On the Selection of the Ridge and Raise Factors

Catalina Garcia*, Roman Salmeron Gomez, Jose Garcia Perez and Maria Del Mar Lopez Martin

Departamento de Metodos Cuantitativos para la Economia y la Empresa, Universidad de Granada, Campus Universitario de La Cartuja, 18071 (Granada), Spain; cbgarcia@ugr.es, romansg@ugr.es, jgarcia@ual.es, mariadelmarlopez@ugr.es

Abstract

Objectives: To analyze the relation of the ridge and raise factors with the squared coefficient of correlation and present, from this relation, a methodology to select the adequate value of the ridge and raise factors. Methods/Statistical Analysis: Two independent variables have been simulated with a given coefficient of correlation between 0.95 and 0.999 and a gap of 0.001 to obtain a dependent variable. Then, it has been selected the value of the ridge and raise factor that leads to values of VIF lesser than 10 for every value of the given coefficient of correlation. Findings: Apart to propose a way to select the raise and ridge factor, it is concluded that the relation between the ridge factor and the squared coefficient of correlation is linear while the relation of the raise factor with the squared coefficient of correlation is potential. Application/Improvements: The procedure presented in this paper allows adequately selecting the value of the ridge and raise factor that mitigates the collinearity. This contribution can be applied in many different fields where collinearity is present.

Keywords: Multicollinearity, Raise Regression, Ridge Regression

1. Introduction

Given the following standardized linear model

\[ y = X\beta + u \]

\[ y = \beta_1 x_1 + \beta_2 x_2 + u \]

(1)

with \( n \) observations and two exogenous variables, the relation between the exogenous variables can lead to near collinearity. Although the estimators will still be lineal, unbiased and optimal, the estimation will be unstable. A measure widely applied to diagnose the existence of collinearity is the Variance Inflator Factor (VIF) whose definition is given by the following expression:

\[ VIF_i = \frac{1}{1 - R_i^2} \]

(2)

where \( R_i^2 \) is the coefficient of determination of the regression between \( x_i \) on the rest of exogenous variables. In the case of two exogenous variables, \( R_i^2 \) coincides with the squared coefficient of correlation, \( \rho^2 \). Traditionally it is considered\(^4\) that there is an important problem of collinearity when the VIF is higher than 10, implying that \( \rho \geq 0.9 \) taking into account expression (2).

The ridge estimator is the most applied methodology to estimate models with collinearity\(^3\). The basis of ridge regression is to introduce a parameter \( k \geq 0 \), known as ridge parameter, in the diagonal of the matrix \( X'X \) trying to avoid the singularity of this matrix.

The raise estimation was presented\(^10\) as an alternative methodology to estimate models with collinearity. This method treats the collinearity from a geometrical point of view introducing the raise parameter \( (\lambda \geq 0) \) that allows to separate the angle between the vectors associated to the related exogenous variables. Note that the ridge and the raise estimators introduce a constant, \( k \) and \( \lambda \) respectively, opening a not trivial question about the selection of these parameters.

It was proposed a procedure\(^10\) to estimate \( k \) beginning with a small value and increasing it to get stability in the estimations. Ever since have been many attempts to obtain a good estimator of \( k \). There is a review of different methods to estimate \( k \) collecting 28 different methods and proposing five new methods\(^15\). They also compared the behavior of these values of \( k \) in relation to the vari-
 ance and the Mean Squared Error (MSE). Regarding raise regression, two criteria have been proposed\textsuperscript{14}, by following the methodology proposed to the ridge parameter, to select the value of \( k \) that leads to the lowest MSE.

From another point of view, it has been related the value of \( k \) in a model with two exogenous variables to the grade of correlation between them\textsuperscript{12}. These authors proposed to select for each value of the coefficient of correlation between the independent variables, the value of \( k \) that leads to stable estimations for the parameters. However, in the application of this methodology, it was obtained values of the VIF lesser than one which is not compatible with the theoretical concept of VIF given by expression (2).

There are more examples in the scientific literature where the VIF expression is erroneously extended from OLS to the ridge regression leading to values lesser than 1\textsuperscript{16,24}. It has been developed an extension\textsuperscript{22} of the VIF which always leads to values of VIF equal or higher than 1, verifying also the conditions of being decreasing in \( k \) and continuous (i.e. when \( k = 0 \), the results coincide to OLS results). This expression will be applied in this paper to analyze the relation between the squared coefficient of correlation (\( \rho^2 \)) and the ridge parameter (\( k \)). Thus, by using a simulation for each value of the squared coefficient of correlation (\( \rho^2 \)) is selected the value of \( k \) that makes the VIF lesser than 10. Analogously, the procedure is repeated to analyze the relation between the squared coefficient of correlation (\( \rho^2 \)) and the raise parameter (\( \lambda \)).

Before presenting the contribution of this paper, we proceed to review the main characteristics of the ridge and raise estimators, taking special interest in the expression of the VIF.

### 1.1 The Ridge Estimator and the VIF

The ridge estimator, proposed\textsuperscript{3,4} to mitigate the collinearity, is given by:

\[
\hat{\beta}_k(k) = (X'X + kI)^{-1}X'Y
\]

where \( k \geq 0 \). In a standardized model with two exogenous variables it is verified\textsuperscript{14,22} that:

\[
X'X = \begin{pmatrix}
1 & \rho_{12} \\
\rho_{12} & 1
\end{pmatrix}, \quad (X'X)^{-1} = \frac{1}{1 - \rho_{12}^2} \begin{pmatrix}
1 & -\rho_{12} \\
-\rho_{12} & 1
\end{pmatrix}
\]

where \( \rho_{12} \) is the coefficient of correlation between the exogenous variables. These expressions allowed Marquard\textsuperscript{23} to state that the elements of the main diagonal of \( (X'X)^{-1} \) are the variance inflation factors associated to exogenous variables. This idea was extended directly to ridge regression\textsuperscript{22}:

\[
\text{“The parameter variance inflation factors for the ridge estimator are the diagonal elements of } [X'X + kI^{-1}][X'X][X'X + kI^{-1}] \text{ where } (X'X) \text{ is in correlation form”}
\]

From this statement, the following expression was proposed\textsuperscript{22} to obtain the VIF in ridge regression:

\[
VIF = \frac{(1+k)^2 - 2(1+k)r_{12}^2 + r_{12}^4}{(1+k)^2 - r_{12}^4} \quad (5)
\]

This extension of the VIF traditionally applied in OLS to the ridge estimator, although widely applied in scientific literature, is not correct. In fact, this widely applied expression does not verify a behavior coherent with the concept of VIF due to the following reasons\textsuperscript{22}:

- The relation between VIF and \( r_{12} \) is not strictly increasing\textsuperscript{22}.
- The expression (5) can lead to values of VIF lesser than one which is incompatible with the theoretical definition of VIF given by expression (2).
- The expression (5) is not decreasing in \( k \).

It seems that\textsuperscript{22} applied expression (5) for the calculations in the simulations and this could be the reason why they obtain values of VIF lesser than one.

To avoid this problem, it is suggested to part from the augmented model\textsuperscript{22,23}. Thus, the expression of the ridge estimator (3), can be rewritten as:

\[
\hat{\beta}_k(k) = (X_A'X_A)^{-1}X_A'Y_A = (X'X + kI)^{-1}X'Y
\]

being \( X_A = \begin{pmatrix} X \\ \sqrt{k}I \end{pmatrix} \) and \( Y_A = \begin{pmatrix} Y \\ 0 \end{pmatrix} \), where \( I \) is the identity matrix and is the null matrix with order \( p=2 \). Thus, the following expression of the VIF associated to the ridge estimator for the case of two exogenous variables was presented\textsuperscript{22}:

\[
VIF(k) = \frac{\sqrt{n+2}(1+k) - k}{(n+2)^2(1+k)^2 - 2(n+2)(1+k)(1-k)} \quad (7)
\]

This expression verifies all the conditions desirables for a coherent VIF\textsuperscript{22} and it will be the one applied in this paper.

### 1.2 The Raise Estimator and the VIF

The raise estimator was presented\textsuperscript{14} as an alternative procedure to estimate a model with collinearity. This method
parts from the idea that if vectors $x_1$ and $x_2$ are very close, it is possible to diminish the correlation between them increasing the angle between both vectors. Thus, the VIF will be diminished and consequently, the collinearity.

By raising the vector $x_1$, the model to estimate is given by:

$$y = \beta_1 \tilde{x}_1 + \beta_2 x_2 + u$$

where $\tilde{x}_1 = x_1 + \lambda e_1$, $e_1 = x_1 - \rho x_2$ and $e_1 \perp x_2$, being $e_1$ the residual obtained from $x_1 = \hat{a}x_2 + \hat{a}$ shown in Figure 1.

The expression of the VIF associated to the raised estimator is given by the following expression:

$$VIF(\lambda) = \frac{(1 + \lambda)^2 (1 - \rho^2) + \rho^2}{(1 + \lambda)^2 (1 - \rho^2)}$$

(9)

From (7) and (9) it is possible to conclude that VIF(K) and VIF(\lambda) are always equal or higher than one, decreasing in $k$ and $\lambda$, respectively, and continuous for $k = 0$ and $\lambda = 0$. Then, both expressions verify the desirable properties of a Variance Inflator Factor\textsuperscript{22,25,26}.

Once reviewed the main characteristics of the ridge and raise estimator, the paper presents in Section 2 a Montecarlo simulation to analyze the relation between the ridge and raise factors with the squared coefficient of correlation and some empirical applications in Section 3. To finish, Section 4 summarizes the main conclusions of this paper.

![Figure 1. Representation of raise regression.](image)

2. Monte Carlo Simulation

This section presents a Monte Carlo simulation to analyze the relation between the ridge and raise factors with the coefficient of correlation. The simulation, previously applied in the scientific literature\textsuperscript{7,27-29}, parts from the generating of the following independent variables:

$$x_1 = \sqrt{1 - \gamma^2} z_1 + \gamma z_2$$

$$x_2 = \sqrt{1 - \gamma^2} z_2 + \gamma z_2$$

(10)

where $z_1$ and $z_2$ are independent random numbers distributed by a normal distribution $N(0,100)$ and $\gamma$ is the coefficient of correlation between $x_1$ and $x_2$. The dependent variable $y$ is generated from the following relation:

$$y = x_1 + x_2 + u$$

(11)

being $u$ random numbers distributed by a normal distribution $N(0,1)$.

The size sample is considered to be 30 and 60. Thus, for each simple size two independent variables have been generated from expression (10) requiring a coefficient of correlation varying between 0.950 and 0.999 with a gap of 0.001. The dependent variable has been obtained from expression (11). Note that after the simulation, the coefficient of correlation estimated for the sample ($\hat{\rho}$) will slightly vary from the established as population coefficient of correlation.

In contrast to the paper presented by\textsuperscript{17}, this paper only takes into account values of the coefficient of correlation higher than 0.95. The justification is based on expression (2) where is possible to observe that values of the coefficient of correlation lesser than 0.95 lead to values of VIF lesser than 10 and, in this case, the high collinearity will not be present.

Once the independent and dependent variables have been simulated, the first value of $k$ and $\lambda$ that lead to a value of the VIF lesser than 10 has been selected. Next, it is analyzed the relation between the coefficient of correlation and the selected value of $k$ and $\lambda$. The following subsections present the regressions that will allow to select the value of $k$ and $\lambda$, respectively, for a given coefficient of correlation.

2.1 Relation of the Ridge Estimator and the Squared Coefficient of Correlation

From the simulation procedure previously presented, it has been selected the first value of $k$ that leads to a value of VIF lesser than 10 for every value of the coefficient of correlation. These values are displayed in Table 1. Figure 2 shows the scatter plot between the squared coefficient of correlation and the values of $k$ for $n$ equal to 30 and 60. Note that both functions are clearly linear and almost coincide.
Table 2 and Table 3 displays the model regression that explains the value of $k$ from the squared coefficient of correlation for $n$ equal to 30 and 60, respectively. Note that in both cases the coefficient of determination is approximately one, the coefficient are individually significant and the model is globally significant. Indeed, the estimators of the parameters are very similar as expected from Figure 2.

**Table 1.** (First part) Relation between the squared coefficient of correlation obtained from the simulation and the corresponding value of $k$ which leads to VIF<10 for $n$ equal to 30 and 60

| $\rho$ | $\rho^2$ | $k$ | $\rho^2$ | $k$ |
|---|---|---|---|---|
| 0.950 | 0.901070 | 0.000599 | 0.903645 | 0.002027 |
| 0.951 | 0.905488 | 0.003044 | 0.905126 | 0.002847 |
| 0.952 | 0.906191 | 0.003434 | 0.906884 | 0.003820 |
| 0.953 | 0.908317 | 0.004608 | 0.909395 | 0.005207 |
| 0.954 | 0.910431 | 0.005774 | 0.910577 | 0.005859 |
| 0.955 | 0.913036 | 0.007209 | 0.911845 | 0.006559 |
| 0.956 | 0.914137 | 0.007815 | 0.921418 | 0.011824 |
| 0.957 | 0.950673 | 0.027725 | 0.916331 | 0.009030 |
| 0.958 | 0.918804 | 0.010381 | 0.918518 | 0.010232 |
| 0.959 | 0.920304 | 0.011204 | 0.920227 | 0.011171 |
| 0.960 | 0.922984 | 0.012673 | 0.922915 | 0.012645 |
| 0.961 | 0.924470 | 0.013486 | 0.924298 | 0.013403 |
| 0.962 | 0.926034 | 0.014342 | 0.926174 | 0.014430 |
| 0.963 | 0.927588 | 0.015191 | 0.928287 | 0.015585 |
| 0.964 | 0.930574 | 0.016821 | 0.929389 | 0.016187 |
| 0.965 | 0.932295 | 0.017759 | 0.931428 | 0.017301 |
| 0.966 | 0.933288 | 0.018300 | 0.934218 | 0.018822 |
| 0.967 | 0.936118 | 0.019840 | 0.935245 | 0.019382 |
| 0.968 | 0.937140 | 0.020396 | 0.938022 | 0.020892 |
| 0.969 | 0.939163 | 0.021495 | 0.939461 | 0.021675 |
| 0.970 | 0.941677 | 0.022859 | 0.941740 | 0.022912 |
| 0.971 | 0.943845 | 0.024034 | 0.943593 | 0.023917 |
| 0.972 | 0.945019 | 0.024670 | 0.945265 | 0.024823 |
| 0.973 | 0.947551 | 0.026039 | 0.947420 | 0.025989 |
| 0.974 | 0.949526 | 0.027106 | 0.949515 | 0.027122 |

**2.2 Relation of the Raise Estimator and the Squared Coefficient of Correlation**

From the simulation procedure previously presented, it has been selected the first value of $\lambda$ that leads to a value of VIF lesser than 10 for every value of the coefficient of correlation. These values are displayed in Table 4. Figure 3 shows the scatter plot between the squared coefficient of correlation and the values of $\lambda$ for $n$ equal to 30 and 60. Note that the relation is potential of type $\lambda = \alpha (\rho')^\beta$. Then, the regression model will be estimated as $\log \lambda = \log \alpha + \beta \log \rho^2 + u$, coinciding $\beta$ of the potential model with the regression coefficient of transformed data and obtaining $\alpha$ as the antilog($\alpha$).

**Table 2.** (Second part) Relation between the squared coefficient of correlation obtained from the simulation and the corresponding value of $k$ which leads to VIF<10 for $n$ equal to 30 and 60

| $\rho$ | $\hat{\rho}^2$ | $k$ | $\hat{\rho}^2$ | $k$ |
|---|---|---|---|---|
| 0.975 | 0.950960 | 0.027880 | 0.951674 | 0.028288 |
| 0.976 | 0.952981 | 0.028970 | 0.954229 | 0.029667 |
| 0.977 | 0.954912 | 0.030010 | 0.955795 | 0.030510 |
| 0.978 | 0.956502 | 0.030866 | 0.957844 | 0.031613 |
| 0.979 | 0.959051 | 0.032236 | 0.959888 | 0.032712 |
| 0.980 | 0.961947 | 0.033791 | 0.960923 | 0.033269 |
| 0.981 | 0.963326 | 0.034530 | 0.962504 | 0.034118 |
| 0.982 | 0.965461 | 0.035674 | 0.964921 | 0.035414 |
| 0.983 | 0.966274 | 0.036109 | 0.967430 | 0.036758 |
| 0.984 | 0.968436 | 0.037266 | 0.969322 | 0.037770 |
| 0.985 | 0.971711 | 0.038727 | 0.970325 | 0.038307 |
| 0.986 | 0.972391 | 0.039379 | 0.972710 | 0.039581 |
| 0.987 | 0.974299 | 0.040396 | 0.975289 | 0.040957 |
| 0.988 | 0.976304 | 0.041464 | 0.977697 | 0.042241 |
| 0.989 | 0.977790 | 0.042256 | 0.978812 | 0.042834 |
| 0.990 | 0.981105 | 0.044018 | 0.981157 | 0.044081 |
| 0.991 | 0.982176 | 0.044587 | 0.983593 | 0.045376 |
| 0.992 | 0.984338 | 0.045734 | 0.984825 | 0.046029 |
| 0.993 | 0.987196 | 0.047249 | 0.986825 | 0.047090 |
| 0.994 | 0.989093 | 0.048253 | 0.988284 | 0.047864 |
| 0.995 | 0.991003 | 0.049263 | 0.990277 | 0.048918 |
| 0.996 | 0.992742 | 0.050181 | 0.993608 | 0.050680 |
| 0.997 | 0.995062 | 0.051406 | 0.994012 | 0.050894 |
| 0.998 | 0.996048 | 0.051925 | 0.996683 | 0.052303 |
| 0.999 | 0.998847 | 0.053400 | 0.998847 | 0.052979 |

Tables 5 and 6 show the regression of the value of $\lambda$ depending of the squared coefficient of correlation for $n$
equal to 30 and 60, respectively. Note that in both cases the coefficient of determination is approximately one, the coefficient are individually significant and the model is globally significant.

Figure 2. Scatter plot between \( k \) and \( \rho^2 \) for \( n \) equal to 30 and 60.

### Table 3. Regression of \( k \) in function of squared coefficient of correlation for \( n \) equal to 30

| \( n=30 \) | Coefficient | Standard error | T-statistic | P-value |
|-----------|-------------|----------------|-------------|---------|
| Constant  | -0.4854904  | 0.00051066     | -950.708684 | 2.898E-104 |
| \( \rho^2 \) | 0.53972914  | 0.00053669     | 1005.66914  | 1.952E-105 |
| R\(^2\)=0.99995254 | F=1011370.42 | p-value=1.952E-105 |

### Table 4. Regression of \( k \) in function of squared coefficient of correlation for \( n \) equal to 60

| \( n=60 \) | Coefficient | Standard error | T-statistic | P-value |
|-----------|-------------|----------------|-------------|---------|
| Constant  | -0.48577367 | 0.00049611     | -979.166291 | 7.034E-105 |
| \( \rho^2 \) | 0.54004815  | 0.00052158     | 1035.40551  | 4.82E-106 |
| R\(^2\)=0.99995523 | F=1072064.56 | p-value= 4.82E-1 |

### Table 5 (First part). Relation between the squared coefficient of correlation obtained from the simulation and the corresponding value of \( \lambda \) which leads to VIF<10 for \( n \) equal to 30 and 60

| \( \rho \)   | \( \rho^2 \) | \( \lambda \) | \( \lambda \) |
|-------------|-------------|-------------|-------------|
| 0.950       | 0.901070    | 9.999007    | 0.020856    |
| 0.951       | 0.905488    | 9.998996    | 0.029638    |
| 0.952       | 0.906191    | 9.999008    | 0.040322    |
| 0.953       | 0.908317    | 9.999006    | 0.056095    |
| 0.954       | 0.910431    | 9.999000    | 0.063743    |
| 0.955       | 0.913036    | 9.998996    | 0.072109    |

### Table 6 (Second part). Relation between the squared coefficient of correlation obtained from the simulation and the corresponding value of \( \lambda \) which leads to VIF<10 for \( n \) equal to 30 and 60

| \( \rho \)   | \( \rho^2 \) | \( \lambda \) | \( \lambda \) |
|-------------|-------------|-------------|-------------|
| 0.956       | 0.914137    | 9.998999    | 0.959905    | 0.141484 |
| 0.957       | 0.950673    | 9.998996    | 0.957252    | 0.103183 |
| 0.958       | 0.918804    | 9.999003    | 0.958394    | 0.119225 |
| 0.959       | 0.920304    | 0.132793    | 0.959285    | 0.132196 |
| 0.960       | 0.922984    | 0.154009    | 0.960685    | 0.153449 |
| 0.961       | 0.924470    | 0.166242    | 0.961404    | 0.164814 |
| 0.962       | 0.926034    | 0.179506    | 0.962379    | 0.180711 |
| 0.963       | 0.927588    | 0.193092    | 0.963476    | 0.199344 |
| 0.964       | 0.930574    | 0.220438    | 0.964048    | 0.209383 |
| 0.965       | 0.932295    | 0.237001    | 0.965105    | 0.228582 |
| 0.966       | 0.933288    | 0.246830    | 0.966550    | 0.256244 |
| 0.967       | 0.936118    | 0.276081    | 0.967081    | 0.26866 |
| 0.968       | 0.937140    | 0.287114    | 0.968515    | 0.296853 |
| 0.969       | 0.939163    | 0.309753    | 0.969258    | 0.313187 |
| 0.970       | 0.941677    | 0.339472    | 0.970433    | 0.340244 |
| 0.971       | 0.943845    | 0.366656    | 0.971387    | 0.363413 |
| 0.972       | 0.945019    | 0.382031    | 0.972248    | 0.385316 |
| 0.973       | 0.947551    | 0.416882    | 0.973355    | 0.415023 |
| 0.974       | 0.949526    | 0.445855    | 0.974431    | 0.445678 |
On the Selection of the Ridge and Raise Factors

The selection of the ridge and raise factors, for 60 cities where $\rho = 0.984$, $\gamma_1 = -0.077$ and $\gamma_2 = -0.177$. The original model estimated by OLS is given by:

$$\hat{y} = 3.0609879x_1 - 3.1890121x_2$$

The model is globally significant ($F_{\text{exp}} = 28.4071$) and the estimated parameters of the two exogenous variables are also individually significant, $t_{\text{exp}}(\hat{\beta}_1) = 5.0695$ and $t_{\text{exp}}(\hat{\beta}_2) = -5.3269$. The coefficient of determination is equal to 0.32875 and the VIF is equal to 31.502 denoting severe collinearity. In order to estimate this model by using ridge or raise regressions is essential to select and appropriate value for the ridge and raise factors, respectively.

From the regressions presented in Section 3 is possible to relate the ridge and raise factors with the squared coefficient of correlation. Thus, by using the regression shown in Table 3, it is obtained a value of $k = 0.0372$ which leads to a value of VIF equal to 9.9999. The model estimated by ridge regression for $k = 0.0372$ will be expressed as:

$$\hat{y} = 0.87701566x_1 - 1.002683587701566x_2$$

This model is not globally significant and this fact is based on the property of the coefficient of determination of the ridge regression that is decreasing in function of $k$. For this reason the $F_{\text{exp}}$ of the global significance test is also decreasing and the model will not be globally significant.

On the other hand, using the regression shown in Table 6, it is obtained a value of $\lambda = 0.9756981$ leading to a value of VIF equal to 31.502. The model estimated by raise regression for $\lambda = 0.9756981$ will be expressed as:

$$\hat{y} = 1.54931965x_1 - 1.70153054x_2$$

This model maintains constant the value of the coefficient of determination for all values of $\lambda$. Consequently, the model remains being globally significant and the estimated parameters will be individually significant.

Note that the methodology proposed allows to select an adequate value of $k$ and $\lambda$ that in both cases lead to a value of VIF lesser than 10 concluding that the collinearity has been mitigated.

Figure 3. Scatter plot between $\lambda$ and $\rho^2$ for n equal to 30 and 60.

3. Empirical Application

To illustrate the contribution of this paper, the following empirical application is presented by using an example previously applied in the scientific literature. In this example, the mortality rate $y$, is related to the nitrogen oxide pollution potential, $x_1$ and the hydrocarbon pollution potential, $x_2$, for 60 cities where $\rho = 0.984$, $\gamma_1 = -0.077$ and $\gamma_2 = -0.177$. The original model estimated by OLS is given by:

$$\hat{y} = 3.0609879x_1 - 3.1890121x_2$$

The model is globally significant ($F_{\text{exp}} = 28.4071$) and the estimated parameters of the two exogenous variables are also individually significant, $t_{\text{exp}}(\hat{\beta}_1) = 5.0695$ and $t_{\text{exp}}(\hat{\beta}_2) = -5.3269$. The coefficient of determination is equal to 0.32875 and the VIF is equal to 31.502 denoting severe collinearity. In order to estimate this model by using ridge or raise regressions is essential to select and appropriate value for the ridge and raise factors, respectively.

From the regressions presented in Section 3 is possible to relate the ridge and raise factors with the squared coefficient of correlation. Thus, by using the regression shown in Table 3, it is obtained a value of $k = 0.0372$ which leads to a value of VIF equal to 9.9999. The model estimated by ridge regression for $k = 0.0372$ will be expressed as:

$$\hat{y} = 0.87701566x_1 - 1.002683587701566x_2$$

This model is not globally significant and this fact is based on the property of the coefficient of determination of the ridge regression that is decreasing in function of $k$. For this reason the $F_{\text{exp}}$ of the global significance test is also decreasing and the model will not be globally significant.

On the other hand, using the regression shown in Table 6, it is obtained a value of $\lambda = 0.9756981$ leading to a value of VIF equal to 31.502. The model estimated by raise regression for $\lambda = 0.9756981$ will be expressed as:

$$\hat{y} = 1.54931965x_1 - 1.70153054x_2$$

This model maintains constant the value of the coefficient of determination for all values of $\lambda$. Consequently, the model remains being globally significant and the estimated parameters will be individually significant.

Note that the methodology proposed allows to select an adequate value of $k$ and $\lambda$ that in both cases lead to a value of VIF lesser than 10 concluding that the collinearity has been mitigated.
4. Conclusions

When a model presents collinearity it is appropriate to use, alternatively to OLS estimator, the ridge or raise estimations. In both cases, it is necessary to select an appropriate value for the ridge and raise factors, respectively. This paper analyzes the relation of the ridge and raise factors with the squared coefficient of correlation. Firstly, two independent variables were simulated with a given coefficient of correlation between 0.95 and 0.999 with a gap of 0.001. A dependent variable has been obtained from them. Then, it is proposed to select the value of the ridge and raise factor that leads to values of VIF lesser than 10 for every value of the given coefficient of correlation. The expression of the VIF used in this procedure verifies the properties of being continuous, decreasing in $k$ and $\lambda$ and being always higher than or equal to 1. It is concluded that the relation between the ridge factor and the squared coefficient of correlation is linear while the relation of the raise factor with the squared coefficient of correlation is potential. The contribution of this paper is illustrated with an empirical example showing that the procedure presented in this paper allows to adequately select the value of the ridge and raise factor that mitigate the collinearity. This contribution can be applied in many different fields where collinearity is present. As future research line, it is pretended to extend this analysis to models with $p \geq 2$ exogenous variables.

5. References

1. Theil H. Principles of econometrics. New York: Wiley; 1971.
2. Brien RM. A caution regarding rules of thumb for variance inflation factors. Quality and Quantity. 2007; 41:673–90. Crossref
3. Hoerl AE, Kennard RW. Ridge regression biased estimation for no orthogonal problems, Technometrics. 1970; 12:55–67. Crossref
4. Hoerl AE, Kennard RW. Ridge regression applications to no orthogonal problems. Technometrics. 1970; 12:69–82. Crossref
5. Hoerl AE, Kennard RW, Wballdin KF. Ridge regression some simulations. Communications in Statistics. 1975; 4(2):105–23. Crossref
6. Brown WG, Beattie BR. Improving estimates of economics parameters by use ridge regression with production function applications. American Journal of Agricultural Economics. 1975; 57(1):21–32. Crossref
7. McDonald GG, Galarneau DI. Monte Carlo evaluation of some ridge-type estimator. Journal of the American Statistical Association. 1975; 70(350):407–16. Crossref
8. Hemmerle WJ. An explicit solution for generalized ridge regression. Technometrics. 1975; 17:309–14. Crossref
9. Hocking RR, Speed FM, Lynn MJ. A class of biased estimators in linear regression. Technometrics. 1976; 18:425–38. Crossref
10. Garcia-Ferrer A. El problema de la multicolinealidad en los modelos de regresion lineales. Algunas soluciones posibles. Revista Espanola de Economia. 1977; 77:120–39.
11. Polverini F. Multicollinerita e stimatori ridge del modello classico di regressione lineare. Giornali degli Ecomoisti e Annali di Economia. 1978; 37(1-2):89–112.
12. Fourgeaud C, Gourieux C, Pradel J. Some theoretical results for generalized ridge regression estimators. Journal of Econometrics. 1984; 25(2):191–203. Crossref

### Table 7. Regression of $\lambda$ in function of squared coefficient of correlation for $n$ equal to 30

| $n = 30$ | Coefficient | Standard error | T-statistic | P-value |
|---|---|---|---|---|
| Constant | 0.6331135 | 0.04352952 | 14.5444629 | 3.2385E-19 |
| $\rho^2$ | 45.4525161 | 1.71631001 | 26.4826959 | 2.6942E-30 |
| $R^2$ | 0.93594305 | 0.370133183 | 701.333183 | 2.6942E-30 |

### Table 8. Regression of $\lambda$ in function of squared coefficient of correlation for $n$ equal to 60

| $n = 60$ | Coefficient | Standard error | T-statistic | P-value |
|---|---|---|---|---|
| Constant | 0.59709765 | 0.0296132 | 20.1632228 | 4.4666E-25 |
| $\rho^2$ | 43.3826355 | 1.15888387 | 37.4348428 | 3.5284E-37 |
| $R^2$ | 0.9668821 | 0.140136745 | 1401.36745 | 3.5284E-37 |
On the Selection of the Ridge and Raise Factors

13. Casella G. Condition numbers and minimax ridge regression estimators. Journal of the American Statistical Association. 1985; 80(391):753–8. Crossref

14. Garcia CB, Garcia J, Soto J. The raise method, an alternative procedure to estimate the parameters in presence of collinearity. Quality and Quantity. 2011; 45(2):403–23. Crossref

15. Kibria BM, Banik S. Some ridge regression estimators and their performances. Journal of Modern Applied Statistical Methods. 2016; 15(1):1–34.

16. Garcia CB, Garcia J, Salmeron R, Lopez MM. Raise regression, selection of the raise parameter. International Conference on Data Mining DMiN’15; Las Vegas (EEUU). 2015. p. 185–90.

17. Duzan H, Shariff N. Simulation study to verify the appropriate k value for ridge regression in two-variable regression model. Indian Journal of Science and Technology. 2015; 8(12):1–8. Crossref

18. Marquardt DW, Snee RD. Ridge regression in practice. The American Statistician. 1975; 29(1):3–20. Crossref

19. Mardikyan S, Cetin E. Efficient choice of biasing constant for ridge regression. International Journal of Contemporary Mathematic Sciences. 2008; 3(11):527–36.

20. Ijomah MA, Nwakuya, MT. Ridge trace a boost to ridge regression estimate in the presence of multicollinearity. Science African. 2011; 10:42–7.

21. Irfan M, Javed M, Raza MA. Comparison of shrinkage regression methods for remedy of multicollinearity problem. Middle-East Journal of Scientific Research. 2013; 14(4):570–9.

22. Garcia CB, Garcia J, Lopez MM, Salmeron R. Collinearity revisiting the variance inflation factor in ridge regression. Journal of Applied Statistics. 2015; 42(3):648–61. Crossref

23. Marquardt DW. Generalized inverses, ridge regression, biased lineal estimation and nonlinear estimation. Technometrics. 1970; 12(3):590–612. Crossref

24. Zhang J, Ibrahim M. A simulation study on spss ridge regression and ordinary least squares regression procedures for multicollinearity. Journal of Applied Statistics. 2005; 32:571–88. Crossref

25. Garcia J, Salmeron R, Garcia CB, Lopez MM. The raise estimators. Estimation, inference and properties. Communications in Statistics, Theory and Methods. 2015; 46(13):6446–62.

26. Garcia J, Salmeron R, Garcia CB, Lopez MM. Standardization of variables and collinearity diagnostic in ridge regression. International Statistical Review. 2015; 84(2):245–66. Crossref

27. Wichern DW, Dean W, Churchill GA. A comparison of ridge estimators. Technometrics. 1978; 20(3):301–11. Crossref

28. Gibbons, DG. A simulation study of some ridge estimators. Journal of the American Statistical Association. 1981; 76(373):131–9. Crossref

29. Salmeron R, Garcia J, Lopez MM, Garcia C. Collinearity diagnostic applied in ridge estimation through the variance inflation factor. Journal of Applied Statistics. 2016; 43(10):1831–49. Crossref

30. McDonald GC. Tracing ridge regression coefficients. Wiley Interdisciplinary Reviews: Computational Statistics. 2010; 2(6):695–703. Crossref

31. McDonald GC, Schwing RC. Instabilities of regression estimates relating air pollution to mortality. Technometrics. 1973; 15(3):463–81. Crossref