MTest: a Bootstrap Test for Multicollinearity

Morales-Oñate, Victor [1,2,*] ; **Morales Oñate, Bolívar** [3]

[1] *Universidad de las Américas, Departamento de Economía, Quito, Ecuador*
[2] *Universidad San Francisco de Quito, Colegio de Administración y Economía, Quito, Ecuador*
[3] *Universidad Técnica de Ambato, Facultad de Ingeniería en Sistemas, Electrónica e Industrial, Ambato, Ecuador*

Abstract: A nonparametric test based on bootstrap for detecting multicollinearity is proposed: MTest. This test gives statistical support to two of the most famous methods for detecting multicollinearity in applied work: Klein's rule and Variance Inflation Factor (VIF for essential multicollinearity). As part of the procedure, MTest generates a bootstrap distribution for the coefficient of determination which: i) lets the researcher assess multicollinearity by setting a statistical significance α, or more precisely, an achieved significance level (ASL) for a given threshold, ii) using a pairwise Kolmogorov-Smirnov (KS) test, establishes a guide for an educated removal of variables that are causing multicollinearity. In order to show the benefits of MTest, the procedure is computationally implemented in a function for linear regression models. This function is tested in numerical experiments that match the expected results. Finally, this paper makes an application of MTest to real data known to have multicollinearity problems and successfully detects multicollinearity with a given ASL.

Keywords: MTest, Multicollinearity, NonParametric Statistics, Simulation

MTest: una Prueba *bootstrap* para Multicolinealidad

Resumen: Se propone una prueba no paramétrica basada en bootstrap para detectar multicolinealidad: MTest. Esta prueba brinda soporte estadístico a dos de los métodos más famosos para detectar multicolinealidad en trabajo aplicado: la regla de Klein y el Factor de Inflación de Varianza (VIF por multicolinealidad esencial). Como parte del procedimiento, MTest genera una distribución bootstrap para el coeficiente de determinación que: i) permite al investigador evaluar la multicolinealidad al establecer una significancia estadística α, o más precisamente, un nivel de significancia alcanzado (ASL) para un umbral dado, ii) utilizando una prueba de Kolmogorov-Smirnov (KS) por parejas, establece una guía para una eliminación informada de las variables que están causando multicolinealidad. Para mostrar los beneficios de MTest, el procedimiento se implementa computacionalmente en una función para modelos de regresión lineal. Esta función se prueba en experimentos numéricos que coinciden con los resultados esperados. Finalmente, este documento hace una aplicación de MTest a datos reales que se sabe que tienen problemas de multicolinealidad y detecta con éxito la multicolinealidad con un ASL dado.

Palabras clave: MTest, Multicolinealidad, Estadística no paramétrica, Simuluación

1. INTRODUCTION

When predictors of a regression model are correlated, multicollinearity appears. This may be a problem depending on the *degree* of correlation in the dataset which may be stated using the determinant of the predictors. If the predictors are linearly dependent, the determinant of the correlation matrix is equal to 0 (perfect multicollinearity); if the determinant is equal to 1, there is no multicollinearity (Stein, 1975).

Testing for multicollinearity has been studied from parametric approaches and *rule of thumb* proposals. Farrar and Glauber (1967) is a seminal work in the first case. They propose three different ways to do the test.

In the first one, if the determinant of the correlation of the matrix of the predictors is *almost* equal to 1, then there is no multicollinearity. It relies on having observations that come from an orthogonal, multivariate-normal distribution. This excludes the case when dummy variables are included. The second test they propose starts by computing the principal minors of the correlation matrix of the predictors. Each minor is divided by the determinant of the correlation of the matrix, this quotient has an F-distribution if the underlying distributions are normal. In the third case, they use the partial correlation coefficients, r_{ij} of the determinant of the correlation of the matrix of the predictors. r_{ij} are compared to their off-diagonal elements and use a t-test to make a comparison.

The first and second proposals in Farrar and Glauber (1967) share the fact that the underlying distributions are normal or multivariate-normal. This may be a limitation in real life applications where dummy variables, skewness and asymmetry are present in data. Then, a nonparametric approach can be useful to over-

victor.morales@uv.cl
Recibido: 21/12/2021
Aceptado: 20/01/2023
Publicado en línea: 01/05/2023
10.33333/rp.vol51n2.05

come this issue.

Some authors claim that the *problem of multicollinearity* should not be an issue (Leamer, 1983; Achen, 1982). They emphasize that even in the presence of multicollinearity, estimators keep being best linear unbiased estimator (BLUE) which is actually the case. So they say that the *real* problem is a matter of sample size. They claim that if the sample size is large enough, then multicollinearity would not be a problem. Sample size is sometimes restricted to the field of study. In economics, for example, time series of some indicators are not long enough and statistical modeling may be challenging. Nonetheless, this is changing with the availability of open source projects to access to economic data such as the World Development Indicators (The World Bank, 2021). Around the time authors like Leamer (1983) and Achen (1982) stated their concerns, the scientific community lacked of more data or computing power, so their claim needed careful attention. But things have change in regards to computing power and data.

With the emerging of the Big Data era, sample size is becoming less of an issue, but the problem of multicollinearity remains (Dinov, 2016). The problem is that the presence of multicollinearity makes the estimated variances and covariances inflated. In this context, wider confidence intervals are obtained which makes it easier not to reject the null hypothesis of the coefficients in the predictors. It follows that there may be one or more coefficients with no statistical significance but with a high coefficient of determination (Gujarati et al., 2012).

Jaya et al. (2020) make a comparison of different machine learning techniques in regression such as Ridge and Lasso to obtain the technique that avoids multicollinearity. In order to detect multicollinearity however, they use Variance Inflation Factors (VIF), which can be seen as a *rule of thumb* approach.

In R there are several packages that try to detect multicollinearity: Imdadullah et al. (2016) and Salmerón-Gómez et al. (2021b) are two of the most recent ones. Salmerón-Gómez et al. (2021b) propose a detection method based on a perturbation of the observations but they do not perturb dummy variables. Imdadullah et al. (2016) make a review of these methods and create an R package to make overall (determinant, R-squared, among others) and individual (Klein's rule, VIF, among others) multicollinearity diagnosis.

Klein's rule and VIF can be derived from a global and auxiliary coefficients of determination of a regressions model and a *rule of thumb* is applied. In this sense, both methods consider the coefficients of determination fixed when it can actually be considered a random variable (Carrodus and Giles, 1992; Koerts and Abrahamse, 1969). If by bootstrapping we let the global and auxiliary coefficients of determination be random variables, the detection of multicollinearity goes one step further. This work takes the two of the most widely used individual *rule of thumb* methods, Klein's rule and VIF, and places them in a bootstrap context in order to have a statistical test to assess multicollinearity. This proposal makes it possible to have a null hypothesis and an alternative hypothesis for the presence of multicollinearity, and depending on the significance level, the researcher could reach a conclusion.

This paper is organized as follows. In Section 2, we detail the VIF, Klein's rule, their relationship and recall the bootstrap concept. Section 3 states the null and alternative hypothesis of the proposed MTest with it corresponding procedure. In Section 4, we set up a simulation study to analize MTest under a controlled situation. In Section 5, we apply MTest to a widely used dataset known to have multicollinearity issues. Finally, in Section 6 we give some conclusions.

2. Notation and Concepts

2.1 General Setup

Consider the regression model

$$Y_i = \beta_0 X_{0i} + \beta_1 X_{1i} + \cdots + \beta_p X_{pi} + u_i \tag{1}$$

where $i = 1, \ldots, n$, $X_{j,i}$ are the predictors with $j = 1, \ldots, p$, $X_0 = 1$ for all i and u_i is the gaussian error term.

In order to describe Klein's rule and VIF methods, we need to define *auxiliary regressions* associated to model (1). An example of an auxiliary regressions is:

$$X_{2i} = \gamma_1 X_{1i} + \gamma_3 X_{3i} + \cdots + \gamma_p X_{pi} + u_i.$$

In general, there are p auxiliary regressions and the dependent variable is omitted in each auxiliary regression. Let R_g^2 be the coefficient of determination of (1) and R_j^2 the jth coefficient of determination of the jth auxiliary regression.

2.2 VIF

A common way practitioners compute the VIF method is by:

$$VIF_j = \frac{1}{1 - R_j^2} \tag{2}$$

for every auxiliary regression $j = 1, \ldots, p$. It states that multicollinearity is generated by covariate X_j if $VIF_j > 10$[1] (Gujarati et al., 2012). The case $j = 0$ is not considered since (2) detects approximate multicollinearity of the essential type. This means that the intercept is excluded. For more details and methods on detecting essential and non-essential multicollinearity, see Salmerón-Gómez et al. (2020).

2.3 Klein's rule

Klein's rule compares the R_j^2 coefficient of determination of the jth auxiliary regression with R_g^2. The rule states that if $R_j^2 > R_g^2$ then the X_j variable originates multicollinearity.

Klein stated that *Intercorrelation ... is not necessarily a problem unless it is high relative to the over-all degree of multiple correlation* ... (Klein, 1962). It means that there is an implicit threshold (R_g^2) and Klein's rule should be used along with VIF as they may be complementary since there will be cases when VIF report a multicollinearity problem, but Klein's rule does not.

2.4 Relationship between the VIF and Klein's rule

It is possible to link both methods by the expression

$$10 = \frac{1}{1 - R_j^2}$$

[1] Some literature also uses 3 or 5 as a threshold.

$$R_j^2 = 0.90$$

which means that according to the VIF method, variable X_j originates multicollinearity if $R_j^2 \geq 0.90$.
Different contradictions that may exist between the Klein's rule and the VIF. For example, if $R_g^2 = 0.85$ and some $R_j^2 = 0.88$, Klein's method would detect multicollinearity but VIF will not. As another example, if $R_j^2 = 0.94$ and $R_g^2 = 0.98$, VIF will detect multicollinearity but Klein will not.
It should be noted that the value 0.90 is not fixed in all applications. For example, works like Marcoulides and Raykov (2019) tests several values as a VIF threshold. Our proposal, MTest has 0.90 threshold by default but can be changed if the researcher decides another one.

2.5 Bootstrap

The bootstrap resampling technique was introduced by Efron (1992), in his seminal work *Bootstrap methods: Another look at the jackknife*. It is a computationally intensive method, without strict structural assumptions in the underlying random process that generates the data. It is used to obtain approximations of the distribution of an estimator, of the bias, the variance, standard error and confidence intervals. In a normal experiment, repeating the experiment enables us to compute standard errors, the bootstrap *principle* lets us *simulate* the replication by resampling.
If the resampling mechanism is chosen appropriately, then the resampling, together with the sample in question, is expected to reflect the original relationship between the population and the sample. The advantage is that we can now avoid the problem of having to deal with the population, and instead we can use the sample and resamples, to address statistical inference questions regarding the unknown quantities in the population. The bootstrap principle addresses the problem of not having complete knowledge of the population, to make an inference about the estimator $\hat{\theta}$, schematically:

- The first step consists of the construction of an estimator of an unknown probability distribution F, $F(\hat{F})$ from the available observations $X_1,\ldots,X_n$, which provides a representative image of the population[2].
- The next step consists of the generation of random variables $X_1^*,\ldots,X_n^*$ of the estimator $\hat{F}$, which fulfills the role of the sample for the bootstrap version of the original problem.

Therefore, the bootstrap version of the estimator $\hat{\theta}$ based on the original sample $X_1,\ldots,X_n$ is given by $\hat{\theta}^*$, obtained by substituting $X_1^*,\ldots,X_n^*$.
The above setup is known as the *nonparametric* bootstrap, but it also has a *parametric* approach where one can assume F belonging to a parametric model $\{F_\theta : \theta \in \Theta\}$ where Θ is the parameter space. In this case, $F = F_{\hat{\theta}}$ where $\hat{\theta}$ is an estimator of θ. For more details on parametric and nonparametric bootstrap, see Godfrey (2009).
It must be noted that the bootstrap also has some drawbacks. Horowitz (2001) highlights two important problems. The first one is instrumental variables estimation with ill correlated instruments and predictors. In this case, bootstrap approximations fail to be useful. The second problem is when the variance of the bootstrap estimator is high. This work does not fit in either of these cases since it is not related to instrumental variables and the variance of the bootstrap estimator is finite and well behaved.

[2]Let X be a random variable, F the cumulative distribution function and $\hat{F}$ is a non parametric functional estimator of F.

3. MTEST

Given a regression model, Mtest is based on computing estimates of R_g^2 and R_j^2 from n_{boot} bootstrap samples obtained from the dataset, $R_{g_{boot}}^2$ and $R_{j_{boot}}^2$ respectively.
Therefore, in the context of MTest, the VIF rule translates into:

$$H_0 : \mu_{R_{j_{boot}}^2} \geq 0.90,$$

and

$$H_a : \mu_{R_{j_{boot}}^2} < 0.90.$$

We seek an achieved significance level (ASL)

$$\text{ASL} = \text{Prob}_{H_0}\{\mu_{R_{j_{boot}}^2} \geq 0.90\}$$

estimated by

$$\widehat{\text{ASL}}_{n_{boot}} = \#\{\mu_{R_{j_{boot}}^2} \geq 0.90\}/n_{boot}$$

In a similar manner, the Klein's rule translates into:

$$H_0 : \mu_{R_{j_{boot}}^2} \geq \mu_{R_{g_{boot}}^2},$$

and

$$H_a : \mu_{R_{j_{boot}}^2} < \mu_{R_{g_{boot}}^2}.$$

We seek an achieved significance level

$$\text{ASL} = \text{Prob}_{H_0}\{\mu_{R_{j_{boot}}^2} \geq \mu_{R_{g_{boot}}^2}\}$$

estimated by

$$\widehat{\text{ASL}}_{n_{boot}} = \#\{\mu_{R_{j_{boot}}^2} \geq \mu_{R_{g_{boot}}^2}\}/n_{boot}.$$

It should be noted that this set up lets us formulate VIF and Klein's rules in terms of statistical hypothesis testing.

3.1 MTest: the algorithm

$R_{g_{boot}}^2$ and $R_{j_{boot}}^2$ are the distributions of R_g^2 and R_j^2 induced by applying the bootstrap procedure to the dataset. Achieved significance level is computed for the VIF and Klein's rule. In the following we describe the procedure step by step:

1. Create n_{boot} samples from original data with replacement of a given size (n_{sam}).
2. Compute $R_{g_{boot}}^2$ and $R_{j_{boot}}^2$ from each n_{boot} samples. This outputs a $B_{n_{boot}\times(p+1)}$ matrix.
3. Compute $\widehat{\text{ASL}}_{n_{boot}}$ for the VIF and Klein's rule.

Note that the matrix $B_{n_{boot}\times(p+1)}$ allows us to inspect results in detail and make further tests such as boxplots, pariwise Kolmogorov-Smirnov (KS) of the predictors and so on.

4. NUMERICAL EXPERIMENTS

4.1 Experiment 1

4.1.1 Data simulation

In this section, we implement the procedure described in Section 3.1 that allows us to test the hypothesis. We start by simulating 1000 data points according to the following regression:

$$Y_i = 10 - 5X_{1i} + 3X_{2i} + 9X_{3i} + \hat{\varepsilon}_i$$

where $\hat{\varepsilon} \sim N(0,3)$. X_1, X_3 and X_3 are simulated using `MASS` R package (Venables and Ripley, 2002) with the following correlation structure:

$$\begin{pmatrix} 1.00 & -0.945 & 0.3 \\ -0.945 & 1.00 & -0.5 \\ 0.30 & -0.50 & 1.0 \end{pmatrix}.$$

From this correlation structure it is expected that X_{1i} and X_{2i} may cause multicollinearity due to the high correlation between them (-0.945).

The code for the data generation process is detailed in Appendix A. The code for MTest is detailed in Appendix B. The function is defined with four parameters:

- `datos`: A $p+2$ dimensional data frame that includes the dependent variable.
- `nboot`: Number of bootstrap replicates. This is the n_{boot} parameter described in Section 3.1, the default nboot = 500.
- `nsam`: Sample size in the bootstrap procedure. When *nsam* = NULL (default), then nsam = nrow(datos)*3.
- `trace`: Logical. If `TRUE` then the iteration number out of a total of nboot is printed. The default value is `FALSE`.
- `seed`: A numeric value that sets the seed value of the procedure. The default value is `NULL`.
- `valor_vif`: A numeric value that sets threshold for in the VIF rule. The default value is `valor_vif=0.9`

4.1.2 Testing multicollinearity

Table 1 shows the R^2 for the global regression and auxiliary regressions in the first row, VIF values are presented in the second row and Klein's rule can be computed with this information which is shown in the third row. The fourth and fifth row of Table 1 present the MTest results by computing $\widehat{\text{ASL}}_{n_{boot}=1000}$ for VIF and Klein's rules.

From a traditional usage of the rules, the Klein's rule suggests that X_2 is a variable that causes multicollinearity since its auxiliary regression $R^2_{X_2}$ is greater than the global R^2_g. Kleins's rule is presented with $*$ denoting the variables that the rule identifies as a problem and with $\bullet$ the variables that do not.

In the VIF case, if the threshold is equal to 10, this method suggests that X_1 and X_2 causes multicollinearity. Both rules detect multicollinearity problems as expected.

MTest is also presented in Table 1 with ASL values computed with *nboot* $= 1000$, $\widehat{\text{ASL}}_{n_{boot}}$ for VIF and Klein are 1 for predictors X_1, X_2 and 0 for X_3. This implies that we cannot reject the null hypothesis stated in 3.1. In other words, this means that X_1 and X_2 also yield multicollinearity problems according to MTest which confirms the results in the application of traditional VIF and Klein's rules.

Table 1. R^2 for the global regression and auxiliary regressions are presented in the first row. VIF values are presented in the second row and Kleins's rule is presented with $*$ denoting the variables that the rule identifies as a problem and with $\bullet$ the variables that do not.

	Y	X_1	X_2	X_3
R^2	0.9073	0.9308	0.9432	0.5273
VIF		14.4583	17.606315	2.1158
$Klein$		$*$	$*$	$\bullet$
VIF: $\widehat{\text{ASL}}_{n_{boot}}$		1	1	0
Klein: $\widehat{\text{ASL}}_{n_{boot}}$		1	1	0

Boxplots of $R^2_{g_{boot}}$ and $R^2_{j_{boot}}$ are presented in Figure 1. They show that the bootstrap distributions are centered at their means for Y, X_1, X_2 and X_3, respectively are 0.9073, 0.9308, 0.9432, 0.5274. We need to leave all variables in the initial model, the MTest function also gives us a clear idea of the variability in global R^2_g: in our simulation, $0.9041 \leq R^2_g \leq 0.9106$.

This bootstrap replicates open the possibility of applying other testing methods to check how statistically different the values are. For example, we apply a Kolmogorov-Smirnov (KS) test.

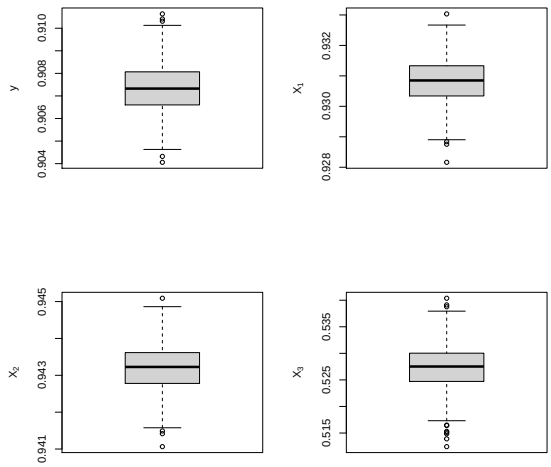

Figure 1. *Boxplot* results of the bootstrap procedure contained in $B_{n_{boot}\times(p+1)}$

A pairwise KS matrix of p-values is presented in Tables 2 and 3, we set the significance level $\alpha = 0.05$. The code for computing the pairwise KS matrix is detailed in Appendix C. Table 2 shows the pairwise p-values for the equality hypothesis of the n_{boot} replicates of variables in the rows compared to the variables in the columns. For example, the pairwise p-value between X_2 and X_3 (0) is lower

than α which rejects the equality hypothesis between them. Note that all the off-diagonal p-values also reject the hull hypothesis.

Table 2. Pairwise KS p-values of bootstrap samples $B_{n_{boot}\times(p+1)}$. `two.sided` is used as the alternative hypothesis.

	Y	X_1	X_2	X_3
Y	1	0	0	0
X_1	0	1	0	0
X_2	0	0	1	0
X_3	0	0	0	1

Table 3 shows the pairwise p-values of $B_{n_{boot}\times(p+1)}$. The null hypothesis is that the Cumulative Distribution Function of the variable in the row does lie below that of the variable in the column. In other words, we are testing if R^2 values in the rows are greater than the ones in the columns. Note that the first column is similar to testing $H_0 : \mu_{R^2_{X_{j_{boot}}}} \geq \mu_{R^2_{Y_{boot}}}$. For example, the first column tells us that X_1 and X_2 are greater than Y which is consistent with our previous findings.

Table 3. Pairwise KS p-values of bootstrap samples $B_{n_{boot}\times(p+1)}$. `greater` is used as the alternative hypothesis.

	Y	X_1	X_2	X_3
Y	1	0	0	1
X_1	1	1	0	1
X_2	1	1	1	1
X_3	0	0	0	1

Once candidate variables that may be causing multicollinearity are identified, some researchers decide to remove one or more of them. Note that results in Table 3 can also guide our decision on choosing whether X_1 or X_2 should be removed from the regression. If we take the sum over the rows in Table 3, this value gives us a metric that could be used to prioritize the removal of the variables. This is a metric of *how much* the variable in the row is greater than the one in the column. In this example, the row sum for X_2 is 4 and for X_1 is 3. It then suggests that X_2 is the one that should be removed.

4.2 Experiment 2

This experiment studies MTest vs rule of thumb VIF in (2) and contrasts its results. Following Salmerón-Gómez et al. (2018) and Salmerón-Gómez et al. (2021a), data is simulated as:

$$X_j = \sqrt{1-\lambda^2}W_j + \lambda W_p,$$

where $j = 2,\ldots,p$ with $p = 3,4,5$, $W_j \sim N(10,100)$, $\lambda \in \{0.8, 0.82, 0.84, \ldots, 0.98\}$ and $n \in \{20, 100, 200\}$. This setup let us specify different grades of collinearity (λ), different sample sizes (n) and different number of covariates (p).

Table 4 shows results for $p = 4$ and Appendix D for $p = 3$ (Table 9) and $p = 4$ (Table 10). We can see that all VIF troubling values are also detected by MTest, but there are cases where MTest detects multicollinearity and VIF does not (VIF threshold is 10). For example, with $\alpha = 0.05$, if $\lambda = 0.94$ and $n = 100$, MTest detects X_2 as a variable with potential multicollinearity but VIF does not ($VIF = 8.5$). All such these cases are in bold.

Table 4. Simulation results for VIF: $\widehat{\text{ASL}}_{n_{boot}}$ and VIF for $p = 4$. Different sample sizes (n) and grades of collinearity (λ).

		VIF: $\widehat{\text{ASL}}_{n_{boot}}$				VIF			
n	λ	X_1	X_2	X_3	X_4	X_1	X_2	X_3	X_4
20	0.80	0.00	0.00	0.00	0.02	4.2	3.1	3.9	5.8
	0.82	0.00	0.00	0.00	0.04	4.5	3.4	3.3	6.0
	0.84	0.00	0.00	0.00	**0.12**	5.1	3.8	3.7	6.9
	0.86	0.00	0.00	0.00	**0.28**	5.8	4.3	4.3	8.0
	0.88	**0.05**	0.00	0.00	**0.51**	6.7	5.0	5.1	9.6
	0.90	**0.21**	0.03	0.04	0.78	8.0	6.0	6.2	11.9
	0.92	**0.52**	**0.17**	**0.21**	0.94	9.9	7.5	7.8	15.3
	0.94	0.86	0.58	0.69	0.99	13.2	10.0	10.7	21.0
	0.96	1.00	0.95	0.99	1.00	19.6	15.0	16.4	32.5
	0.98	1.00	1.00	1.00	1.00	38.9	30.1	34.0	67.6
100	0.80	0.00	0.00	0.00	0.00	3.1	2.7	2.7	6.5
	0.82	0.00	0.00	0.00	**0.05**	2.8	3.1	3.3	8.1
	0.84	0.00	0.00	0.00	**0.30**	3.1	3.4	3.6	9.3
	0.86	0.00	0.00	0.00	0.81	3.5	3.9	4.1	10.8
	0.88	0.00	0.00	0.00	0.99	4.0	4.4	4.7	12.8
	0.90	0.00	0.00	0.00	1.00	4.7	5.3	5.6	15.6
	0.92	0.00	0.00	0.00	1.00	5.7	6.5	6.9	19.8
	0.94	0.01	**0.13**	**0.23**	1.00	7.5	8.5	9.0	26.8
	0.96	0.83	0.97	0.99	1.00	11.1	12.4	13.3	40.8
	0.98	1.00	1.00	1.00	1.00	21.7	24.4	26.2	82.6
200	0.80	0.00	0.00	0.00	0.00	3.5	2.9	2.3	7.4
	0.82	0.00	0.00	0.00	0.00	3.4	3.1	3.2	8.3
	0.84	0.00	0.00	0.00	**0.19**	3.7	3.4	3.6	9.5
	0.86	0.00	0.00	0.00	0.94	4.2	3.9	4.0	11.1
	0.88	0.00	0.00	0.00	1.00	4.9	4.5	4.7	13.2
	0.90	0.00	0.00	0.00	1.00	5.8	5.4	5.6	16.2
	0.92	0.00	0.00	0.00	1.00	7.2	6.8	7.0	20.8
	0.94	**0.21**	**0.05**	**0.16**	1.00	9.5	9.0	9.3	28.3
	0.96	1.00	1.00	1.00	1.00	14.0	13.6	13.8	43.5
	0.98	1.00	1.00	1.00	1.00	27.7	27.3	27.7	89.0

5. APPLICATION

This section applies the proposed method to a dataset available in Longley (1967) and is used to show problems of multicollinearity. It is a time series from 1947 to 1962 where

- y: number of people employed, in thousands.
- x_1: GNP implicit price deflactor.
- x_2: GNP, millions of dollars.
- x_3: number of people unemployed in thousands.
- x_4: number of people in the armed forces.
- x_5: non institutionalized population over 14 years of age.
- time: year.

The regression model is given by

$$y = \phi_0 + \phi_1 x_1 + \phi_2 x_2 + \phi_3 x_3 + \phi_4 x_4 + \phi_5 x_5 + \phi_6 \text{time} \quad (3)$$

its global $R^2_g = 0.996$ and its estimation is:

Table 5 shows the estimation of coefficients in equation (1). The p-value of the F statistic is closely equal to 0 which means that we can reject the null hypothesis that $\phi_1, \phi_2, \ldots, \phi_p = 0$. As mentioned in Gujarati et al. (2012), one symptom of the presence of multicollinearity is when we have a high R^2 and a few significant individual predictors. Which is the case in this application since out the six predictors, only three are statistically significant.

Table 5. Coefficient estimation of the application model: Equation (3).

	ϕ_i in Equation (3)
(Intercept)	67271.28*
	(23237.42)
x_1	-2.05
	(8.71)
x_2	-0.03
	(0.03)
x_3	-1.95**
	(0.48)
x_4	-0.96**
	(0.22)
x_5	0.05
	(0.23)
time	1585.16*
	(482.68)
R^2	0.9955
Adj. R^2	0.9921
Num. obs.	15
RMSE	295.62

Figure 2 shows the bootstrap replicates of the dependent variable (*global*) and the predictors (x_1 to *time*). Boxplots suggest that x_2, x_5 and time are candidate variables that generate multicollinearity problems. Furthermore, x_1, x_3 and x_4 seem not to be greater than the dependent variable.

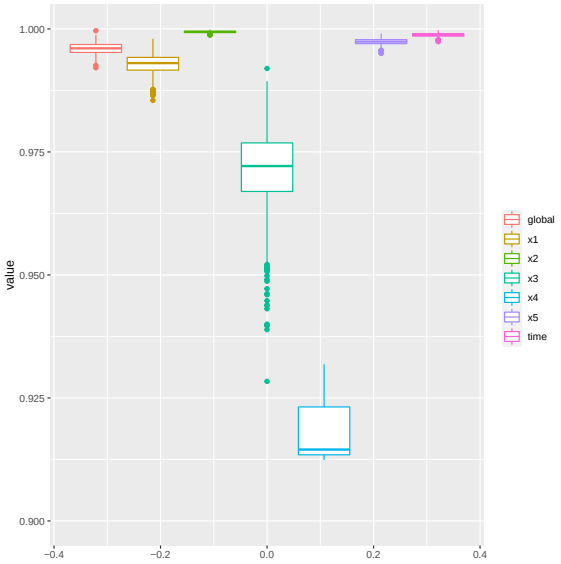

Figure 2. *Boxplot* of MTest results

Table 6 shows the results in a similar manner that the one shown in 1. According to the traditional VIF rule, all the variables yield multicollinearity problems except in x_4 (VIF threshold equal to 10). Traditional Klein's rule shows a multicollinearity problem in predictors x_2, x_5 and time.

The MTest given by VIF: $\widehat{\text{ASL}}_{n_{boot}}$ shows multicollinearity problems in the same predictors that the traditional VIF did, but the value 0.003 gives us a confidence metric to reject the null hypothesis $H_0 : \mu_{R^2_{x_{4boot}}} \geq 0.90$.

Setting $\alpha = 0.05$, the MTest given by Klein's: $\widehat{\text{ASL}}_{n_{boot}}$ shows multicollinearity problems in predictors x_1,x_2, x_5 and time. That is, x_1 was identified as a predictor with potential multicollinearity problems but the traditional Klein did not. Note also that setting $\alpha = 0.10$, we could reject the null hypothesis $H_0 : \mu_{R^2_{x_{1boot}}} \geq \mu_{R^2_{g_{boot}}}$.

Table 6. R^2 for the global regression and auxiliary regressions are presented in the first row. VIF values are presented in the second row and Kleins's rule is presented with $*$ denoting the variables that the rule identifies as a problem and with $\bullet$ the variables that do not.

	x_1	x_2	x_3	x_4	x_5	time
R^2	0.992	0.999	0.969	0.741	0.997	0.999
VIF	130	1491	32	4	348	746
Klein	$\bullet$	*	$\bullet$	$\bullet$	*	*
VIF: $\widehat{\text{ASL}}_{n_{boot}}$	1	1	1	0.003	1	1
Klein: $\widehat{\text{ASL}}_{n_{boot}}$	0.065	1	0	0	0.907	0.998

Table 7 shows the pairwise p-values of $B_{n_{boot}\times(p+1)}$ from the application data. The null hypothesis is that the Cumulative Distribution Function of the variable in the row does lie below of the variable in the column. The first column is similar to testing $H_0 : \mu_{R^2_{x_{3boot}}} \geq \mu_{R^2_{y_{boot}}}$. It tells us that the predictors that are greater than the response variable are x_2, x_5 and x_6, which is consistent with our intuition derived from Figure 2.

Table 7. Pairwise KS p-values of bootstrap samples $B_{n_{boot}\times(p+1)}$. `greater` is used as the alternative hypothesis.

	y	x_1	x_2	x_3	x_4	x_5	time
y	1	1	0	1	1	0	0
x_1	0	1	0	1	1	0	0
x_2	1	1	1	1	1	1	1
x_3	0	0	0	1	1	0	0
x_4	0	0	0	0	1	0	0
x_5	0.991	1	0	1	1	1	0
time	1	1	0	1	1	1	1

Table 7 can also help us decide which variable should be removed first. For example, in this case x_2, time and x_5 would be the order of removing the predictors. This is achieved by checking the rows of Table 7, this may suggest the ordering of the removal. Respectively, the row sums of x_2, time and x_5 are 7, 6, and 4.991.

In our application, after removing x_2 from the dataset, the p-values of the Klein's rule using MTest are 0.002, 0.000, 0.000, 0.314 and 0.845 for x_1, x_3, x_4, x_5 and *time* respectively. Only after removing x_2, *time* and x_5 multicollinearity was removed.

Nonetheless, this removal recommendation is a purely empirical approach. In this application, we could have divided x_2 by x_1 since this ratio (*realgni*) is a useful predictor as well, it is the real GNP. By doing this, and removing predictors x_5 and *time*, multicollinearity was also removed. This means that the theory behind the predictors could play a very important role when dealing with multicollinearity. The final model after this last consideration is shown in Table 8.

6. CONCLUSIONS

MTest is a bootstrap application for testing multicollinearity problems in the predictors. It lets us have a confidence metric, ASL, that given an α threshold helps us decide whether or not reject the null hypothesis stated in 3. The whole code generated for this article can be found at `https://github.com/vmoprojs/ArticleCodes/tree/master/MTest`.

Table 8. Coefficient estimation of the application model after removing variables causing multicollinearity.

	ϕ_i in Equation (3) after removing variables
(Intercept)	42716.56***
	(710.12)
x_3	-0.68**
	(0.17)
x_4	-0.84**
	(0.22)
realgni	72.01***
	(3.33)
R^2	0.9893
Adj. R^2	0.9864
Num. obs.	15

The application shows consistency with the numerical experiments. Both present MTest as a useful approach to test multicollinearity giving a *boosting* to the traditional rules in the sense that we can now have distributions of $\mu_{R^2_{X_{g_{boot}}}}$ and $\mu_{R^2_{X_{j_{boot}}}}$ which are involved in the testing procedure.

A graphical representation of the bootstrap replicates were found to be very useful. In our application, MTest lets us have boxplots of the predictors and the dependent variable to guide our intuition and later perform a KS test.

The pairwise KS matrix of p-values is a complementary tool for testing multicollinearity derived from MTest. It can also help us decide which predictor has more potential multicollinearity problems.

This work can be extended to generalized linear models in the same manner that Fox and Weisberg (2019) did with the `vif` function contained in `car` package, but the general idea of MTest would remain the same. Another extension of MTest could be possible in the context of Ridge, Lasso or Elastic Net regression. The hyperparameters in these methods are usually selected through cross validation or repeated cross validation, and MTest could be implemented in these procedures.

REFERENCES

Achen, C. H. (1982). *Interpreting and using regression*. Sage.

Carrodus, M. L. and Giles, D. (1992). The exact distribution of R2 when the regression disturbances are autocorrelated. *Economics Letters, 4*(38), 375-380. https://doi.org/10.1016/0165-1765(92)90021-P

Dinov, I. D. (2016). Methodological challenges and analytic opportunities for modeling and interpreting Big Healthcare Data.*Gigascience, 5*(1). https://doi.org/10.1186/s13742-016-0117-6

Efron, B. (1992). Bootstrap methods: another look at the jackknife. *Breakthroughs in statistics* (pp. 569-593). Springer, New York, NY. https://doi.org/10.1214/aos/1176344552

Farrar, D. E. and Glauber, R. R., (1967). Identities for negative moments of quadratic forms in normal variables. *The Review of Economic and Statistics, 49*, 92-107. https://doi.org/10.1016/j.spl.2008.12.004

Fox, J. and Weisberg, S. (2019). *An R Companion to Applied Regression*, Third Edition. Thousand Oaks CA: Sage. https://socialsciences.mcmaster.ca/jfox/Books/Companion/

Godfrey, L. (2009). *Bootstrap tests for regression models*. Springer.

Gujarati, D. N. and Porter, D. C. and Gunasekar, S., (2012). *Basic econometrics)*. McGraw-Hill, United States.

Horowitz, J. L. (2001). *The bootstrap*. In Handbook of econometrics (Vol. 5, pp. 3159-3228). Elsevier.

Imdadullah, M. and Aslam, M. and Altaf, S. (2016). mctest: An R Package for Detection of Collinearity among Regressors. textitThe R Journal, 8(2), 499–509. https://doi.org/10.32614/RJ-2016-062

Jaya, I. G. N. M. and Ruchjana, B. and Abdullah, A. (2020). Comparison Of Different Bayesian And Machine Learning Methods In Handling Multicollinearity Problem: A Monte Carlo Simulation Study. *ARPN J. Eng. Appl. Sci, 15*(18), 1998-2011.

Klein, L.R. (1962). *An Introduction to Econometrics*. Prentic-Hall, Englewood, Cliffs, N. J, 101.

Koerts, J. and Abrahamse, A. P. J. (1969). *On the theory and application of the general linear model*. Rotterdam University Press.

Leamer, E. E., (1983). Model choice and specification analysis. *Handbook of econometrics, 1*, 285-330.

Longley, J. W. (1967). An appraisal of least squares programs for the electronic computer from the point of view of the user. *Journal of the American Statistical association, 62*(319), 819-841. https://doi.org/10.1080/01621459.1967.10500896

Marcoulides, K. M. and Raykov, T. (2019). Evaluation of variance inflation factors in regression models using latent variable modeling methods. *Educational and psychological measurement, 79*(5), 874-882. https://doi.org/10.1177/0013164418817803

Salmerón-Gómez, R. and García-García, C. and García-Pérez, J. (2018). Variance Inflation Factor and Condition Number in multiple linear regression. *Journal of Statistical Computation and Simulation, 88*(12), 2365-2384. https://doi.org/10.1080/00949655.2018.1463376

Salmerón-Gómez, R. and García-García, C. and García-Pérez, J. (2020). Detection of Near-Multicollinearity through Centered and Noncentered Regression. *Mathematics, 8*(6), 931-948. https://doi.org/10.3390/math8060931

Salmerón-Gómez, R. and García-García, C. and García-Pérez, J. (2021a). Obtaining a threshold for the Stewart index and its extension to ridge regression. *Computational Statistics, 36*, 1011–1029. https://doi.org/10.1007/s00180-020-01047-2

Salmerón-Gómez, R. and García-García, C. and García-Pérez, J. (2021b). A guide to using the r package "multicoll" for detecting multicollinearity. *Computational Economics, 57*(2), 529-536. https://doi.org/10.1007/s10614-019-09967-y

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

The World Bank, (2021). World Development Indicators. https://data.worldbank.org/ Accessed: 2010-11-16.

Venables, W. N. and Ripley, B. D. (2002) *Modern Applied Statistics with S*. Fourth Edition. Springer, New York. ISBN 0-387-95457-0

BIOGRAPHY

Víctor Morales Oñate, PhD in Statistics from the University of Valparaíso, Master in Applied Mathematics (USFQ), Economics (FLACSO), Philosophy (PUCV), Financial Risks (UNIR) and Engineer in Economics and Finance from the National Polytechnic School. He works in Data Analytics in the private sector, is a researcher and teacher of graduate programs.

Bolívar Morales Oñate, Telecommunications Engineer graduated from the National Polytechnic School and Master in Applied Mathematics from the San Francisco de Quito University. He works in statistical process control and is a professor at the Technical University of Ambato.

Appendix A. Data Generation Code

```
# ***** Data Generation
library(MASS)
rm(list = ls())
graphics.off()

# Sample size:
n <- 10000
# vector means
medias <- c(0,0,0)
# Correlation structure:
rho_12 <- -.945 # -.94
rho_13 <- .3
rho_23 <- -.5
# Coefficients of the regression:
betas <- c(10,-5,3,9)
s.d <- 3 # deviation of the residual

(Sigma <- matrix(c(1,rho_12,rho_13
,rho_12,1,rho_23,rho_13,rho_23,1),3,3))
set.seed(247)
# Predictors simulation:
X <- mvrnorm(n = n, medias, Sigma)
M <- cbind(1,X)
# Output simulation
y <- M %*% betas + rnorm(n,0,s.d)
datos <- data.frame(y,X)
```

Appendix B. Code for computing MTest

```
Mtest <- function(datos, nboot = 500,
                  nsam = NULL,trace = TRUE,
                  seed = NULL,
                  valor_vif = 0.9)
{
  if(is.null(nsam)){nsam = nrow(datos)*3}

  vals <- 1:nrow(datos)

  if(!is.null(seed)) {set.seed(seed)}

  sol.rsq <- NULL
  sol.vif <- NULL
  i = 1
  while(i <=nboot)
  {
    sam <- sample(vals,nsam,replace = TRUE)
    aux <- datos[sam,]
    maux <- lm(y~.,data = aux)
    sm  <- summary(maux)
    vif.vals <- vif(maux)
    Raux <- (vif.vals-1)/vif.vals

    s1 <- c(sm$r.squared,Raux)
    sol.rsq <- rbind(sol.rsq,s1)
```

```
    sol.vif <- rbind(sol.vif,vif.vals)

    if(trace)
    {
      cat("Iteration",i,"out of ",nboot,"\n")
    }
    i = i+1
  }

  pval_vif <- NULL
  for(j in 2:ncol(sol.rsq))
  {
    pval_vif <- c(pval_vif,
    sum(sol.rsq[,j]>valor_vif)/nboot)
  }
  names(pval_vif) <-
  colnames(sol.rsq)[2:ncol(sol.rsq)]
  pval_klein <- NULL
  for(z in 2:ncol(sol.rsq))
  {
    pval_klein <- c(pval_klein,
    sum(sol.rsq[,1]<sol.rsq[,z])/nboot)
  }
  names(pval_klein) <-
  colnames(sol.rsq)[2:ncol(sol.rsq)]

  colnames(sol.rsq) <- c("global",
  paste(names(datos)[-1],sep =""))
  rownames(sol.rsq) <- 1:nrow(sol.rsq)
  return(list(Bvals= sol.rsq,
  pval_vif = pval_vif,pval_klein=pval_klein))
}
```

Appendix C. Code for the pairwise KS matrix of p-values

```
pairwise.ks.test <- function(X,
alternative="two.sided")
{
  #Returns the p value of the
  #pairwise KS test of X columns
  n <- ncol(X)
  sol <- matrix(NA, ncol = n,
  nrow = n)
  for(i in 1:(n))
  {
    for(j in (1):n)
    {
      # print(c(i,j))
      a <- suppressWarnings(
      ks.test(X[,i],X[,j],
      alternative = alternative))
      # print(a$p.value)
      sol[i,j] <- a$p.value
    }
  }
  if(alternative=="less")
  {print("alternative hypothesis:
  the CDF of x lies below that of y.
   Rows are `x` and Columns are `y`")}
  if(alternative=="greater")
  {print("alternative hypothesis:
  the CDF of x lies above that of y.
  Rows are `x` and Columns are `y`")}
  if(alternative=="two.sided")
  {print("alternative hypothesis:
  two-sided")}
  colnames(sol) <- colnames(X)
  rownames(sol) <- colnames(X)
  return(sol)
}
```

Appendix D. Results of Experiment 3 when $p = 3$ and $p = 5$

Table 9. Simulation results for VIF: $\widehat{\text{ASL}}_{n_{boot}}$ and VIF for $p = 3$. Different sample sizes (n) and grades of collinearity (λ).

		VIF: $\widehat{\text{ASL}}_{n_{boot}}$			VIF		
n	λ	X_1	X_2	X_3	X_1	X_2	X_3
20	0.80	0.00	0.00	0.00	4.8	2.6	4.9
	0.82	0.00	0.00	0.00	4.3	2.9	3.9
	0.84	0.00	0.00	0.00	4.7	3.2	4.3
	0.86	0.01	0.00	0.00	5.3	3.5	4.9
	0.88	0.02	0.00	0.01	6.1	4.1	5.7
	0.90	**0.08**	0.00	0.04	7.1	4.8	6.8
	0.92	0.27	0.01	**0.24**	8.7	5.9	8.4
	0.94	0.70	**0.13**	0.71	11.3	7.8	11.1
	0.96	0.97	0.79	0.99	16.5	11.6	16.5
	0.98	1.00	1.00	1.00	31.8	23.0	32.4
100	0.80	0.00	0.00	0.00	3.1	3.8	6.2
	0.82	0.00	0.00	0.00	2.8	2.4	4.1
	0.84	0.00	0.00	0.00	3.1	2.6	4.7
	0.86	0.00	0.00	0.00	3.5	3.0	5.5
	0.88	0.00	0.00	0.00	4.0	3.5	6.5
	0.90	0.00	0.00	0.02	4.8	4.2	8.0
	0.92	0.00	0.00	0.59	5.9	5.3	10.1
	0.94	0.01	0.00	1.00	7.7	7.2	13.8
	0.96	0.90	0.78	1.00	11.4	10.9	21.3
	0.98	1.00	1.00	1.00	22.5	22.6	43.9
200	0.80	0.00	0.00	0.00	3.7	3.2	6.5
	0.82	0.00	0.00	0.00	3.0	2.9	5.3
	0.84	0.00	0.00	0.00	3.4	3.2	6.1
	0.86	0.00	0.00	0.00	3.8	3.7	7.0
	0.88	0.00	0.00	0.00	4.4	4.2	8.3
	0.90	0.00	0.00	0.56	5.2	5.0	10.1
	0.92	0.00	0.00	1.00	6.4	6.3	12.8
	0.94	0.01	0.00	1.00	8.5	8.3	17.3
	0.96	1.00	1.00	1.00	12.5	12.4	26.3
	0.98	1.00	1.00	1.00	24.8	24.7	53.5

Table 10. Simulation results for VIF: $\widehat{\text{ASL}}_{n_{boot}}$ and VIF for $p = 5$. Different sample sizes (n) and grades of collinearity (λ).

		VIF: $\widehat{\text{ASL}}_{n_{boot}}$					VIF				
n	λ	X_1	X_2	X_3	X_4	X_5	X_1	X_2	X_3	X_4	X_5
20	0.80	0.01	0.00	0.01	0.00	0.65	6.1	4.4	4.7	3.7	10.4
	0.82	0.00	0.00	0.00	0.00	**0.17**	3.4	3.5	2.6	2.7	7.3
	0.84	0.00	0.00	0.00	0.00	**0.36**	3.7	3.8	2.9	3.0	8.4
	0.86	0.00	0.00	0.00	0.00	**0.58**	4.2	4.3	3.3	3.3	9.8
	0.88	0.01	0.00	0.00	0.00	0.82	4.8	4.8	3.8	3.8	11.7
	0.90	0.03	0.02	0.00	0.00	0.97	5.7	5.6	4.6	4.4	14.4
	0.92	**0.09**	**0.08**	0.03	0.00	1.00	7.0	6.8	5.8	5.4	18.4
	0.94	**0.40**	**0.40**	**0.26**	**0.08**	1.00	9.2	8.7	7.8	7.0	25.2
	0.96	0.95	0.93	0.86	0.70	1.00	13.6	12.6	12.0	10.2	38.8
	0.98	1.00	1.00	1.00	1.00	1.00	26.8	23.9	24.9	19.8	80.0
100	0.80	0.00	0.00	0.00	0.00	0.99	4.0	3.5	3.1	3.8	12.5
	0.82	0.00	0.00	0.00	0.00	**0.23**	2.5	3.0	3.0	2.8	9.1
	0.84	0.00	0.00	0.00	0.00	0.67	2.8	3.3	3.4	3.2	10.5
	0.86	0.00	0.00	0.00	0.00	0.97	3.1	3.8	3.8	3.6	12.4
	0.88	0.00	0.00	0.00	0.00	1.00	3.6	4.4	4.4	4.2	14.9
	0.90	0.00	0.00	0.00	0.00	1.00	4.3	5.2	5.3	5.0	18.5
	0.92	0.00	0.00	0.00	0.00	1.00	5.4	6.4	6.6	6.3	23.8
	0.94	0.00	**0.11**	**0.14**	**0.08**	1.00	7.1	8.4	8.7	8.4	32.9
	0.96	0.76	0.95	1.00	0.98	1.00	10.7	12.5	13.1	12.7	51.1
	0.98	1.00	1.00	1.00	1.00	1.00	21.6	24.9	26.3	25.6	106.1
200	0.80	0.00	0.00	0.00	0.00	0.61	3.5	3.3	2.8	3.1	10.2
	0.82	0.00	0.00	0.00	0.00	0.96	3.6	3.3	3.1	3.4	11.4
	0.84	0.00	0.00	0.00	0.00	1.00	4.0	3.7	3.4	3.8	13.1
	0.86	0.00	0.00	0.00	0.00	1.00	4.5	4.2	3.9	4.3	15.3
	0.88	0.00	0.00	0.00	0.00	1.00	5.2	4.8	4.6	4.9	18.3
	0.90	0.00	0.00	0.00	0.00	1.00	6.2	5.8	5.5	5.8	22.5
	0.92	0.00	0.00	0.00	0.00	1.00	7.7	7.2	6.8	7.2	28.7
	0.94	0.56	**0.27**	**0.13**	**0.25**	1.00	10.1	9.5	9.2	9.4	39.1
	0.96	1.00	1.00	1.00	1.00	1.00	14.9	14.3	13.8	13.9	60.0
	0.98	1.00	1.00	1.00	1.00	1.00	29.1	28.5	27.9	27.1	122.6