutt, Felipe, Arturo N. Neyra Flores, Pedro E. Zata Pupuche, Carlos Fabián Falcón, Yolanda Maribel Mercedes Chipana Fernández, Víctor Hugo Fernández Rosas, Francisco Alejandro Espinoza Polo, Gaby Esther Chunga Pingo, Mercy Carolina Merejildo Vera, Carlos Alfredo Cerna Muñoz, Luis Orlando Miranda Diaz, Miguel Ángel Hernández López, Martín Desiderio Vejarano Campos, Erick Delgado Bazán, Zadith Garrido Campaña, José Paredes Carranza, Leyli J. Aguilar Ventura, Graciela M. Monroy Correa, Ruth A. Chicana Becerra, Jhonny Richard Rodriguez Barboza, Mariella M. Quipas Bellizza, Fernando Emilio Escudero Vilchez and Silvia Liliana Salazar Llerena. This open-access article is distributed under a Creative Commons Attribution (CC-BY) 4.0 license.

Introduction

In the context of multiple regression analysis, multicollinearity refers to a scenario where there is a pronounced interconnection among the explanatory variables (Wondola *et al.*, 2020). The presence of collinearity indicates that a substantial part of the information in one or more of these covariates is redundant. Habshah *et al.* (2009) pointed out that collinearity, or non-orthogonality of the design matrix, is an almost linear dependence between two or more covariates. According to Silvey (1969); Belsley *et al.* (1980), in cases where the variables exhibit linear correlations, it is possible for one or more eigenvalues of the model X'X to be relatively tiny. The presence of collinearity causes difficulties in the estimation of model parameters, variable selection and model interpretation. When covariates in a regression model are not orthogonal, inference based on estimates of model parameters can be invalid. Multicollinearity leads to increased variances in the estimated parameters, which might result in the individual predictors appearing statistically insignificant despite the overall model being significant. When multicollinearity is present, it can complicate the estimation of the beta coefficients and their interpretation. As multicollinearity intensifies, the confidence intervals for the regression coefficients become wider and the t-statistics shrink in value. For coefficients to be deemed statistically significant under these conditions, they must be larger, implying that rejecting the null hypothesis becomes more challenging when multicollinearity exists. However, it's important to note that large standard errors can arise from factors other than multicollinearity (Oke *et al.*, 2019).

While the model's predictive performance may remain unaffected. When the focus of the investigation is to determine how the covariates' independent effects differ from one another, the existence of collinearity presents a substantial obstacle. The reason for this phenomenon is that when collinearity is present, the estimates of regression coefficients become less stable, resulting in larger Standard Errors (SEs) for these coefficients. In addition to the collinearity problem, although multiple linear models are widely used, it is well known that atypical observations can have a high impact on parameter estimates, predicted values and estimates of the covariance matrix; Cook (1977). Although there are many procedures used to detect collinearity, they are generally based on ad-hoc practical rules and are often unreliable with unquantifiable error rates. These procedures can be categorized as those based on three key aspects to consider in this study: (i) The correlation among covariates, (ii) The structure of the design matrix and (iii) Descriptive indices such as the condition index discovered by Belsley *et al.* (1980) and the factor of inflation variance (VIF) as discussed in Kutner *et al.* (2005); Fox and Monette (1992); Hair *et al.* (2014).

It is important to note that even these descriptive indices are not without their critics (for example, Gunst, 1984; O'brien, 2007) and new qualitative measures continue to be recommended; see, for example, Chennamaneni *et al.* (2016). Farrar and Glauber (1967) introduced an inferential technique for evaluating collinearity in linear models by examining deviations from orthogonality in the design matrix. However, this method has faced significant criticism from researchers such as O'Hagan and McCabe (1975); Wichers (1975); Haitovsky (1969). Based on the current state of knowledge, it appears that there are no alternative methodologies currently accessible for assessing collinearity in linear models. Subjective diagnostics have become increasingly prevalent in contemporary research. A notable example is the R package mctest, which was introduced by Imdadullah *et al.* (2016). In general, the user is left to rely upon rule-of-thumb criteria to judge the severity of collinearity. Furthermore, if an observation in a linear model has a large value on two or more covariates, artificial collinearity may be induced. The effect of such collinearity in regression models, especially in biological science where covariates are strongly correlated is not totally studied. The aforementioned literature, including Sengupta and Bhimasankaram (1997); Walker and Birch (1988); Mason and Gunst (1985), demonstrates that there exists a resemblance between the outcome and an estimated linear relationship.

The objectives of this study are: (i) To evaluate how the diagnostic measures (individual and overall) are affected by atypical observations; (ii) To assess the performance of the collinearity indices by simulations; (iii) To apply the new indices to real-world morphological and agricultural data sets with different collinearity structures and atypical cases. All numerical evaluations carried out in this study were implemented in the R software (Core Team, 2016).

Materials and Methods

Collinearity Indices

The collinearity diagnostic measures used and implemented in R with the mctest package proposed by (Imdadullah *et al.*, 2016), are described by these authors as detailed below.

Overall Collinearity Diagnostic Measures

Determinant

The matrix $X'X$ will exhibit singularity if it possesses linearly dependent columns or rows. Hence, the

DOI: 10.3844/ojbsci.2024.195.207

determinant of the normalized correlation matrix R, which is obtained by multiplying the transpose of matrix X with X and excluding the intercept term, might serve as an indicator for the presence of collinearity among the regressors. Nevertheless, it is remarkable to note that the determinant of a matrix does not offer insights into the dependency between regressors. Instead, it merely indicates the singularity or departure from orthogonality of a correlation matrix. According to Cooley and Lohnes (1971), the value of $X'X$ on the scale falls within the range of $0 \leq |X'X| \leq 1$. According to Asteriou and Hall (2007), if the determinant of the value $X'X$ is around zero, it indicates a presence of collinearity among the regressors.

R-Squared

R^2 is obtained by doing a regression analysis of all x variables on y. According to Stock and Watson (2010), R^2 exhibits a monotonically non-decreasing relationship with the number of regressors incorporated into the model. In other words, R^2 serves as an indicator of the extent to which the regression accurately captures the data. Conversely, when the R^2 values increase, there is a greater likelihood of the regressors being affected by multicollinearity, as the R^2 is influenced by the regressors sharing their variances (Asteriou and Hall, 2007).

Farrar χ^2

It is the Chi-square test for detecting the strength of collinearity over the complete set of regressors. $\chi^2 = -\left[n - 1 - \frac{1}{6(2p+5)}\right] \times log_e[X'X] \sim \psi^2_{v=\frac{1}{2}p(p-1).}$

Collinearity exists among regressors if $\chi^2 > \chi^2_{\frac{1}{2}p(p-1)}$(Farrar and Glauber, 1967).

Condition Index

$$CI_j = \sqrt{\frac{máx(\lambda_j)}{\lambda_j}}\ j = 1,2,\ldots,p;\ \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$$

Collinearity exists if any of CI_j >10, 15 or 30 (Belsley *et al*., 1980; Chatterjee and Hadi, 2013).

Sum of Reciprocal of Eigenvalues

In an orthogonal system $\sum_{j=1}^{p} \frac{1}{\lambda_j} = p$, therefore, for a sample based correlation matrix R with eigenvalues λ_j, comparing p with $\sum_{j=1}^{p} \frac{1}{\lambda_j}$ can be used to indicate collinearity. If $\sum_{j=1}^{p} \frac{1}{\lambda_j}$ is (say) five times larger than the number of regressors used in the model then collinearity exists among regressors (Chatterjee and Price, 1977; Dillon and Goldstein, 1984).

Theil's Indicator

Theil (1971) proposed a measure of collinearity based on an incremental contribution $(R^2 - R_j^2)$ to the squared multiple correlation, where R_j^2 is the R^2 from auxiliary regression of regressors:

$$m = R^2 - \sum_{i=1}^{p} (R^2 - R^2_{-i})$$

If $m = 0$ then all $X's$ are mutually uncorrelated (no redundancy exists) as the incremental contribution all add up to R^2. However, if $m \sim 1$ then Collinearity exists among regressors.

Red Indicator

In their study, Kovács *et al*. (2005) introduced a novel and synthetic normalized indicator for diagnosing collinearity. This indicator leverages eigenvalues or quantifies the average correlation of the data:

$$Red = \frac{\frac{\sqrt{\sum_{j=1}^{p} (\lambda_j - 1)^2}}{p}}{\sqrt{p-1}}$$

In the event that the value of the *Red* indicator is zero (*Red* = 0), it signifies the lack of redundancy, while a value close to 1 (*Red* ~ 1) indicates the presence of maximal redundancy.

Individual Collinearity Diagnostic Measures

Klein's Rule

If the value of R_j obtained from the auxiliary regression exceeds the total R^2 obtained from the regression of y on all the regressors, it suggests the presence of potential issues with multicollinearity. The decision rule for the discovery of collinearity is., $R^2_{x_j.x_1,x_2,\ldots,x_p} > R^2_{y.x_1,x_2,\ldots,x_p}$ (Klein, 1969).

VIF and Tol

The Variance Inflation Factor (VIF) quantifies the extent to which the variances of the predicted regression coefficients are amplified when there is no connection among the p regressors. The significance of the diagonal elements in the $((X'X)^{-1})$ matrix for identifying multicollinearity is widely recognized:

$$VIF_j = (X'X)^{-1}_{jj} = \frac{1}{1-R_j^2} \text{ and } Tol_j = \frac{1}{VIF_j} = 1 - R_j^2$$

The criticism on *VIF* is that $var(\hat{\beta}_j) = \frac{\sigma^2}{\sum x_j^2} VIF$ depends on σ^2, $\sum x_j^2$ and *VIF*, which shows that a high *VIF* can be counterbalanced by a low σ^2 or high $\sum x_j^2$. So a high *VIF* is neither a necessary nor a sufficient measure of multicollinearity. The value of VIF >3, 5 ,10 or value of $Tol \sim 0$ indicates existence of collinearity among regressors (Neter *et al*., 2004).

Eigenvalues

Kendall (1957); Silvey (1969) proposed use the eigenvalues of the correlation matrix ($X'X$) as a means to assess the existence of multicollinearity. They established that small eigenvalues, which are close to zero, serve as an indication of high collinearity. However, they did not specify the precise threshold for determining the degree of smallness. The presence of one or more lower eigenvalues in the matrix $X'X$ or its corresponding correlation matrix is indicative of collinearity.

CVIF

Curto and Pinto (2011) introduced a novel metric for assessing multicollinearity, which aims to quantify the influence of intercorrelation among independent variables on the variance of the Ordinary Least Squares Estimators (OLSEs):

$$CVIF_j = VIF_j \times \frac{1 - R^2}{1 - R_0^2}$$

where, $R_0^2 = R_{yx1}^2 + R_{yx2}^2 + \cdots + R_{yxp}^2$. Collinearity exists if $CVIF_j \geq 10$.

Leamer's Methods

Leamer in Greene (2002) suggested a measure of the effect of multicollinearity for the j^{th} variable:

$$C_j = \left\{ \frac{\left(\left(\sum_{i=1}^{n} \left(X_{ij} - \bar{X}_j \right)^2 \right)^{-1} \right)}{(X'X)_{jj}^{-1}} \right\}^{\left(\frac{1}{2}\right)}$$

This measure is the square root of the ratio of variances of estimated coefficients $(\hat{\beta}_j)$ when estimated without and with the other regressors. If X_j is uncorrelated with the other regressors C_j would be 1 otherwise will be equal to $\left(1 - R_j^2\right)^{\frac{1}{2}}$, i.e., $C_j \sim 0$ indicates existence of collinearity among regressors.

F and R^2 Relation

The relationship of F-test and R^2 from regressing X_j on the other remaining regressors can be used to detect multicollinearity. The relationship is described as:

$$F_j = \frac{\frac{R_{x_j,x_1,\ldots,x_p}^2}{p-2}}{\frac{1 - R_{x_j,x_1,\ldots,x_p}^2}{n-p+1}} \sim F_{(p-2,n-p+1)},$$

where, $F^* = F_{p-2,n-p+1}$. If $F_j > F^*$, then it means that the regressor X_j is collinear with other regressors and it should be dropped from the model (Gujarati and Porter, 2003).

Farrar w

It is an *F*-test for locating the regressors which are collinear with others and it makes use of multiple correlation coefficients among regressors:

$$w_j = \frac{R_j^2}{1 - R_j^2}\left(\frac{n-p}{p-1}\right) \sim F_{(n-p,p-1)}$$

If $w_j > F_{(n-p,p-1)}$, there is indication of considerable collinearity (Farrar and Glauber, 1967).

Most of the overalland individual measures to detect multicollinearity described above are included in the *R* mctest package, which mainly implements functions for detecting multicollinearity between covariates using the omcdiag () functions in the case of general measures and imcdiag () for individual measurements (Imdadullah *et al*., 2016).

Simulation Studies

Simulation I

The primary objective of the initial Monte Carlo simulation study is to accomplish the following: (a) Demonstrate the application of collinearity tests; (b) Determine the accuracy rate of correctly identifying collinearity cases using collinearity indices; (c) Compute the Mean Squared Error (MSE) of the regression coefficient estimators; and (d) Compare various widely-used overall and individual collinearity measures. The commonly utilized comprehensive measures include the Farrar-Glauber (FG) test, Determinant of the matrix *X'X* (DE), Red Indicator (RI), Sum of Reciprocals of eigenvalues (SR), Theil Indicator (TI) and Condition Number (CN). On the other hand, the prevalent individual measures consist of *VIF*, Tolerance Limit (TL), *WI* and *FI* statistics, Leamer Indicator (LI), Corrected VIF (CVIF) and Klein Indicator (KI). It should be noted that the standard indices mentioned are implemented in the *R* package mctest. For more comprehensive information, please refer to the study conducted by Imdadullah *et al.* (2016) and the references provided therein. The simulation is grounded on the linear regression model, which is formally stated as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

where the random error ε is generated from the *N*(0,1) distribution.

Three covariates X_1, X_2 and X_3, where \$$X_3 = kX_2$, with $k \in \{1/4, 2\}$ were considered. Three distributions are used to generate X_1 and X_2: uniform, normal and exponential. We set $\beta_0 = 0$, $\beta_1 = 1$, $\beta_2 = 1$ and $\beta_3 = 1$, considering

10000 simulations and six sample sizes: $n \in \{7, 10, 20, 30, 50, 100\}$. Furthermore, we assume the linear model with heteroscedastic and homoscedastic errors.

Simulation II

A second simulation study is conducted to consider the effects of outlier contamination on the percentage of correctly identified collinearity cases by two of the current indicators, overall measure *FG* and individual measure *FI*.

In the scenario, a linear model includes three covariates, labeled X_1, X_2 and X_3,. The first two covariates, X_1 and X_2, originate from a normal distribution. The third covariate, X_3, is defined as a multiple of kX_2, specifically with $k \in \{1/4, 2\}$ were considered.

The random errors ε, are generated from the $N(0,1)$ distribution but are contaminated at random with 5, 10, 15 and 20% of outliers which are generated from the $N(0,4)$ distribution. The simulations are carried out for sample size $n \in \{7, 10, 20, 30, 50, 100\}$.

Application to Real-World Data Sets

Corn Data

To manage corn production, it is important to estimate the yield potential. To do this, the grain yield, Y, is considered as a function of the covariates: Distance between rows, X_1, number of corncobs per m^2, X_2 and number of grains per corncob, X_3. The objective is to build a model with the yield of corn as the response and using the other measurements as covariates. The fitted model can then be used to predict corn yield in future years.

Fig. 1: *Cachama* (*Colossoma macropomum*)

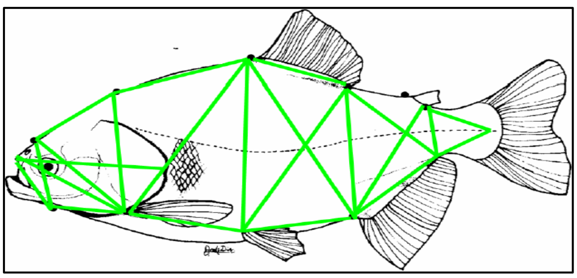

Fig. 2: Landmarks used for extracting truss measurements from *C. macropomum*

Table 1: Truss measurements from *C. macropomum* specimens

Tip of snout to end of epiphyseal sulcus
Tip of snout to insertion of pectoral fin
Anterior edge of the epiphyseal sulcus to the end of the epiphyseal sulcus
Anterior edge of the epiphyseal sulcus at the insertion of the pectoral fin
Anterior edge of the epiphyseal sulcus when articulating
Articulate to insertion of pectoral fin
Posterior edge of epiphyseal sulcus to end of dorsal fin
Posterior edge of the epiphyseal sulcus at the insertion of the pelvic fin
Posterior edge of the epiphyseal sulcus to the insertion of the pectoral fin
Posterior edge of the epiphyseal groove when articulating
Insertion of pectoral fin to insertion of pelvic fin
Dorsal fin base
Anterior edge of dorsal fin to anterior edge of anal fin
Anterior edge of dorsal fin to insertion of pelvic fin
Anterior edge of dorsal fin to insertion of pectoral fin
Insertion of pelvic fin to end of anal fin
Posterior edge of dorsal fin to the fatty fin
Posterior edge of dorsal fin to posterior edge of anal fin
Posterior edge of dorsal fin to anterior edge of anal fin
Posterior edge of dorsal fin to insertion of pelvic fin
Anal fin base
Posterior edge of the fatty fin to the last scale of the lateral line
Posterior edge of fatty fin to posterior edge of anal fin
Posterior edge of the fatty fin to the anterior border of the anal fin
Posterior edge of the fatty fin to the anterior border of the anal fin
Eye diameter
Head length
Fat fin base

Fish Morphology

The present study involved the analysis of 92 specimens of *Colossoma macropomum* (refer to Fig. 1) obtained from artificial ponds located at the Papelón fish station in Venezuela. The specimens had an average weight of 600 g. The study employed the "Truss protocol" or "trusses" approach proposed by Strauss and Bookstein (1982). This method enables a comprehensive reconstruction of the shape by utilizing the distances between homologous anatomical landmarks, as presented in Table 1 and Fig. 2. The landmarks are connected by distances that create a sequence of uninterrupted quadrilaterals, each with its own internal diagonals (refer to Fig. 2). This arrangement enables the identification of variations in shape along the vertical, horizontal and oblique orientations.

Results and Discussion

Simulation I

Tables 2-7 report the percentage of correctly identified collinearity cases for the overall and individual collinearity measures, whereas Table 8 presents the empirical mses of the estimators for the regression coefficients. From Tables 2-7, using uniform, normal and exponential distributions for X_1 and X_2 and with heteroscedastic and homoscedastic errors, note that for the FG (overall) and Fi (individual) collinearity indices, the percentage of cases of collinearity correctly identified exceeds the values for all the other measures and that the percentage increases as n increases.

In Table 8, using uniform, normal and exponential distributions for X_1 and X_2, observe that the empirical MSE of the estimators of the regression coefficients decreases as the sample size increases, which shows the empirical consistency of the OLS estimators of the regression coefficients. The three scenarios considered (uniform, normal and exponential distributions) produce very particular results in relation to the empirical MSE of $\hat{\beta}_3$, which measures the effect of the covariate X_3, expressed as a linear combination of X_1 and X_2. This estimator ($\hat{\beta}_3$) has an MSE close to zero, in addition to being the smallest in comparison to the MSE of the other three estimators ($\hat{\beta}_0$, $\hat{\beta}_1$and $\hat{\beta}_2$). In summary, this simulation study quantifies the effect of the degree of collinearity on the collinearity measures. In particular, the FG and Fi collinearity indices more robust as both sample size and collinearity degree increase. This is a major advantage since collinearity is a matter of degree and not simply presence or absence of collinearity. Likewise, the results show the superiority of these indices compared to the other used measures.

Table 2: Percentage of correctly identified Collinearity cases, for various values of k and n, where X_1 and X_2 follow uniform distributions, $X_3 = kX_2$ and with heteroscedastic errors

$X_3 = kX_2$	Collinearity measurement	Index or test	% of correct collinearity n = 7	n = 10	n = 20	n = 30	n = 50	n = 100
$k = 1/4$	Overall	FG	0.0057	0.0353	0.1527	0.2677	0.4821	0.8385
		Det	0.0005	0.0000	0.0000	0.0000	0.0000	0.0000
		Red Ind	0.2621	0.1148	0.0125	0.0013	0.0000	0.0000
		Sum lambda	0.0494	0.0061	0.0000	0.0000	0.0000	0.0000
		Theil	0.0422	0.0066	0.0000	0.0000	0.0000	0.0000
		CN	0.0687	0.0161	0.0003	0.0000	0.0000	0.0000
	Individual	VIF	0.0105	0.0009	0.0000	0.0000	0.0000	0.0000
		TOL	0.0105	0.0009	0.0000	0.0000	0.0000	0.0000
		Wi	0.0095	0.0037	0.0009	0.0006	0.0013	0.0082
		Fi	0.2411	0.3411	0.6017	0.7543	0.9132	0.9940
		Leamer	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
		CVIF	0.0132	0.0168	0.0295	0.0356	0.0441	0.0495
		Klein	0.0152	0.0024	0,000	0.0441	0.0495	0.0000
$k = 2$	Overall	FG	0.3348	0.9734	1.0000	1.0000	1.0000	1.0000
		Det	0.0581	0.0078	0.0000	0.0000	0.0000	0.0000
		Red Ind	0.9822	0.9888	0.9883	0.9997	1.0000	1.0000
		Sum lambda	0.8534	0.8452	0.8698	0.8819	0.9296	0.9737
		Theil	0.1886	0.0621	0.0022	0.0000	0.0000	0.0000
		CN	0.9970	0.9989	1.0000	1.0000	1.0000	1.0000
	Individual	VIF	0.6819	0.6281	0.5488	0.5065	0.4552	0.3862
		TOL	0.6819	0.6281	0.5488	0.5065	0.4552	0.3862
		Wi	0.6536	0.8497	0.9990	1.0000	1.0000	1.0000
		Fi	0.9943	0.9999	1.0000	1.0000	1.0000	1.0000
		Leamer	0.0295	0.0037	0.0000	0.0000	0.0000	0.0000
		CVIF	0.0009	0.0002	0.0001	0.0000	0.0000	0.0000
		Klein	0.1386	0.1052	0.0356	0.0130	0.0013	0.0000

Table 3: Percentage of correctly identified Collinearity cases, for various values of k and n, where X_1 and X_2 follow uniform distributions, $X_3 = kX_2$ and with homoscedastic errors

$X_3 = kX_2$	Collinearity measurement	Index or test	% of correct collinearity n = 7	n = 10	n = 20	n = 30	n = 50	n = 100
$k = 1/4$	Overall	FG	0.0052	0.0242	0.0862	0.1361	0.2369	0.5006
		Det	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000
		Red Ind	0.2202	0.0836	0.0067	0.0009	0.0000	0.0000
		Sum lambda	0.0421	0.0047	0.0000	0.0000	0.0000	0.0000
		Theil	0.0518	0.0119	0.0002	0.0000	0.0000	0.0000
		CN	0.0192	0.0024	0.0000	0.0000	0.0000	0.0000
	Individual	VIF	0.0073	0.0006	0.0000	0.0000	0.0000	0.0000
		TOL	0.0073	0.0006	0.0000	0.0000	0.0000	0.0000
		Wi	0.0065	0.0027	0.0002	0.0000	0.0002	0.0001
		Fi	0.1939	0.2541	0.4412	0.5643	0.7347	0.9196
		Leamer	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
		CVIF	0.0177	0.0214	0.0324	0.0388	0.0449	0.0619
		Klein	0.0230	0.0060	0.0001	0.0000	0.0000	0.0000
$k = 2$	Overall	FG	0.1401	0.8255	1.0000	1.0000	1.0000	1.0000
		Det	0.0177	0.0017	0.0000	0.0000	0.0000	0.0000
		Red Ind	0.9157	0.9084	0.9085	0.9122	0.9331	0.9684
		Sum lambda	0.5979	0.4828	0.2997	0.2087	0.1133	0.0313
		Theil	0.1618	0.0555	0.0023	0.0000	0.0000	0.0000
		CN	0.8662	0.8411	0.8365	0.8374	0.8769	0.9214
	Individual	VIF	0.3787	0.2427	0.0702	0.0292	0.0040	0.0000
		TOL	0.3787	0.2427	0.0702	0.0292	0.0040	0.0000
		Wi	0.3506	0.4866	0.9282	0.9963	1.0000	1.0000
		Fi	0.9641	0.9970	1.0000	1.0000	1.0000	1.0000
		Leamer	0.0077	0.0004	0.0000	0.0000	0.0000	0.0000
		CVIF	0.0063	0.0025	0.0001	0.0000	0.0000	0.0000
		Klein	0.1274	0.0935	0.0286	0.0097	0.0008	0.0000

Table 4: Percentage of correctly identified Collinearity cases, for various values of k and n, where X_1 and X_2 follow normal distributions, $X_3 = kX_2$ and with heteroscedastic errors

$X_3 = kX_2$	Collinearity measurement	Index or test	% of correct collinearity n = 7	n = 10	n = 20	n = 30	n = 50	n = 100
$k = 1/4$	Overall	FG	0.0248	0.2409	0.8410	0.9745	0.9994	1.0000
		Det	0.0032	0.0000	0.0000	0.0000	0.0000	0.0000
		Red Ind	0.5676	0.4372	0.2055	0.1054	0.0296	0.0008
		Sum lambda	0.1798	0.0635	0.0031	0.0006	0.0000	0.0000
		Theil	0.0531	0.0074	0.0001	0.0000	0.0000	0.0000
		CN	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000
	Individual	VIF	0.0724	0.0180	0.0003	0.0000	0.0000	0.0000
		TOL	0.0724	0.0180	0.0003	0.0000	0.0000	0.0000
		Wi	0.0663	0.0574	0.1277	0.2829	0.7156	0.9984
		Fi	0.6328	0.8448	0.9920	0.9998	1.0000	1.0000
		Leamer	0.0013	0.0000	0.0000	0.0000	0.0000	0.0000
		CVIF	0.0030	0.0018	0.0008	0.0002	0.0001	0.0000
		Klein	0.0049	0.0002	0.0000	0.0000	0.0000	0.0000
$k = 2$	Overall	FG	0.9335	0.9998	1.0000	1.0000	1.0000	1.0000
		Det	0.5883	0.4866	0.2877	0.0230	0.0991	0.0245
		Red Ind	0.9997	0.9999	1.0000	1.0000	1.0000	1.0000
		Sum lambda	0.9974	0.9995	1.0000	1.0000	1.0000	1.0000
		Theil	0.2354	0.0804	0.0027	0.0002	0.0000	0.0000
		CN	0.2376	0.1287	0.0275	0.0067	0.0003	0.0000
	Individual	VIF	0.9970	0.9977	0.9999	1.0000	1.0000	1.0000
		TOL	0.9970	0.9977	0.9999	1.0000	1.0000	1.0000
		Wi	0.9895	0.9995	1.0000	1.0000	1.0000	1.0000
		Fi	0.9998	1.0000	1.0000	1.0000	1.0000	1.0000
		Leamer	0.4959	0.4111	0.2436	0.1746	0.0842	0.0219
		CVIF	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
		Klein	0.1533	0.1063	0.0370	0.0168	0.0014	0.0000

Table 5: Percentage of correctly identified Collinearity cases, for various values of k and n, where X_1 and X_2 follow normal distributions, $X_3 = kX_2$ and with homoscedastic errors

$X_3 = kX_2$	Collinearity measurement	Index or test	% of correct collinearity n = 7	n = 10	n = 20	n = 30	n = 50	n = 100
$k = 1/4$	Overall	FG	0.0043	0.0234	0.0844	0.1374	0.2411	0.4974
		Det	0.0003	0.0000	0.0000	0.0000	0.0000	0.0000
		Red Ind	0.2238	0.0833	0.0059	0.0006	0.0000	0.0008
		Sum lambda	0.0447	0.0034	0.0000	0.0000	0.0000	0.0000
		Theil	0.0555	0.0130	0.0001	0.0000	0.0000	0.0000
		CN	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	Individual	VIF	0.0079	0.0004	0.0000	0.0000	0.0000	0.0000
		TOL	0.0079	0.0004	0.0000	0.0000	0.0000	0.0000
		Wi	0.0072	0.0014	0.0000	0.0003	0.0002	0.0004
		Fi	0.1918	0.2526	0.4338	0.5633	0.7420	0.9201
		Leamer	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000
		CVIF	0.0195	0.0217	0.0319	0.0373	0.0418	0.0620
		Klein	0.0277	0.0063	0.0000	0.0000	0.0000	0.0000
$k = 2$	Overall	FG	0.1396	0.7817	0.9994	1.0000	1.0000	1.0000
		Det	0.0187	0.0011	0.0000	0.0000	0.0000	0.0000
		Red Ind	0.8805	0.8795	0.8773	0.8887	0.9102	0.9465
		Sum lambda	0.5559	0.4653	0.3052	0.2232	0.1305	0.0433
		Theil	0.1631	0.0532	0.0018	0.0002	0.0000	0.0000
		CN	0.0147	0.0011	0.0000	0.0000	0.0000	0.0000
	Individual	VIF	0.3514	0.2446	0.0912	0.0407	0.0081	0.0001
		TOL	0.3514	0.2446	0.0912	0.0407	0.0081	0.0001
		Wi	0.3277	0.4657	0.8842	0.9918	1.0000	1.0000
		Fi	0.9385	0.9935	1.0000	1.0000	1.0000	1.0000
		Leamer	0.0090	0.0009	0.0000	0.0000	0.0000	0.0000
		CVIF	0.0081	0.0058	0.0012	0.0002	0.0000	0.0000
		Klein	0.1293	0.0862	0.0261	0.0105	0.0008	0.0000

Table 6: Percentage of correctly identified Collinearity cases, for various values of k and n, where X_1 and X_2 follow exponential distributions, $X_3 = kX_2$ and with heteroscedastic errors

$X_3 = kX_2$	Collinearity measurement	Index or test	% of correct collinearity n = 7	n = 10	n = 20	n = 30	n = 50	n = 100
$k = 1/4$	Overall	FG	0.0033	0.0175	0.0395	0.0462	0.0510	0.0675
		Det	0.0003	0.0000	0.0000	0.0000	0.0000	0.0000
		Red Ind	0.1988	0.0582	0.0018	0.0002	0.0000	0.0008
		Sum lambda	0.0389	0.0030	0.0000	0.0000	0.0000	0.0000
		Theil	0.1004	0.0315	0.0012	0.0000	0.0000	0.0000
		CN	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
	Individual	VIF	0.0044	0.0001	0.0000	0.0000	0.0000	0.0000
		TOL	0.0044	0.0001	0.0000	0.0000	0.0000	0.0000
		Wi	0.0038	0.0007	0.0000	0.0000	0.0002	0.0004
		Fi	0.1553	0.1682	0.2513	0.2881	0.3397	0.4255
		Leamer	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
		CVIF	0.0312	0.0268	0.0200	0.0101	0.0031	0.0002
		Klein	0.0663	0.0294	0.0027	0.0000	0.0001	0.0000
$k = 2$	Overall	FG	0.0076	0.0608	0.2801	0.4617	0.7446	0.9763
		Det	0.0010	0.0000	0.0000	0.0000	0.0000	0.0000
		Red Ind	0.3127	0.1548	0.0294	0.0073	0.0003	0.0000
		Sum lambda	0.0727	0.0106	0.0001	0.0000	0.0000	0.0000
		Theil	0.1174	0.0411	0.0032	0.0001	0.0000	0.0000
		CN	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
	Individual	VIF	0.0185	0.0015	0.0000	0.0000	0.0000	0.0000
		TOL	0.0185	0.0015	0.0000	0.0000	0.0000	0.0000
		Wi	0.0156	0.0061	0.0057	0.0087	0.0211	0.1567
		Fi	0.3126	0.4388	0.7391	0.8787	0.9728	0.9996
		Leamer	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000
		CVIF	0.0337	0.0405	0.0502	0.0553	0.0613	0.0676
		Klein	0.0887	0.0534	0.0118	0.0034	0.0001	0.0000

DOI: 10.3844/ojbsci.2024.195.207

Table 7: Percentage of correctly identified Collinearity cases, for various values of k and n, where X_1 and X_2 follow exponential distributions, $X_3 = kX_2$ and with homoscedastic errors

$X_3 = kX_2$	Collinearity measurement	Index or test	% of correct collinearity n = 7	n = 10	n = 20	n = 30	n = 50	n = 100
$k = 1/4$	Overall	FG	0.0045	0.0250	0.0848	0.1368	0.2364	0.4937
		Det	0.0004	0.0000	0.0000	0.0000	0.0000	0.0000
		Red Ind	0.2308	0.0791	0.0072	0.0009	0.0000	0.0000
		Sum lambda	0.0467	0.0041	0.0000	0.0000	0.0000	0.0000
		Theil	0.0576	0.0109	0.0002	0.0000	0.0000	0.0000
		CN	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
	Individual	VIF	0.0066	0.0002	0.0000	0.0000	0.0000	0.0000
		TOL	0.0066	0.0002	0.0000	0.0000	0.0000	0.0000
		Wi	0.0056	0.0012	0.0005	0.0002	0.0002	0.0008
		Fi	0.2001	0.2490	0.4361	0.5461	0.7195	0.9141
		Leamer	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
		CVIF	0.0204	0.0229	0.0284	0.0360	0.0419	0.0579
		Klein	0.0283	0.0069	0.0000	0.0000	0.0000	0.0000
$k = 2$	Overall	FG	0.1465	0.6695	0.9901	0.9997	1.0000	1.0000
		Det	0.0253	0.0024	0.0000	0.0000	0.0000	0.0000
		Red Ind	0.8083	0.7887	0.7723	0.7793	0.8020	0.8554
		Sum lambda	0.4885	0.3990	0.2990	0.2527	0.1782	0.1022
		Theil	0.1611	0.0499	0.0026	0.0002	0.0000	0.0000
		CN	0.0028	0.0001	0.0000	0.0000	0.0000	0.0000
	Individual	VIF	0.3219	0.2332	0.1287	0.0800	0.0314	0.0062
		TOL	0.3219	0.2332	0.1287	0.0800	0.0314	0.0062
		Wi	0.3030	0.3965	0.7696	0.9447	0.9992	1.0000
		Fi	0.8621	0.9655	0.9998	1.0000	1.0000	1.0000
		Leamer	0.0129	0.0011	0.0000	0.0000	0.0000	0.0000
		CVIF	0.0133	0.0092	0.0039	0.0020	0.0002	0.0000
		Klein	0.1229	0.0870	0.0264	0.0095	0.0008	0.0000

Table 8: Empirical MSE of the indicated parameter estimator in a regression model, using the specified n and a distribution for X_1 and X_2

	Uniform				Normal				Exponential			
n	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
7	1.94	0.64	0.69	0.17	2.45	0.70	0.34	0.15	0.81	0.78	1.11	0.14
10	1.35	0.58	0.62	0.12	1.88	0.37	0.28	0.11	0.55	0.41	0.82	0.11
20	0.78	0.35	0.35	0.06	1.10	0.21	0.16	0.06	0.36	0.23	0.47	0.06
30	0.59	0.28	0.27	0.05	0.83	0.16	0.13	0.05	0.27	0.18	0.34	0.05
50	0.45	0.20	0.22	0.04	0.67	0.12	0.10	0.03	0.20	0.13	0.25	0.03
100	0.23	0.09	0.10	0.02	0.46	0.03	0.06	0.01	0.14	0.02	0.14	0.01

Table 9: Percentage of correctly identified collinearity cases in a linear modelo contaminated with the indicated percentage of outliers, for various values of k and n, where X_1 and X_2 follow normal distributions, $X_3 = kX_2$ and with homoscedastic errors

% Outlier	$X_3 = kX_2$	Collinearity measurement	Index or test	% of correct collinearity n = 7	n = 10	n = 20	n = 30	n = 50	n = 100
5	$k = 1/4$	Overall	FG	0.0042	0.0244	0.0832	0.1427	0.2431	0.4933
		Individual	Fi	0.1986	0.2644	0.4361	0.5696	0.7460	0.9214
	$k = 2$	Overall	FG	0.1432	0.7867	0.9996	1.0000	1.0000	1.0000
		Individual	Fi	0.9427	0.9937	1.0000	1.0000	1.0000	1.0000
10	$k = 1/4$	Overall	FG	0.0042	0.0244	0.0832	0.1427	0.2431	0.4933
		Individual	Fi	0.1986	0.2644	0.4361	0.5696	0.7460	0.9214
	$k = 2$	Overall	FG	0.1432	0.7867	0.9996	1.0000	1.0000	1.0000
		Individual	Fi	0.9427	0.9937	1.0000	1.0000	1.0000	1.0000
15	$k = 1/4$	Overall	FG	0.0042	0.0244	0.0832	0.1427	0.2431	0.4933
		Individual	Fi	0.1986	0.2644	0.4361	0.5696	0.7460	0.9214
	$k = 2$	Overall	FG	0.1432	0.7867	0.9996	1.0000	1.0000	1.0000
		Individual	Fi	0.9427	0.9937	1.0000	1.0000	1.0000	1.0000
20	$k = 1/4$	Overall	FG	0.0042	0.0244	0.0832	0.1427	0.2431	0.4933
		Individual	Fi	0.1986	0.2644	0.4361	0.5696	0.7460	0.9214
	$k = 2$	Overall	FG	0.1432	0.7867	0.9996	1.0000	1.0000	1.0000
		Individual	Fi	0.9427	0.9937	1.0000	1.0000	1.0000	1.0000

Table 10: Collinearity diagnostics in a linear model with corn data

Collinearity	Index or test	Value	p-value
Overall	FG	60.2073	< .000001
	Det	0.0197	NS
	Red Ind	0.8520	*
	Sum lambda	41.7184	*
	Theil	0.6679	*
	CN	7.0974	NS
Individual	F_1	38.5740	.0058698
	F_2	370.6573	.0002019
	F_3	287.6991	.0002951

* and p<.05 (Collinearity identified); ns (unidentified collinearity)

Table 11: Overall collinearity diagnosis in patterns of morphological covariance in *C. macropomum* species

Index or test	Collinearity diagnosis
Determinante	*
Farrar-Glauber	*
Red indicator	*
Suma de Lambda	*
Theil indicator	NS
Número de codición	*

* (Collinearity identified); NS (unidentified collinearity)

Table 12: Individual collinearity diagnosis in patterns of morphological covariance in *C. macropomum* species

Landmarks	VIF	F_j
Tip of snout to end of epiphyseal sulcus	*	*
Tip of snout to insertion of pectoral fin	*	*
Anterior edge of the epiphyseal sulcus to the end of the epiphyseal sulcus	*	*
Anterior edge of the epiphyseal sulcus at the insertion of the pectoral fin	*	*
Anterior edge of the epiphyseal sulcus when articulating	*	*
Articulate to insertion of pectoral fin	*	*
Posterior edge of epiphyseal sulcus to end of dorsal fin	*	*
Posterior edge of the epiphyseal sulcus at the insertion of the pelvic fin	*	*
Posterior edge of the epiphyseal sulcus to the insertion of the pectoral fin	NS	*
Posterior edge of the epiphyseal groove when articulating	*	*
Insertion of pectoral fin to insertion of pelvic fin	*	*
Dorsal fin base	*	*
Anterior edge of dorsal fin to anterior edge of anal fin	NS	*
Anterior edge of dorsal fin to insertion of pelvic fin	*	*
Anterior edge of dorsal fin to insertion of pectoral fin	NS	*
Insertion of pelvic fin to end of anal fin	NS	*
Posterior edge of dorsal fin to the fatty fin	*	*
Posterior edge of dorsal fin to posterior edge of anal fin	*	*
Posterior edge of dorsal fin to anterior edge of anal fin	NS	*
Posterior edge of dorsal fin to insertion of pelvic fin	*	*
Anal fin base	*	*
Posterior edge of the fatty fin to the last scale of the lateral line	NS	*
Posterior edge of fatty fin to posterior edge of anal fin	*	*
Posterior edge of the fatty fin to the anterior border of the anal fin	*	*
Posterior edge of the fatty fin to the anterior border of the anal fin	*	*
Eye diameter	*	*
Head length	*	*
Fat fin base	NS	*

* (Collinearity identified); NS (unidentified collinearity)

Simulations II

The results, shown in Table 9, do not provide any evidence for an effect from outliers on collinearity identification using the collinearity indices (individual and overall) since, as the proportion of outliers increases, the percentage of collinearity cases correctly identified remains stable. In summary, the results show the robustness of the FG (overall) and Fi (individual) collinearity indices in presence of outliers.

Aplication to Real-World Data Sets

Corn Data

Figure 3 displays scatter-plots for all the variables and their corresponding correlations. This figure indicates that Y has moderate or high correlation with each covariate, suggesting that a multiple linear regression model is suitable. However, high correlations are also found between some covariates, indicating the likely presence of collinearity. Table 10 shows the corresponding values of the collinearity diagnostics. We include the currently used general collinearity measures and an individual collinearity measure F_j. The FG test, red indicator, sum lambda and Theil confirming the presence of collinearity. Similarly, since F_i provides p<.01 for each covariate: X_1, X_2 and X_3 it is assumed that these covariates are collinear, as indicated by Farrar and Glauber (1967). This allows us to infer that the three covariates are involved in one or more linear dependency relationships between them. When comparing the indices FG and F_i with the other measures, note that are shown as powerful tools for the study of collinearity, since they verify the presence of collinearity and at the same time identify whether a covariate is collinear or not.

Morphology Fish (C. macropomun)

Table 11, most of the diagnostic measures, except for the Theil indicator, identify that there are redundant characteristics associated with morphological covariation patterns in *C. macropomum* specimens, that is, there is multicollinearity, which can contribute to the entropy of the models used to identify patterns of morphological covariation of this species. Table 12, $VIFs$ can modify most of the distances measured on the lateral profile of these examples are attributed to redundant morphological characteristics. Only morphological characteristics tales like; posterior edge of the epiphyseal sulcus at the insertion of the fin pectoral variables, anterior edge of the dorsal fin to the anterior edge of the anal fin, anterior edge of the dorsal fin at the insertion of the pectoral fin, insertion of the pelvic fin at the anterior edge anal fin, posterior edge of dorsal fin to anterior edge of anal fin, posterior edge of fat fin to last scale of lateral line and base of fat fin not direct redundant morphological information, saber, not son causing multicollinearity (Fig. 4). These variables are associated with morphological covariation patterns that make the difference in the head area, in the area of the bases of the fins of the abdomen and in the anterior part of the fish. The results of the farra-glauber test (individual diagnostic measure of multicollinearity) do not perform well in relation to the identification of the origin of multicollinearity, since it is not capable of identifying non-redundant covariates associated with the morphology of the examples *C. macropomum*.

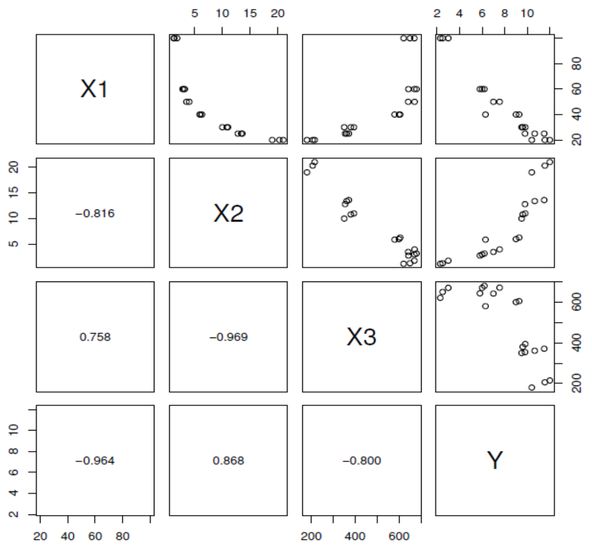

Fig. 3: Scatter-plots and their correlations for the indicated variables with corn data

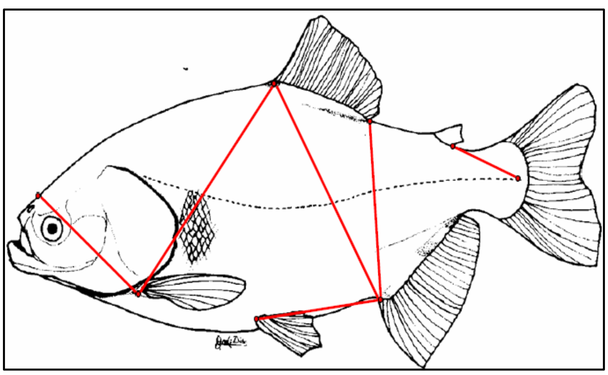

Fig. 4: Non-redundant covariates (landmarks) in the truss protocol on *C. macropomum*

Conclusion

The results do not provide any evidence for an effect from outliers on collinearity identification using the collinearity indices (individual and overall). The FG and F_i collinearity indices more robust as both sample size and collinearity degree increase. On the fitted models on corn data and fish morphology the most of overall collinearity indices confirmed the presence of collinearity. However, the *VIF* (individual measure) had a better performance on the fitted model on the morphology of *C. macropomum.* These results suggest an effect of the number of model parameters (p) on the performance of the collinearity indices (individual and general), therefore a more exhaustive study that considers models with a greater number of parameters is recommended.

Acknowledgment

The authors express their gratitude to the four reviewers for their insightful feedback and recommendations, which have undoubtedly enhanced the clarity and caliber of this manuscript.

Funding Information

The authors have not received any financial support or funding to report.

Author's Contributions

All authors equally contributed in this study.

Ethics

All the protocols of ethical research conduct were strictly adhered to throughout the study and there are no conflicts of interest to report.

References

Asteriou, D., & Hall, S. G. (2007). Applied Econometrics: a modern approach, revised edition. *Hampshire: Palgrave Macmillan*, *46*(2), 117-155.

Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). The condition number. *Regression diagnostics: Identifying Influential Data and Sources of Collinearity*, *100*, 104. https://doi.org/10.1002/0471725153

Chennamaneni, P. R., Echambadi, R., Hess, J. D., & Syam, N. (2016). Diagnosing harmful collinearity in moderated regressions: A roadmap. *International Journal of Research in Marketing*, *33*(1), 172-182. https://doi.org/10.1016/j.ijresmar.2015.08.004

Chatterjee, S., & Hadi, A. S. (2013). *Regression analysis by example*. John Wiley and Sons. https://doi.org/10.1002/0470055464

Dillon, W. R., & Goldstein, M. (1984). *Multivariate analysis: Methods and applications*. *New York (NY): Wiley*, 1984. https://lib.ugent.be/catalog/rug01:000081114

Chatterjee, S. & Price, B. (1977). Regression Analysis by Example. New York: *John Wiley and Sons, Inc.,* pp.19-22. https://doi.org/10.12691/aees-5-2-4

Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics*, *19*(1), 15-18. https://doi.org/10.2307/1268249

Cooley, W. W A., & Lohnes, P. R. A (1971): Multivariate Data Analysis. *John Wiley and Sons,* Australia.

Core Team, R. (2016). The R Project for Statistical Computing. *R Foundation for Statistical Computing, Vienna, Austria*. https://www.r-project.org/

Curto, J. D., & Pinto, J. C. (2011). The corrected vif (cvif). *Journal of Applied Statistics*, *38*(7), 1499-1507. https://doi.org/10.1080/02664763.2010.505956

Farrar, D. E., & Glauber, R. R. (1967). Multicollinearity in regression analysis: the problem revisited. *The Review of Economic and Statistics*, 92-107. https://doi.org/10.2307/1937887

Fox, J., & Monette, G. (1992). Generalized collinearity diagnostics. *Journal of the American Statistical Association*, *87*(417), 178-183. https://doi.org/10.2307/2290467

Greene, W. H (2002): Econometric Analysis. *Prentic-Hall, New Jersey, 5th Ed*. 2002. https://www.ctanujit.org/uploads/2/5/3/9/25393293/econometric_analysis_by_greence.pdf

Gujarati, D. N., & Porter, D. C. (2003). Basic Econometrics, McGraw-Hill. *New York*.

Gunst, R. F. (1984). Comment: Toward a balanced assessment of collinearity diagnostics. *The American Statistician*, *38*(2), 79-82. https://doi.org/10.2307/2683238

Habshah, M., Norazan, M. R., & Rahmatullah Imon, A. H. M. (2009). The performance of diagnostic-robust generalized potentials for the identification of multiple high leverage points in linear regression. *Journal of Applied Statistics*, *36*(5), 507-520. https://doi.org/10.1080/02664760802553463

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2014). An analysis of mental wellbeing and mental psychopathology of gifted students by an emotional and social developmental process perspective.

Haitovsky, Y. (1969). Multicollinearity in regression analysis: Comment. *The Review of Economics and Statistics*, 486-489.https://doi.org/10.2307/1926450

Imdadullah, M., Aslam, M., & Altaf, S. (2016). Mc test: An R Package for Detection of Collinearity among Regressors. *R J.*, *8*(2), 495. https://doi.org/10.32614/RJ-2016-062

Kendall, M. G. (1957). A Course in Multivariate Analysis. *Griffin, London*. https://doi.org/10.1017/S0020268100037422

Klein, L. R. (1969). Estimation of interdependent systems in macroeconometrics. *Econometrica: Journal of Econometric Society*, 171-192. https://doi.org/10.2307/1913529

Kovács, P., Petres, T., & Tóth, L. (2005). A new measure of multicollinearity in linear regression models. *International Statistical Review*, *73*(3), 405-412. https://www.jstor.org/stable/25472683

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied Linear Statistical Models*. McGraw-hill. https://thuvienso.hoasen.edu.vn/handle/123456789/9564

Mason, R. L., & Gunst, R. F. (1985). Outlier-induced collinearities. *Technometrics*, *27*(4), 401-407. https://doi.org/10.2307/1270207

Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (2004). *Applied Linear Statistical Models*. ISBN-13: 978-0-07-310874-2.

O'brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. *Quality & quantity*, *41*, 673-690. https://doi.org/10.1007/s11135-006-9018-6

O'Hagan, J., & McCabe, B (1975): Tests for the severity of multicollinearity in regression analysis: A comment. The Review of Economics and Statistics, 57:368-370. https://doi.org/10.2307/1923927

Oke, J., Akinkunmi, W. B., & Etebefia, S. O. (2019). Use of correlation, tolerance and variance inflation factor for multicollinearity test. *GSJ*, *7*(5). http://www.globalscientificjournal.com/researchpaper/use_of_correlation_tolerance_and_variance_inflation_factor_for_multicollinearity.pdf

Sengupta, D., & Bhimasankaram, P. (1997). On the roles of observations in collinearity in the linear model. *Journal of the American Statistical Association*, *92*(439), 1024-1032. https://doi.org/10.2307/2965567

Silvey, S. D. (1969). Multicollinearity and imprecise estimation. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *31*(3), 539-552. https://www.jstor.org/stable/2984357

Stock, J. H., & Watson, M. W. (2010). Modeling Inflation after the Crisis. *Introduction to Econometrics. Pearson Addison-Wesley*, 3rd Ed. https://www.princeton.edu/~mwatson/papers/stock-watson_frbkc_2010.pdf

Strauss, R. E., & Bookstein, F. L. (1982). The truss: Body form reconstructions in morphometrics. *Systematic Biology*, *31*(2), 113-135. 135. https://doi.org/10.2307/2413032

DOI: 10.3844/ojbsci.2024.195.207

Theil, H. (1971). An economic theory of the second moments of disturbances of behavioral equations. *The American Economic Review*, *61*(1), 190-194. https://www.jstor.org/stable/1910554

Walker, E., & Birch, J. B. (1988). Influence measures in ridge regression. *Technometrics*, *30*(2), 221-227. https://doi.org/10.2307/1270168

Wichers, C. R. (1975). The detection of multicollinearity: A comment. *The Review of Economics and Statistics*, 366-368. https://doi.org/10.2307/1923926

Wondola, D. W., Aulele, S. N., & Lembang, F. K. (2020, February). Partial least square (PLS) method of addressing multicollinearity problems in multiple linear regressions (case studies: Cost of electricity bills and factors affecting it). In *Journal of Physics: Conference Series* (Vol. *1463*, No. 1, p. 012006). IOP Publishing. https://doi.org/10.1088/1742-6596/1463/1/012006

Appendix 1

R code for collinearity diagnosis (individual and overall) in agricultural trials.

```
> library('mctest')
> x <- Data.morfometria[, -1]
> y <- Data.morfometria[, 1]
> omcdiag (x, y, detr = 0.001, red = 0.6, conf = 0.99,
theil = 0.6, cn = 15)
> omcdiag (x, y, Inter = FALSE)
> omcdiag (x, y)
> imcdiag(x, y, corr = TRUE)
> imcdiag(x, y)
```