Revista Electrónica de Comunicaciones y Trabajos de ASEPUMA. **Rect@**
Volumen 19(2018). Páginas 45 a 62.
DOI: 10.24309/recta.2018.19.1.04

ABOUT THE LIMITS OF RAISE REGRESSION TO REDUCE CONDITION NUMBER WHEN THREE EXPLANATORY VARIABLES ARE INVOLVED

ANTONIO FRANCISCO ROLDÁN LÓPEZ DE HIERRO

aroldan@ugr.es
Universidad de Granada, Departamento de Didáctica de la Matemática
Campus Universitario de La Cartuja, Universidad de Granada. 18071 Granada (España)

ROMÁN SALMERÓN GÓMEZ

romansg@ugr.es
Universidad de Granada, Departamento de Métodos Cuantitativos para la Economía y la Empresa
Campus Universitario de La Cartuja, Universidad de Granada. 18071 Granada (España)

CATALINA GARCÍA GARCÍA

cbgarcia@ugr.es
Universidad de Granada, Departamento de Métodos Cuantitativos para la Economía y la Empresa
Campus Universitario de La Cartuja, Universidad de Granada. 18071 Granada (España)

Recibido (28/11/2017)
Revisado (04/06/2018)
Aceptado (20/06/2018)

RESUMEN: Este trabajo muestra que la regresión alzada puede considerarse como una metodología apropiada para reducir la multicolinealidad aproximada que aparece de forma natural en los problemas de regresión lineal. Cuando se trata de tres variables explicativas, su aplicación reduce el número de condición de la matriz asociada al conjunto de datos. Sin embargo, este procedimiento tiene un umbral: aunque las columnas de dicha matriz se pueden separar, se demuestra que el número de condición nunca será menor que una constante que se puede obtener fácilmente utilizando los elementos de la matriz inicial. Finalmente, la contribución se ilustra a través de un ejemplo empírico.

Palabras Clave: Multicolinealidad, regresión alzada, número de condición, autovalores, transformación de datos.

ABSTRACT: This manuscript shows that the raise regression can be considered as an appropriate methodology in order to reduce the approximate multicollinearity that naturally appears in problems of linear regression. When three explanatory variables are involved, its application reduces the condition number of the matrix associated to data set. Nevertheless, this procedure has a threshold: although the columns of X can be separated, it is proved that the condition number will never be less than a constant that can be easily worked out by using the elements of the initial matrix. Finally, the contribution is illustrated through an empirical example.

Keywords: Multicollinearity, raise regression, condition number, eigenvalue, transformation data.

1. Introduction

Regression analysis is a powerful methodology to describe the relationship between a response variable (usually denoted by Y) and one or more explicative variables (denoted by $X_1, X_2, \ldots, X_n$). Although researchers can consider a large list of possible models in order to study how variables $X_1, X_2, \ldots, X_n$ explain the variable Y, linear models are, undoubtedly, the most used in practice. Their simplicity and applicability lead most of researchers to use them at least as a first approach.

When studying a variable depending another ones, it is usual to involve a large number of independent variables $X_1, X_2, \ldots, X_n$ (as many variables as we can handle in practice). In some cases it appears a problem of *multicollinearity* because there is a high correlation among the input variables $X_1, X_2, \ldots, X_n$ (even they could be linearly dependent). This can be interpreted as such variables are measuring the same phenomena but in different ways. It is well-known that the existence of multicollinearity affects to the estimation by ordinary least squares (OLS) of the model as well as the interpretation of the obtained results. Ridge estimation (RE) (see e.g. Hoerl and Kennard[1,2]) is commonly applied as alternative method to analyze data by reducing the effects of multicollinearity. Usually labeled in statistic and econometric applications it is applied in many different fields such as medicine, physics and chemistry (see McDonald[3]). Recently other alternative methodologies have been proposed in order to (partially or totally) palliate the problem of multicollinearity. For instance, García *et. al.*[4] proposed to raise an independent variable to mitigate the effects of the linear dependence between explicative variables. This methodology is known as raise regression and was more developed by Salmerón *et. al.*[5].

Whatever been the applied methodology, it is important to check that its application has mitigated the collinearity. This fact justifies the development of measures to determine the presence of multicollinearity. A widely used measure in the literature is the variance inflator factor (VIF) (see e.g. Marquardt[6], Theil[7], Fox and Monette[8] and O'Brien[9]). Salmerón *et. al.*[5] presented the application of the VIF in the raise regression. However, this measure is not to always applicable, for example, when any of the independent variables is qualitative or there is an interaction term obtained from a dichotomic variable. This fact justifies the study of the condition number to be applied after the application or the raise regression. García *et. al.*[10] showed, in an empiric study with computational simulations, that the condition number decreases with the application of the raise regression and presented some algebraical problems.

In this paper we establish the limits of such improvement when three explicative variables are involved (including the constant term). In this way, we show that, if one variable is appropriately raised, we can reduce the condition number of the data matrix up to a concrete limit (and not beyond such limit). Therefore, the problem of multicollinearity can be partially overpassed by using this process.

A second question of interest is the transformation data required to appropriately calculate the condition number. In ordinary least squares (OLS), Belsley *et. al.*[11] shown that the basic data should be scaled to have equal length since it is assured that the matrix with orthogonal variables presents a condition number equal to 1 (the minimum value possible). However, this condition is also verified if the data are typified or standardized. The difference is that in these last cases the model has no independent term. Thus, although a lesser condition number will be obtained (see Belsley[12]) the influence of the independent term will be lost. In this paper, this question is treated for the raise regression.

The paper is structured as follows: Section 2 reviews how to calculate the condition number in OLS and introduces the raise regression. Section 3 develops the calculation of the condition number in the raise regression transforming the data to be unit length and standardized. Finally, Section 4 illustrates the contribution of this paper through an empirical example.

2. Preliminaries about the problem of multicollinearity

This section introduces the problem of multicollinearity and some basic preliminaries. From now on, let $\mathcal{M}_{m\times n}(\mathbb{R})$ denote the family of matrices with m rows and n columns over the field of all real numbers $\mathbb{R}$, and let $\mathcal{M}_n(\mathbb{R}) = \mathcal{M}_{n\times n}(\mathbb{R})$ the set of all real square matrices of order n. It will be supposed that the vectors $u \in \mathbb{R}^n$ are given in columns.

2.1. *The statement of the problem of approximate multicollinearity*

Let us consider a multiple linear model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + U, \tag{1}$$

in which a real random variable Y (called the *dependent* or *response variable*) is explained by means of two real random variables X_1 and X_2 (called the *independent* or *explicative variables*). Implicitly, the constant variable $X_0 \equiv 1$ is considered. Given a random sample $\{(x_{1i}, x_{2i}, y_i)\}_{i=1}^n$, it is possible to establish the following matrix X associated to the independent variables X_1 and X_2, and Y for response variable

$$X = \begin{pmatrix} 1 & x_{11} & x_{21} \\ 1 & x_{12} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} \end{pmatrix} \in \mathcal{M}_{n\times 3}(\mathbb{R}) \quad \text{and} \quad Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \in \mathcal{M}_{n\times 1}(\mathbb{R}).$$

When no confusion is possible, we will also denote the columns of X by $X_0 \equiv 1$, X_1 and X_2, respectively. The OLS estimation of the real coefficients β_0, β_1 and β_2 is given by

$$\widehat{\beta} = \left(X^T X\right)^{-1} X^T Y, \tag{2}$$

where the matrix $X^T X \in \mathcal{M}_3(\mathbb{R})$ plays a crucial role in this process: it must be invertible. Such property holds if, and only if, the columns of X are linearly independent vectors. If this property is not fulfilled, that is, if the columns of X are linearly dependent, it appears a problem of *multicollinearity*. In some cases, although $X^T X$ can be invertible, its determinant can be near to zero. In such cases, although the columns of X are linearly independent, they are near to be dependent: this is the problem of *approximate multicollinearity*. Under this condition, although estimation (2) can be computed, the estimation of model (1) will be unstable.

2.2. *The condition number*

A simple but effective procedure to diagnose approximate multicollinearity is based on the *condition number* of the matrix X, which is given by

$$k_X = \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}},$$

where $\lambda_{\max}$ is the maximum eigenvalue and $\lambda_{\min}$ is the minimum eigenvalue of a definite positive matrix obtained as $X^T X$ by following three possible transformations of X: (1) Normalizing the columns of X (obtaining a matrix $X_{u\ell}$); (2) standardizing the data (matrix X_s); and (3) typifying the data. Taking into account that $k_X \geq 1$ (because $0 < \lambda_{\min} \leq \lambda_{\max}$), the aim to transform the data is to consider a matrix such that its condition number is 1 when data are orthogonal. For instance, if X is the diagonal matrix $X = \operatorname{diag}(1, 1, 10)$, the eigenvalues of $X^T X = \operatorname{diag}(1, 1, 100)$ verify $\lambda_{\max}/\lambda_{\min} = 100$, and a condition number like 10 would not detect that data are orthogonal. Hence, it is necessary to consider a transformation of data in such a way that condition number is 1 when data are orthogonal. In the following lines, we describe each transformation.

Normalizing the columns of X, we consider the modified matrix

$$X_{u\ell} = \begin{pmatrix} \frac{1}{\sqrt{n}} & \frac{x_{11}}{\sqrt{\sum x_{1i}^2}} & \frac{x_{21}}{\sqrt{\sum x_{2i}^2}} \\ \frac{1}{\sqrt{n}} & \frac{x_{12}}{\sqrt{\sum x_{1i}^2}} & \frac{x_{22}}{\sqrt{\sum x_{2i}^2}} \\ \vdots & \vdots & \vdots \\ \frac{1}{\sqrt{n}} & \frac{x_{1n}}{\sqrt{\sum x_{1i}^2}} & \frac{x_{2n}}{\sqrt{\sum x_{2i}^2}} \end{pmatrix},$$

where we agree that, from now on, sums are given from $i = 1$ to $i = n$. Notice that, after this normalization, columns of $X_{u\ell}$ are unitary with respect to Euclidean metric on $\mathbb{R}^n$, and the matrix $X_{u\ell}^T X_{u\ell}$ has the form

$$X_{u\ell}^T X_{u\ell} = \begin{pmatrix} 1 & \frac{\sum x_{1i}}{\sqrt{n}\sqrt{\sum x_{1i}^2}} & \frac{\sum x_{2i}}{\sqrt{n}\sqrt{\sum x_{2i}^2}} \\ \frac{\sum x_{1i}}{\sqrt{n}\sqrt{\sum x_{1i}^2}} & 1 & \frac{\sum x_{2i}\ x_{1i}}{\sqrt{\sum x_{1i}^2}\sqrt{\sum x_{2i}^2}} \\ \frac{\sum x_{2i}}{\sqrt{n}\sqrt{\sum x_{2i}^2}} & \frac{\sum x_{2i}\ x_{1i}}{\sqrt{\sum x_{1i}^2}\sqrt{\sum x_{2i}^2}} & 1 \end{pmatrix}.$$

Matrix $X_{u\ell}^T X_{u\ell}$ is real, symmetric and definite positive, so its three eigenvalues, λ_1, λ_2 and λ_3, are strictly positive. For convenience, we agree that they will be ordered in the following way

$$0 < \lambda_1 \leq \lambda_2 \leq \lambda_3.$$

With this notation, the condition number of X is

$$k_X = \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}} = \sqrt{\frac{\lambda_3}{\lambda_1}} \geq 1.$$

It is usual to accept that the collinearity is *moderate* for values of the condition number between 20 and 30, *high* for values between 30 and 100, and *unacceptable* for values higher than 100.

The previous methodology is not the unique transformation to get that a matrix with orthogonal variables has a condition number equal to 1. It is also possible to standardize the data since, in this case, the matrix used to calculate the condition number coincides with the correlation matrix.

$$X_s = \begin{pmatrix} \frac{x_{11} - \overline{X}_1}{\sqrt{n\mathrm{Var}(X_1)}} & \frac{x_{21} - \overline{X}_2}{\sqrt{n\mathrm{Var}(X_2)}} \\ \frac{x_{12} - \overline{X}_1}{\sqrt{n\mathrm{Var}(X_1)}} & \frac{x_{22} - \overline{X}_2}{\sqrt{n\mathrm{Var}(X_2)}} \\ \vdots & \vdots \\ \frac{x_{1n} - \overline{X}_1}{\sqrt{n\mathrm{Var}(X_1)}} & \frac{x_{2n} - \overline{X}_2}{\sqrt{n\mathrm{Var}(X_2)}} \end{pmatrix}, \quad X_s^T X_s = \begin{pmatrix} 1 & \mathrm{corr}(X_1, X_2) \\ \mathrm{corr}(X_1, X_2) & 1 \end{pmatrix}.$$

Notice that this transformation implies the elimination of the independent term in model (1).

The third possible alternative is to typify the data. However, in this case, the condition number will coincide with the one obtained from standardized data. Thus, throughout this manuscript, we will only consider and compare the results obtained by employing the first two transformations of data: unit length and standardization.

2.3. *Raise regression for reducing approximate multicollinearity*

In order to overcome the above-mentioned drawbacks when collinearity appears, raise regression was introduced in García *et. al.* [4], and in Salmerón *et. al.* [5]. To correct the problem of approximate multicollinearity before proceeding to the estimation, let assume that we are interested in raising an independent variable (for instance, X_1) by using the other ones (in this case, X_0 and X_2). Hence we will separate them through the following auxiliary regression

$$X_1 = \widehat{\alpha}_0 + \widehat{\alpha}_2 X_2 + E, \tag{3}$$

whose estimation by OLS is $\hat{\alpha} = (X_2^T X_2)^{-1} X_2^T X_1$. The vector $E = (e_1, e_2, \ldots, e_n)^T \in \mathbb{R}^n$ cannot be zero because we assume that columns of X are not linearly dependent, and it satisfies $E \perp X_0 \equiv 1$ and $E \perp X_2$, so

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} x_{2i}\, e_i = 0. \tag{4}$$

The *raise vector*, denoted by $\widetilde{X}_1(t)$, is defined as

$$\widetilde{X}_1(t) = X_1 + tE, \qquad \text{where } t \in [0, +\infty)\,.$$

Let us show that

$$\widetilde{X}_1(t) \neq 0 \qquad \text{for all } t \in [0, +\infty)\,. \tag{5}$$

Reasoning by contradiction, suppose that there is $t_0 \in [0, +\infty)$ such that $\widetilde{X}_1(t_0) = 0$. Then $-t_0 E = X_1 = \widehat{\alpha}_0 + \widehat{\alpha}_2 X_2 + E$, so $\widehat{\alpha}_0 X_0 + \widehat{\alpha}_2 X_2 + (1 + t_0)\, E = 0$. Thus E is a linear combination of X_0 and X_2, which is a contradiction because $E \perp X_0$ and $E \perp X_2$. This contradiction guarantees that (5) holds.

The raise method will be obtained by substituting in model (1) the vector X_1 by the raise vector $\widetilde{X}_1(t)$, that is, the raise method will be the OLS regression with the vectors $\widetilde{X}_1(t)$ and X_2 instead of X_1 and X_2. Then, the model to estimate will be given by

$$Y = \beta_0\,(t) + \beta_1(t)\, \widetilde{X}_1(t) + \beta_2(t)\, X_2 + W, \tag{6}$$

where the estimated parameters depend on t: they will be called *raise estimators* and they will be denoted as $\widehat{\beta}_0\,(t)$, $\widehat{\beta}_1(t)$ and $\widehat{\beta}_2(t)$.

By using this methodology, correlation between vectors $\widetilde{X}_1(t)$ and X_2 will be less than the correlation between vectors X_1 and X_2, so the problem of multicollinearity is partially reduced.

Example 1 *Next, let apply this technique to a set of data previously considered by Hurvich*[13]*. Assume that* $X_1 =$ *"number of households" and* $X_2 =$ *"number of owner-occupied households" are used as explanatory variables (both of them are measured in units of 10000 households) in order to study the evolution of the monthly sales of backyard satellite antennas (variable* Y*). Table 1 contains the data set obtained in nine randomly selected districts. The following linear model can be computed by using the mentioned data set (with standard deviation in parentheses)*

$$\begin{gathered}\widehat{Y} = -2.38163 + 2.40179\, X_1 + 1.4435\, X_2, \quad R^2 = 0.9279,\ F_{2,6} = 38.63. \\ \qquad\quad (10.913) \qquad (2.221) \qquad (3.525)\end{gathered} \tag{7}$$

Since X_1 *and* X_2 *are linearly independent variables and the associated coefficient of determination of the linear model (7) is 92.79%, one can believe that variables* X_1 *and* X_2 *explain variable* Y *in an appropriate way. However, none of the estimated coefficients are significantly different to zero while the model is globally significant which means that a possible problem of multicollinearity appears. Also, the determinant of the correlation matrix, 0.0286, is very close to zero and the*

Tabla 1. Data set.

Satellite antenna sales			
District	Sales (Y)	Households (X_1)	Owner-occupied households (X_2)
1	50	14	11
2	73	28	18
3	32	10	5
4	121	30	20
5	156	48	30
6	98	30	21
7	62	20	15
8	51	16	11
9	80	25	17

condition number of matrix X, calculated for unit length data, $k_X = 36.343$, also confirms the existence of this problem. As a consequence, collinearity between variables X_1 and X_2 is high.

Thus, this setting seems to be ideal to use the raise regression. The following results are obtained by raising every variable for $t = 5$:

$$\underset{\quad (10.6103)\quad (0.3702)\quad (0.8311)}{\widehat{Y} = -4.7154 + 0.4003\,\widetilde{X}_1(5) + 4.5741 X_2,} \qquad R^2 = 0.9279,\ F_{2,6} = 38.63.$$

$$\underset{\quad (10.0927)\quad (0.5237)\quad (0.5875)}{\widehat{Y} = -0.9448 + 3.1489 X_1 + 0.2406\,\widetilde{X}_2(5),} \qquad R^2 = 0.9279,\ F_{2,6} = 38.63.$$

Note that the estimated coefficient of the not raised variable is now individually significant while the coefficient of determination and the global significance test are not modified. It is possible to find more information about this estimation method in Salmerón et al.[5].

On the other hand, if the model is estimated from standardized data the condition number will be lesser than 20 ($k_X = 11.741$) suggesting that the problem of collinearity is solved. However, this conclusion will be a contradiction to the rest of evidences since the symptoms of multicollinearity still persist as it is shown in the estimated model

$$\underset{\quad (0.6)\qquad (0.6)}{\widehat{Y} = 0.7007\,X_1 + 0.2654\,X_2,} \qquad R^2 = 0.9279,\ F_{2,7} = 45.07.$$

This example confirms that the Belsey's statement[12] *about centered data (the problem is that a low condition number for centered data need not indicate the absence of ill conditioning) is also satisfied for standardized data.*

3. About the limits of reducing the condition number by raise regression

This section studies the behavior of the condition number in the raise regression supposing two data transformation: unit length and standardization.

3.1. *Unit length transformation*

Before computing the condition number of the model (6), we have to modify the raised matrix

$$X(t) = \begin{pmatrix} 1 & x_{11} + te_1 & x_{21} \\ 1 & x_{12} + te_2 & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{1n} + te_n & x_{2n} \end{pmatrix}, \tag{8}$$

so that it has unit length

$$X_{u\ell}(t) = \begin{pmatrix} \frac{1}{\sqrt{n}} & \frac{x_{11} + te_1}{\sqrt{\sum (x_{1i} + te_i)^2}} & \frac{x_{21}}{\sqrt{\sum x_{2i}^2}} \\ \frac{1}{\sqrt{n}} & \frac{x_{12} + te_2}{\sqrt{\sum (x_{1i} + te_i)^2}} & \frac{x_{22}}{\sqrt{\sum x_{2i}^2}} \\ \vdots & \vdots & \vdots \\ \frac{1}{\sqrt{n}} & \frac{x_{1n} + te_n}{\sqrt{\sum (x_{1i} + te_i)^2}} & \frac{x_{2n}}{\sqrt{\sum x_{2i}^2}} \end{pmatrix}.$$

In this way, the columns of $X_{u\ell}(t)$ are unitary vectors with respect to the Euclidean metric on $\mathbb{R}^n$. This normalization yields to the following matrix

$$X_{u\ell}(t)^T \cdot X_{u\ell}(t) = \begin{pmatrix} 1 & \frac{\sum (x_{1i}+te_i)}{\sqrt{n}\sqrt{\sum (x_{1i}+te_i)^2}} & \frac{\sum x_{2i}}{\sqrt{n}\sqrt{\sum x_{2i}^2}} \\ \frac{\sum (x_{1i}+te_i)}{\sqrt{n}\sqrt{\sum (x_{1i}+te_i)^2}} & 1 & \frac{\sum x_{2i}\,(x_{1i}+te_i)}{\sqrt{\sum x_{2i}^2}\sqrt{\sum (x_{1i}+te_i)^2}} \\ \frac{\sum x_{2i}}{\sqrt{n}\sqrt{\sum x_{2i}^2}} & \frac{\sum x_{2i}\,(x_{1i}+te_i)}{\sqrt{\sum x_{2i}^2}\sqrt{\sum (x_{1i}+te_i)^2}} & 1 \end{pmatrix}.$$

Notice that, by (4),

$$\sum_{i=1}^{n} (x_{1i} + te_i) = \sum_{i=1}^{n} x_{1i} + t \sum_{i=1}^{n} e_i = \sum_{i=1}^{n} x_{1i}$$

and

$$\sum_{i=1}^{n} x_{2i}\,(x_{1i} + te_i) = \sum_{i=1}^{n} x_{2i}\,x_{1i} + t \sum_{i=1}^{n} x_{2i}\,e_i = \sum_{i=1}^{n} x_{2i}\,x_{1i}$$

do not depend on t. Hence, for all $t \in [0, +\infty)$,

$$B(t) = X_{u\ell}(t)^T \cdot X_{u\ell}(t) = \begin{pmatrix} 1 & a(t) & b \\ a(t) & 1 & c(t) \\ b & c(t) & 1 \end{pmatrix},$$

where

$$a(t) = \frac{\sum x_{1i}}{\sqrt{n}\sqrt{\sum (x_{1i} + te_i)^2}}, \qquad b = \frac{\sum x_{2i}}{\sqrt{n}\sqrt{\sum x_{2i}^2}}, \tag{9}$$

$$c(t) = \frac{\sum x_{2i}\,x_{1i}}{\sqrt{\sum x_{2i}^2}\sqrt{\sum (x_{1i} + te_i)^2}}. \tag{10}$$

Notice that numerators in (9)-(10) do not depend on t. We remark that

$$[\ \exists\ t_0 \geq 0 : a\,(t_0) = c\,(t_0)\ \ ] \quad \Leftrightarrow \quad \frac{\sum x_{1i}}{\sqrt{n}} = \frac{\sum x_{2i}\, x_{1i}}{\sqrt{\sum x_{2i}^2}}$$

$$\Leftrightarrow \quad [\ a\,(t) = c(t) \text{ for all } t \geq 0\ ]\,. \tag{11}$$

Furthermore, since $E \neq 0$,

$$\lim_{t\to+\infty} a\,(t) = \lim_{t\to+\infty} c\,(t) = 0. \tag{12}$$

The characteristic polynomial of each matrix $B\,(t)$ is:

$$p_{B(t)}\,(\lambda) = \det\left(B\,(t) - \lambda I_3\right) = \begin{vmatrix} 1-\lambda & a\,(t) & b \\ a\,(t) & 1-\lambda & c\,(t) \\ b & c\,(t) & 1-\lambda \end{vmatrix}$$
$$= -\lambda^3 + 3\lambda^2 + \left(a(t)^2 + b^2 + c(t)^2 - 3\right)\lambda$$
$$+ \left(2a(t)\, b\, c(t) - a(t)^2 - b^2 - c(t)^2 + 1\right).$$

As each matrix $B\,(t)$ must be symmetric and positive definite, it has three real eigenvalues that are strictly positive. If we denote by $\lambda_1\,(t)$, $\lambda_2\,(t)$ and $\lambda_3\,(t)$ the three positive eigenvalues of $B\,(t)$ such that

$$0 < \lambda_1\,(t) \leq \lambda_2\,(t) \leq \lambda_3\,(t)\,,$$

then we have that

$$p_{B(t)}\,(\lambda) = -\left(\lambda - \lambda_1\,(t)\right)\left(\lambda - \lambda_2\,(t)\right)\left(\lambda - \lambda_3\,(t)\right).$$

In other words,

$$\lambda^3 - 3\lambda^2 + \left(3 - a(t)^2 - b^2 - c(t)^2\right)\lambda + \left(a(t)^2 + b^2 + c(t)^2 - 2a(t)\, b\, c(t) - 1\right)$$
$$= \left(\lambda - \lambda_1\,(t)\right)\left(\lambda - \lambda_2\,(t)\right)\left(\lambda - \lambda_3\,(t)\right), \tag{13}$$

for all $\lambda \in \mathbb{R}$ and all $t \geq 0$. In particular

$$\lambda_1\,(t) + \lambda_2\,(t) + \lambda_3\,(t) = 3 \qquad \text{for all } t \geq 0,$$

which means that functions $\lambda_1, \lambda_2, \lambda_3 : [0, +\infty) \to (0, +\infty)$ are bounded because, for all $t \geq 0$ and all $j \in \{1, 2, 3\}$,

$$0 < \lambda_j\,(t) < \lambda_1\,(t) + \lambda_2\,(t) + \lambda_3\,(t) = 3. \tag{14}$$

Theorem 1 *Under the previous considerations,*

$$\lim_{t\to+\infty} \lambda_1\,(t) = 1 - |\,b\,|\,, \quad \lim_{t\to+\infty} \lambda_2\,(t) = 1 \quad \textit{and} \quad \lim_{t\to+\infty} \lambda_3\,(t) = 1 + |\,b\,|\,. \tag{15}$$

In particular,

$$\lim_{t\to+\infty} k_{X(t)} = \sqrt{\frac{1+|\,b\,|}{1-|\,b\,|}}\,.$$

Proof 1 *Notice that*

$$p_{B(t)}\,(1-b) = -\,(1-b)^3 + 3\,(1-b)^2 + \left(a(t)^2 + b^2 + c(t)^2 - 3\right)(1-b)$$
$$+ \left(2a(t)\, b\, c(t) - a(t)^2 - b^2 - c(t)^2 + 1\right)$$
$$= -b\, a(t)^2 + 2a(t)\, b\, c(t) - b\, c(t)^2 = -b\left(a\,(t) - c\,(t)\right)^2,$$

and, similarly, $p_{B(t)}(1+b) = b\left(a(t) - c(t)\right)^2$. *Furthermore,*

$$\lim_{\lambda\to-\infty} p_{B(t)}(\lambda) = +\infty \quad \textit{and} \quad \lim_{\lambda\to+\infty} p_{B(t)}(\lambda) = -\infty.$$

Taking into account (11), there is $t_0 \in [0,+\infty)$ *such that* $a(t_0) = c(t_0)$ *if, and only if,* $a(t) = c(t)$ *for all* $t \geq 0$. *Hence, we consider the following four cases.*

Case 1) $b = 0$. *In this case, the characteristic polynomial of* $B(t)$ *is:*

$$\begin{aligned} p_{B(t)}(\lambda) &= -\lambda^3 + 3\lambda^2 + \left(a(t)^2 + c(t)^2 - 3\right)\lambda + \left(1 - a(t)^2 - c(t)^2\right) \\ &= -(\lambda-1)\left(\lambda - 1 - \sqrt{a(t)^2 + c(t)^2}\right)\left(\lambda - 1 + \sqrt{a(t)^2 + c(t)^2}\right), \end{aligned}$$

whose roots are

$$\begin{aligned} &\lambda_1(t) = 1 - \sqrt{a(t)^2 + c(t)^2}, \qquad \lambda_2(t) = 1 \qquad \textit{and} \\ &\lambda_3(t) = 1 + \sqrt{a(t)^2 + c(t)^2}. \end{aligned}$$

Then (12) guarantees that (15) holds because $b = 0$.

Case 2) There is $t_0 \in [0,+\infty)$ such that $a(t_0) = c(t_0)$. *This case is equivalent to suppose that* $a(t) = c(t)$ *for all* $t \geq 0$. *Therefore*

$$\begin{aligned} p_{B(t)}(\lambda) &= -\lambda^3 + 3\lambda^2 + \left(a(t)^2 + b^2 + c(t)^2 - 3\right)\lambda \\ &\qquad + \left(2a(t)bc(t) - a(t)^2 - b^2 - c(t)^2 + 1\right) \\ &= -\lambda^3 + 3\lambda^2 + \left(2a(t)^2 + b^2 - 3\right)\lambda + \left(2a(t)^2 b - 2a(t)^2 - b^2 + 1\right) \\ &= -(\lambda - 1 + b)\left(\lambda - 1 - \tfrac{b - \sqrt{8a(t)^2 + b^2}}{2}\right)\left(\lambda - 1 - \tfrac{b + \sqrt{8a(t)^2 + b^2}}{2}\right). \end{aligned}$$

Since $a(t) \to 0$ *as* $t \to +\infty$,

$$1 + \frac{b - \sqrt{b^2}}{2} = \begin{cases} 1, & \textit{if } b \geq 0, \\ 1 - b, & \textit{if } b < 0, \end{cases}$$

and

$$1 + \frac{b + \sqrt{b^2}}{2} = \begin{cases} 1 + b, & \textit{if } b \geq 0, \\ 1, & \textit{if } b < 0, \end{cases}$$

then the three eigenvalues of $B(t)$ *converge, as* $t \to +\infty$, *to* $1 - b$, 1 *and* $1 + b$, *respectively.*

Case 3) $b > 0$ and $a(t) \neq c(t)$ for all $t \geq 0$. *Since*

$$\begin{aligned} &p_{B(t)}(0) = \lambda_1(t)\,\lambda_2(t)\,\lambda_3(t) > 0, \\ &p_{B(t)}(1 - |b|) = p_{B(t)}(1 - b) = -b\left(a(t) - c(t)\right)^2 < 0, \\ &p_{B(t)}(1 + |b|) = p_{B(t)}(1 + b) = b\left(a(t) - c(t)\right)^2 > 0, \\ &\lim_{\lambda\to+\infty} p_{B(t)}(\lambda) = -\infty, \end{aligned}$$

the alternate of the sign and the continuity of the polynomial function implies that

$$\lambda_1(t) \in (0, 1-b), \quad \lambda_2(t) \in (1-b, 1+b) \quad \textit{and} \quad \lambda_3(t) \in (1+b, +\infty).$$

Therefore, each matrix $B(t)$ *has three simple eigenvalues that satisfy*

$$0 < \lambda_1(t) < 1 - b < \lambda_2(t) < 1 + b < \lambda_3(t).$$

Taking limit as $t \to +\infty$ in (13), the bounded functions λ_1, λ_2 and λ_3 must converge to the roots of the polynomial

$$\begin{aligned}\lim_{t\to+\infty} \big[&-\lambda^3 + 3\lambda^2 + \left(a(t)^2 + b^2 + c(t)^2 - 3\right)\lambda \\ &+ \left(2a(t)\, b\, c(t) - a(t)^2 - b^2 - c(t)^2 + 1\right)\big] \\ = &-\lambda^3 + 3\lambda^2 + \left(b^2 - 3\right)\lambda + \left(1 - b^2\right) = -\left(\lambda - 1\right)\left(\lambda + b - 1\right)\left(\lambda - b - 1\right).\end{aligned}$$

Then (15) holds.

Case 4) $b < 0$ and $a\left(t\right) \neq c\left(t\right)$ for all $t \geq 0$. *In this case, since*

$$\begin{aligned}&p_{B(t)}\left(0\right) = \lambda_1\left(t\right)\lambda_2\left(t\right)\lambda_3\left(t\right) > 0,\\ &p_{B(t)}\left(1 - |\, b\,|\right) = p_{B(t)}\left(1 + b\right) = b\left(a\left(t\right) - c\left(t\right)\right)^2 < 0,\\ &p_{B(t)}\left(1 + |\, b\,|\right) = p_{B(t)}\left(1 - b\right) = -b\left(a\left(t\right) - c\left(t\right)\right)^2 > 0,\\ &\lim_{\lambda\to+\infty} p_{B(t)}\left(\lambda\right) = -\infty,\end{aligned}$$

we can repeat the arguments of the third case. □

The following result shows that raise regression is useful to reduce condition number. However, this process has a finite threshold.

Corollary 1 *Under the hypothesis of Theorem 1, for all $t \in [0, +\infty)$,*

$$k_{X(t)} \geq \lim_{s\to+\infty} k_{X(s)} = \sqrt{\frac{\sqrt{n\sum x_{2i}^2} + |\sum x_{2i}|}{\sqrt{n\sum x_{2i}^2} - |\sum x_{2i}|}} = k_\infty.$$

In particular, $k_X \geq k_\infty$.

Proof 2 *As we have shown in the four cases of the proof of the previous theorem, $\lambda_1\left(t\right) \leq 1 - |\, b\,|$ and $\lambda_3\left(t\right) \geq 1 + |\, b\,|$ for all $t \geq 0$, so*

$$k_{X(t)} = \sqrt{\frac{\lambda_3\left(t\right)}{\lambda_1\left(t\right)}} \geq \sqrt{\frac{1 + |\, b\,|}{1 - |\, b\,|}},$$

where

$$\sqrt{\frac{1 + |\, b\,|}{1 - |\, b\,|}} = \sqrt{\frac{1 + \left|\dfrac{\sum x_{2i}}{\sqrt{n}\sqrt{\sum x_{2i}^2}}\right|}{1 - \left|\dfrac{\sum x_{2i}}{\sqrt{n}\sqrt{\sum x_{2i}^2}}\right|}} = \sqrt{\frac{\sqrt{n\sum_{i=1}^{n} x_{2i}^2} + \left|\sum_{i=1}^{n} x_{2i}\right|}{\sqrt{n\sum_{i=1}^{n} x_{2i}^2} - \left|\sum_{i=1}^{n} x_{2i}\right|}}.$$

In particular,

$$k_X = k_{X(0)} \geq \sqrt{\frac{1 + |\, b\,|}{1 - |\, b\,|}} = \sqrt{\frac{\sqrt{n}\sqrt{\sum x_{2i}^2} + |\sum x_{2i}|}{\sqrt{n}\sqrt{\sum x_{2i}^2} - |\sum x_{2i}|}}.$$

□

Thus, for unit length data, $k_{X(t)}$ is continuous in $t = 0$ ($k_{X(0)} = k_X$) and always equal or higher than 1[1].

[1] If $k_\infty < 1$ then $|b| < 0$, which is not possible.

Remark 1 *For the sake of completeness, notice that* $|\,b\,| < 1$ *because of the Cauchy-Schwarz inequality applied to vectors* $u = X_0 = (1, 1, \ldots, 1)^T$ *and* $v = X_2 = (x_{21}, x_{22}, \ldots, x_{2n})^T$*. Indeed,*

$$\begin{aligned} &|\,(1,1,\ldots,1)\cdot(x_{21},x_{22},\ldots,x_{2n})\,| \le \|\,(1,1,\ldots,1)\,\|_2\;\|\,(x_{21},x_{22},\ldots,x_{2n})\,\|_2 \\ &\quad\Leftrightarrow\quad \left|\sum x_{2i}\,\right| \le \sqrt{n}\,\sqrt{\sum x_{2i}^2} \quad\Leftrightarrow\quad |\,b\,| \le 1, \end{aligned}$$

and the equality can only hold when u *and* v *are linearly dependent, but this is impossible since we assumed that the columns of* X *are linearly independent.*

3.2. *Decreasingness of condition number*

In this subsection we study the monotonicity of the function $k_{X(t)} : [0, +\infty) \to [1, +\infty)$. In particular we show that it is decreasing, so its limit is, in fact, its infimum. We remark that we will use the notions *increasingness* and *decreasingness* in a strict way, that is, a function $f : [0, +\infty) \to \mathbb{R}$ is *increasing* if $0 \le t_1 < t_2$ implies that $f\,(t_1) < f\,(t_2)$.

In order to study the monotonicity of the following functions, it is usual to assume that

$$\sum x_{1i} \ge 0, \quad \sum x_{2i} \ge 0 \quad \text{and} \quad \sum x_{1i}x_{2i} \ge 0. \tag{16}$$

These conditions follows from two facts: on the one hand, socioeconomic variables are usually non-negative; on the other hand, if this is not the case, we can replace variables X_1 and X_2 by raised variables $\mathbf{X}_1 = X_1 + c_1$ and $\mathbf{X}_2 = X_2 + c_2$, where constants c_1 and c_2 are such that $c_1 \ge x_{1i}$ and $c_2 \ge x_{2i}$ for all $i \in \{1, 2, \ldots, n\}$. Then model (1) can be replaced by

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + U = \beta_0 + \beta_1\,(\mathbf{X}_1 - c_1) + \beta_2\,(\mathbf{X}_2 - c_2) + U \\ &= (\beta_0 - \beta_1 c_1 - \beta_2 c_2) + \beta_1 \mathbf{X}_1 + \beta_2 \mathbf{X}_2 + U \\ &= \gamma_0 + \beta_1 \mathbf{X}_1 + \beta_2 \mathbf{X}_2 + U. \end{aligned}$$

If one of the constants given in (16) is zero, the eigenvalues of each matrix $B\,(t)$ can be easily computed.

Proposition 1 *If at least one of the following conditions holds:*

$$\begin{cases} b = 0, \\ \text{or there is } t_1 \in [0, +\infty) \text{ such that } a\,(t_1) = 0, \\ \text{or there is } t_2 \in [0, +\infty) \text{ such that } c\,(t_2) = 0, \end{cases}$$

then the three eigenvalues of each matrix $B\,(t)$ *are, for all* $t \in [0, +\infty)$*,*

$$\begin{aligned} &\lambda_1\,(t) = 1 - \sqrt{a\,(t)^2 + b^2 + c\,(t)^2}, \qquad \lambda_2\,(t) = 1, \\ &\lambda_3\,(t) = 1 + \sqrt{a\,(t)^2 + b^2 + c\,(t)^2}. \end{aligned}$$

Proof 3 *Notice that condition "there is* $t_1 \in [0, +\infty)$ *such that* $a\,(t_1) = 0$*" is equivalent to "*$a\,(t) = 0$ *for all* $t \in [0, +\infty)$*" because, in any case,* $\sum x_{1i} = 0$*. The same is true for function* c*. If at least one of the previous conditions holds, then* $2\,a(t)\,b\,c(t) = 0$ *for all* $t \in [0, +\infty)$*. Therefore, the*

characteristic polynomial of $B(t)$ *is*

$$\begin{aligned} p_{B(t)}(\lambda) &= -\lambda^3 + 3\lambda^2 + \left(a(t)^2 + b^2 + c(t)^2 - 3\right)\lambda \\ &\qquad + \left(2a(t)\,b\,c(t) - a(t)^2 - b^2 - c(t)^2 + 1\right) \\ &= -\lambda^3 + 3\lambda^2 + \left(a(t)^2 + b^2 + c(t)^2 - 3\right)\lambda + \left(1 - a(t)^2 - b^2 - c(t)^2\right) \\ &= -(\lambda - 1)\left(\lambda - 1 - \sqrt{a(t)^2 + b^2 + c(t)^2}\right) \\ &\qquad \cdot \left(\lambda - 1 + \sqrt{a(t)^2 + b^2 + c(t)^2}\right), \end{aligned}$$

whose roots are given by the above-mentioned functions. □

The previous result guarantees that if one of the numbers considered in (16) is zero, we can work out the three eigenvalues of each matrix $B(t)$, so the condition number of $B(t)$ is, for all $t \in [0, +\infty)$,

$$k_{X(t)} = \sqrt{\frac{\lambda_3(t)}{\lambda_1(t)}} = \sqrt{\frac{1 + \sqrt{a(t)^2 + b^2 + c(t)^2}}{1 - \sqrt{a(t)^2 + b^2 + c(t)^2}}}.$$

As functions a^2 and c^2 decrease to zero, then the function $t \mapsto k_{X(t)}$ is decreasing.

From now on, suppose that

$$\sum x_{1i} > 0, \quad \sum x_{2i} > 0 \quad \text{and} \quad \sum x_{1i}x_{2i} > 0. \tag{17}$$

In particular, henceforth, $b > 0$.

Proposition 2 *The function* $t \mapsto \sum (x_{1i} + te_i)^2$ *is increasing on* $[0, +\infty)$*, so the functions* a *and* c *are* C^∞ *and decreasing on* $[0, +\infty)$*.*

Furthermore, the function

$$t \mapsto 2\,a(t)\,b\,c(t)$$

is positive and decreasing on $[0, +\infty)$*, and the function*

$$t \mapsto -\left(a(t)^2 + b^2 + c(t)^2\right)$$

is negative and increasing on $[0, +\infty)$*, and both functions are* C^∞*.*

Proof 4 *Taking into account (4),* $n\,\mathrm{Var}(E) = \sum e_i^2$*, so (3) and (4) imply that* $\sum x_{1i}e_i = n\,\mathrm{Var}(E)$ *and* $\sum (x_{1i} + te_i)^2 = \sum x_{1i}^2 + n\,\mathrm{Var}(E)t\,(2 + t)$*. In particular, the function* $t \mapsto \sum (x_{1i} + te_i)^2$ *is* C^∞*, increasing on* $[0, +\infty)$ *and strictly positive. As a result, the functions* a *and* c *are decreasing on* $[0, +\infty)$*, and both are* C^∞ *on the same interval. Furthermore, it can be checked that*

$$2\,a(t)\,b\,c(t) = \frac{2\left(\sum x_{1i}\right)\left(\sum x_{2i}\right)\left(\sum x_{2i}\,x_{1i}\right)}{n^2\left(\sum x_{2i}^2\right)\mathrm{Var}(E)} \cdot \frac{1}{t^2 + 2t}, \tag{18}$$

$$\begin{aligned} -\left(a(t)^2 + b^2 + c(t)^2\right) &= -\frac{\left(\sum x_{2i}\right)^2}{n\left(\sum x_{2i}^2\right)} \\ &\quad - \frac{\left(\sum x_{1i}\right)^2\left(\sum x_{2i}^2\right) + n\,\left(\sum x_{2i}\,x_{1i}\right)^2}{n^2\,\left(\sum x_{2i}^2\right)\mathrm{Var}(E)} \cdot \frac{1}{t^2 + 2t}, \end{aligned}$$

so the first function is decreasing on $[0, +\infty)$ *and the second one is increasing on* $[0, +\infty)$*, and both functions are* C^∞ *on* $[0, +\infty)$*.* □

Next, let consider the functions $\mu_1, \mu_2, \mu_3 : [0, +\infty) \to \mathbb{R}$ given by

$$\mu_i(t) = \lambda_i(t) - 1 \quad \text{for all } t \in [0, \infty).$$

By (14), for all $t \in [0,\infty)$, $\mu_1(t) \leq \mu_2(t) \leq \mu_3(t)$ and $\mu_1(t)+\mu_2(t)+\mu_3(t) = 0$. Using the notation $\mu = \lambda - 1$, the previous functions are the solutions of the characteristic polynomial:

$$\begin{aligned}
&-\mu^3 + \left(a(t)^2 + b^2 + c(t)^2\right)\mu + 2\,a(t)\,b\,c(t) \\
&\quad = \begin{vmatrix} -\mu & a(t) & b \\ a(t) & -\mu & c(t) \\ b & c(t) & -\mu \end{vmatrix} = \det\left(B(t) - (\mu+1)I_3\right) \\
&\quad = -\left(\mu - \mu_1(t)\right)\left(\mu - \mu_2(t)\right)\left(\mu - \mu_3(t)\right) \\
&\quad = -\mu^3 + \left(\mu_1(t) + \mu_2(t) - \mu_3(t)\right)\mu^2 - \left[\,\mu_1(t)\,\mu_2(t) + \mu_1(t)\,\mu_3(t)\right. \\
&\qquad\qquad \left. + \mu_2(t)\,\mu_3(t)\right]\mu + \mu_1(t)\,\mu_2(t)\,\mu_3(t).
\end{aligned}$$

In particular

$$\begin{cases} \mu_1(t)\,\mu_2(t) + \mu_1(t)\,\mu_3(t) + \mu_2(t)\,\mu_3(t) = -\left(a(t)^2 + b^2 + c(t)^2\right), \\ \mu_1(t)\,\mu_2(t)\,\mu_3(t) = 2\,a(t)\,b\,c(t). \end{cases} \tag{19}$$

Theorem 2 *Under (17), the function $t \mapsto k_{B(t)}$ is decreasing on $[0,+\infty)$.*

Proof 5 *Taking into account that $b > 0$, in the proof of Theorem 1 we showed that if there is $t_0 \in [0,+\infty)$ such that $a(t_0) = c(t_0)$ then*

$$\lambda_1(t) = 1 - b, \qquad \lambda_2(t) = 1 + \frac{b - \sqrt{8a(t)^2 + b^2}}{2},$$
$$\lambda_3(t) = 1 + \frac{b + \sqrt{8a(t)^2 + b^2}}{2}.$$

In this case, the result follows from the fact that the function a (and also a^2) is decreasing on $[0,+\infty)$. Next, suppose that $a(t) \neq c(t)$ for all $t \in [0,+\infty)$. In the above mentioned proof we also showed that

$$\begin{aligned}
&p_{B(t)}(0) = \lambda_1(t)\,\lambda_2(t)\,\lambda_3(t) > 0, \\
&p_{B(t)}(1-b) = -b\left(a(t) - c(t)\right)^2 < 0, \\
&p_{B(t)}(1+b) = b\left(a(t) - c(t)\right)^2 > 0, \\
&\lim_{\lambda \to +\infty} p_{B(t)}(\lambda) < 0.
\end{aligned}$$

In particular,

$$\lambda_1(t) \in (0, 1-b), \quad \lambda_2(t) \in (1-b, 1+b) \quad \text{and} \quad \lambda_3(t) \in (1+b, 3).$$

This property guarantees that the eigenvalues of each matrix $B(t)$ are simple, so functions $\lambda_1, \lambda_2, \lambda_3 : [0,+\infty) \to (0,+\infty)$ are at least C^1 on $[0,+\infty)$. Then functions $\mu_1, \mu_2, \mu_3 : [0,+\infty) \to \mathbb{R}$ also are C^1 on $[0,+\infty)$. Since

$$\mu_1(t)\,\mu_2(t)\,\mu_3(t) = 2\,a(t)\,b\,c(t) > 0 \quad \text{for all } t \in [0,+\infty),$$

then functions μ_1, μ_2 and μ_3 has constant sign on $[0,+\infty)$. In fact,

$$\mu_1(t) \in (-1, -b), \quad \mu_2(t) \in (-b, b) \quad \text{and} \quad \mu_3(t) \in (b, 2).$$

As μ_1 is negative, μ_3 is positive and $\mu_1(t)\,\mu_2(t)\,\mu_3(t) = 2\,a(t)\,b\,c(t) > 0$ then necessarily the function μ_2 is negative on $[0,+\infty)$. Hence

$$\begin{aligned}
&\mu_1(t) \in (-1, -b), \quad \mu_2(t) \in (-b, 0), \quad \mu_3(t) \in (b, 2) \quad \text{and} \\
&\mu_1(t) < \mu_2(t) < 0 < \mu_3(t).
\end{aligned}$$

Next, we prove that

$$\mu_3'(t) \neq 0 \quad \textit{for all } t \in [0, +\infty).$$

Reasoning by contradiction, suppose that there is $t_0 \in [0, +\infty)$ *such that* $\mu_3'(t_0) = 0$. *Since* $\mu_1'(t) + \mu_2'(t) + \mu_3'(t) = 0$ *we deduce that* $\mu_2'(t_0) = -\mu_1'(t_0)$. *In this case, by (18), (19) and Proposition 2,*

$$\begin{aligned} 0 > \left.\frac{\partial}{\partial t}\right|_{t=t_0} [\ 2\,a(t)\,b\,c(t)\ ] &= \left.\frac{\partial}{\partial t}\right|_{t=t_0} [\ \mu_1(t)\,\mu_2(t)\,\mu_3(t)\ ] \\ &= \mu_1'(t_0)\,\mu_2(t_0)\,\mu_3(t_0) + \mu_1(t_0)\,\mu_2'(t_0)\,\mu_3(t_0) + \mu_1(t_0)\,\mu_2(t_0)\,\mu_3'(t_0) \\ &= \mu_1'(t_0)\,\mu_2(t_0)\,\mu_3(t_0) + \mu_1(t_0)\,(-\mu_1'(t_0))\,\mu_3(t_0) \\ &= \mu_1'(t_0)\,\mu_3(t_0)\,(\mu_2(t_0) - \mu_1(t_0)). \end{aligned}$$

As $\mu_2(t_0) - \mu_1(t_0) > 0$ *and* $\mu_3(t_0) > 0$, *we deduce that*

$$\mu_1'(t_0) < 0. \tag{20}$$

On the other hand, also by (18), (19) and Proposition 2,

$$\begin{aligned} 0 < \left.\frac{\partial}{\partial t}\right|_{t=t_0} &\left[\ -\left(a(t)^2 + b^2 + c(t)^2\right)\ \right] \\ = \left.\frac{\partial}{\partial t}\right|_{t=t_0} &[\ \mu_1(t)\,\mu_2(t) + \mu_1(t)\,\mu_3(t) + \mu_2(t)\,\mu_3(t)\ ] \\ = \mu_1'(t_0)\,\mu_2(t_0) &+ \mu_1(t_0)\,\mu_2'(t_0) + \mu_1'(t_0)\,\mu_3(t_0) + \mu_1(t_0)\,\mu_3'(t_0) \\ &+ \mu_2'(t_0)\,\mu_3(t_0) + \mu_2(t_0)\,\mu_3'(t_0) \\ = \mu_1'(t_0)\,\mu_2(t_0) &+ \mu_1(t_0)\,(-\mu_1'(t_0)) + \mu_1'(t_0)\,\mu_3(t_0) + (-\mu_1'(t_0))\,\mu_3(t_0) \\ = \mu_1'(t_0)\,\mu_2(t_0) &- \mu_1(t_0)\,\mu_1'(t_0) = \mu_1'(t_0)\,(\mu_2(t_0) - \mu_1(t_0)), \end{aligned}$$

but from this inequality we deduce that $\mu_1'(t_0) > 0$, *which contradicts (20). As a result, we deduce that* $\mu_3'(t) \neq 0$ *for all* $t \in [0, +\infty)$. *Reasoning in the same way, we can also deduce that* $\mu_1'(t) \neq 0$ *for all* $t \in [0, +\infty)$. *As a consequence,* μ_1 *and* μ_3 *are strictly monotone functions. Taking into account that*

$$\mu_1(t) \in (-1, -b) \quad \textit{and} \quad \mu_3(t) \in (b, 2) \quad \textit{for all } t \in [0, +\infty),$$

$$\lim_{t\to+\infty} \mu_1(t) = \lim_{t\to+\infty} \lambda_1(t) - 1 = -b, \quad \lim_{t\to+\infty} \mu_3(t) = \lim_{t\to+\infty} \lambda_3(t) - 1 = b,$$

we conclude that μ_1 *is increasing and* μ_3 *is decreasing on* $[0, +\infty)$. *As a consequence, the function*

$$t \mapsto k_{X(t)} = \sqrt{\frac{\lambda_3(t)}{\lambda_1(t)}} = \sqrt{\frac{1 + \mu_3(t)}{1 + \mu_1(t)}}$$

is decreasing on $[0, +\infty)$. □

3.3. *Standardization transformation*

In this case, before computing the condition number of the model (6), we have to modify the raised[2] matrix (8) by

$$X_s(t) = \begin{pmatrix} \dfrac{x_{11} + t\,e_1 - \overline{X}_1}{\sqrt{n\ \mathrm{Var}(\widetilde{X}_1(t))}} & \dfrac{x_{21} - \overline{X}_2}{\sqrt{n\ \mathrm{Var}(X_2)}} \\ \dfrac{x_{12} + t\,e_2 - \overline{X}_1}{\sqrt{n\ \mathrm{Var}(\widetilde{X}_1(t))}} & \dfrac{x_{22} - \overline{X}_2}{\sqrt{n\ \mathrm{Var}(X_2)}} \\ \vdots & \vdots \\ \dfrac{x_{1n} + t\,e_n - \overline{X}_1}{\sqrt{n\ \mathrm{Var}(\widetilde{X}_1(t))}} & \dfrac{x_{2n} - \overline{X}_2}{\sqrt{n\ \mathrm{Var}(X_2)}} \end{pmatrix}.$$

so that $X_s(t)^T X_s(t)$ is the correlation matrix

$$X_s(t)^T X_s(t) = \begin{pmatrix} 1 & \mathrm{corr}(\widetilde{X}_1(t), X_2) \\ \mathrm{corr}(\widetilde{X}_1(t), X_2) & 1 \end{pmatrix},$$

where, since $\mathrm{cov}(E, X_2) = 0$,

$$\begin{aligned} \mathrm{corr}(\widetilde{X}_1(t), X_2) &= \frac{\mathrm{cov}(\widetilde{X}_1(t), X_2)}{\sqrt{\mathrm{Var}(\widetilde{X}_1(t))}\sqrt{\mathrm{Var}(X_2)}} \\ &= \frac{\mathrm{cov}(X_1, X_2)}{\sqrt{\mathrm{Var}(X_1) + (t^2 + 2t)\cdot \mathrm{Var}(E)}\sqrt{\mathrm{Var}(X_2)}}. \end{aligned}$$

The condition number is

$$\begin{aligned} k_{X(t)} &= \sqrt{\frac{1 + \mathrm{corr}(\widetilde{X}_1(t), X_2)^2}{1 - \mathrm{corr}(\widetilde{X}_1(t), X_2)^2}} \\ &= \sqrt{\frac{(\mathrm{Var}(X_1) + (t^2 + 2t)\cdot \mathrm{Var}(E))\cdot \mathrm{Var}(X_2) + \mathrm{cov}(X_1, X_2)^2}{(\mathrm{Var}(X_1) + (t^2 + 2t)\cdot \mathrm{Var}(E))\cdot \mathrm{Var}(X_2) - \mathrm{cov}(X_1, X_2)^2}}. \end{aligned}$$

And, in such a case,

$$\begin{aligned} \lim_{t\to+\infty} k_{X(t)} &= \lim_{t\to+\infty} \sqrt{\frac{\left(\frac{\mathrm{Var}(X_1) + (t^2+2t)\cdot \mathrm{Var}(E)}{t^2}\right)\cdot \mathrm{Var}(X_2) + \frac{\mathrm{cov}(X_1, X_2)^2}{t^2}}{\left(\frac{\mathrm{Var}(X_1) + (t^2+2t)\cdot \mathrm{Var}(E)}{t^2}\right)\cdot \mathrm{Var}(X_2) - \frac{\mathrm{cov}(X_1, X_2)^2}{t^2}}} \\ &= \sqrt{\frac{\mathrm{Var}(E)\cdot \mathrm{Var}(X_2)}{\mathrm{Var}(E)\cdot \mathrm{Var}(X_2)}} = 1. \end{aligned}$$

Also, $k_{X(t)}$ is decreasing in t since

$$\frac{\partial\, k_X}{\partial t}(t) = -\frac{\mathrm{Var}(X_2)\cdot \mathrm{cov}(X_1, X_2)^2 \cdot (2t+2)\,\mathrm{Var}(E)}{\sqrt{h(t)}\cdot (g(t)\cdot \mathrm{Var}(X_2) - \mathrm{cov}(X_1, X_2)^2)^2} < 0,$$

where

$$\begin{aligned} h(t) &= \frac{\big(\mathrm{Var}(X_1) + (t^2 + 2t)\cdot \mathrm{Var}(E)\big)\cdot \mathrm{Var}(X_2) + \mathrm{cov}(X_1, X_2)^2}{(\mathrm{Var}(X_1) + (t^2 + 2t)\cdot \mathrm{Var}(E))\cdot \mathrm{Var}(X_2) - \mathrm{cov}(X_1, X_2)^2}, \\ g(t) &= \mathrm{Var}(X_1) + (t^2 + 2t)\cdot \mathrm{Var}(E). \end{aligned}$$

[2]Without loss of generality, we consider that the first variable is raised.

Thus, for standardized data, $k_{X(t)}$ is continuous in $t = 0$ (where $k_{X(0)} = k_X$), it is decreasing on t and it is always greater than or equal to 1.

Remark 2 *If we have considered typified data $X_{typ}(t)$, then we would have obtained the same results because $X_{typ}(t)^T X_{typ}(t) = n X_s(t)^T X_s(t)$, and the condition number would not have changed.*

4. Illustrative example

In this section we illustrate our study by describing an example in which raise regression can be useful in order to reduce the effects of collinearity (see Section 2) and, consequently, the condition number of the matrix associated to the problem.

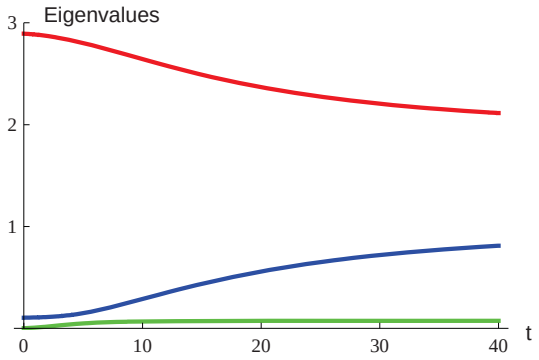

Fig. 1. Evolution of eigenvalues of the matrix $\widetilde{X}(t)$.

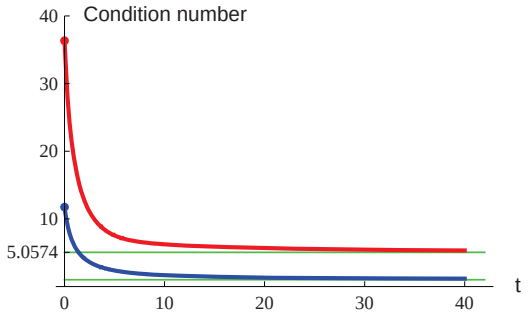

Fig. 2. Evolution of the condition number of the matrix $\widetilde{X}(t) = (\widetilde{X}_1(t), X_2)$ and its limit for unit length (red) and standardized data (blue).

For the data used previously by Hurvich[13], if we raise variable X_1 by considering the new explanatory variable $\widetilde{X}_1(t) = X_1 + tE$, where $t \in [0, +\infty)$, we observe that the highest eigenvalue of $X(t)$ decreases and the lowest one increases when $t \to +\infty$ (see Figure 1). Thus, for unit length data, condition number stabilizes itself around the value:

$$\lim_{t \to +\infty} k_{X(t)} = \sqrt{\frac{23759 + 444\sqrt{2846}}{1855}} \approx 5.0574.$$

Figure 2 shows the evolution of condition number depending on t. For $t = 1$, the condition number is 18.453, so multicollinearity can be considered *moderate*, and for $t = 3$, the condition number is less than 10. For $t = 50$, it is 5.240596, which is very close to its lower bound.

If variable X_2 is raised depending on t, then the limit of the condition number would have been 4.78166 (see Figure 3). Note that the results are very similar to the one obtained when the first variable is raised. In addition, in both cases, the condition number is continuous in $t = 0$ ($k_{X(0)} = k_X$), decreasing in t and it is always greater than the established threshold.

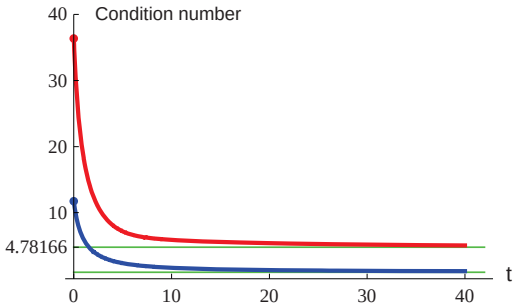

Fig. 3. Evolution of the condition number of the matrix $\widetilde{X}(t) = (X_1, \widetilde{X}_2(t))$ and its limit for unit length (red) and standardized data (blue).

Finally, it is observed that, when data are standardized, the condition number is continuous in $\lambda = 0$ ($k_{X(0)} = k_X$), decreasing in λ and its limit is one when $t \to +\infty$.

5. Conclusions and prospect work

In this manuscript we have described why the raise regression can be considered as an appropriate methodology in order to reduce the approximate multicollinearity that naturally appears in problems of linear estimation when three explanatory variables are involved. In general, its application reduces the condition number of the matrix associated to data set. Nevertheless, this procedure has a threshold: although we can employ values of t arbitrarily large in order to separate the columns of X, the condition number will never be less than a constant that can be easily worked out by using the elements of the associate matrix X.

On the other hand, the problem about that a low condition number for centered or standardized data need not indicate the absence of ill conditioning commented by Belsey[12] is still verified in the raise regression. Therefore, it is preferable to calculate the condition number from normalized data.

Immediately the following questions arise when we employ the raise regression technique:

Open problem 1: does a limit exist on the condition number when more than three explanatory variables are considered? If so, is this limit computable by a simple calculation, directly related to the matrix X? This analysis must be done only with normalized data, that is, with $X_{u\ell}$.

Open problem 2: Does a threshold appear when we consider another measures of the impact of collinearity (like the *variance inflation factor*)?

From our point of view, it is worth considering these problems in future work.

References

1. A. E., Hoerl and R. W., Kennard, Ridge regression: Biased estimation for nonorthogonal problems, *Technometrics*, (1970), 12, 55-67.

2. A. E., Hoerl and R. W., Kennard, Ridge regression: Applications to nonorthogonal problems, *Technometrics*, (1970), 12, 69-82.
3. G.C., McDonald, Ridge regression, *Wiley Interdisciplinary Reviews: Computational Statistics*, (2009), 1, 93–100.
4. C.B. García, J. García and J. Soto, The raise method: An alternative procedure to estimate the parameters in presence of collinearity, *Quality and Quantity*, 45, (2010), 403–423.
5. R. Salmerón, C.B. García, J. García and M.M. López, The raise estimators. Estimation, inference and properties, *Communications in Statistics - Theory and Methods*, 46 (13) (2017), 6446–6462.
6. D.W. Marquardt, Generalized inverses, ridge regression, biased linear estimation and nonlinear estimation, *Technometrics*, 12 (3) (1970), 591-612.
7. H. Theil, *Principles of econometrics* (Wiley, New York, 1971).
8. J. Fox and G. Monette, Generalized collinearity diagnostics, *Journal of the American Statistical Association*, 87 (1992), 178–183.
9. R.M. O'Brien, A caution regarding rules of thumb for variance inflation factors, *Quality and Quantity*, 41 (2007), 673–690.
10. C.B. García, R. Salmerón, J. García and M.M. López, The condition number in the raise regression, *The 4th Advanced Research in Scientific Areas*, (2015), 100–103.
11. D.A. Belsley, E. Kuh and R.E. Welsch, *Regression Diagnostics: Identifying Influential Observations and Sources of Collinearity* (New York: John Wiley, 1980).
12. D.A. Belsley, Demeaning conditioning diagnostics through centering, *The American Statistician*, 38 (2) (1984), 73–77.
13. C. Hurvich, Multicollinearity, In *Handouts about regression (chapter 19)*, available on `http://pages.stern.nyu.edu/~churvich/Regress/Handouts/Chapt19.pdf`.