Hetesi, E. – Kürtösi, Zs. (eds) 2011: The diversity of research at the Szeged Institute of Business Studies. JATEPress, Szeged, pp. 145–154.

The elliptical model of multicollinearity and the Petres' Red indicator

Péter Kovács

One possible method for modelling multicollinearity is to examine the orthogonality of explanatory variables, which is the "stretching" of the space of explanatory variables. The question rightly arises whether multicollinearity can be modelled in a different way.

As a new approach, the elliptical model of multicollinearity can be formulated on the basis of Petres' Red indicator. Parallel with the increase in the extent of the mean correlation of the variables, the "possible eigenvalues" are situated on an m-dimensional sphere with a greater radius. The "possible eigenvalues" are situated on a segment of the m-dimensional sphere in such a way that with a fixed Red value they are located on an (m–1)-dimensional ellipsoid.

Unfortunately, the higher the dimension number of the model, the more conditions have to be given for determining and studying the range of "possible eigenvalues". Therefore, the detailed examination of this range and of the elliptical curves was carried out only for three explanatory variables.

Keywords: redundancy of databases; multicollinearity.

1. Introduction

In the current globalizing world, decision makers have an increased need for information. However, the great increase in the quantity of data is not automatically accompanied by an appropriate increase in information. Contrarily, the problem that decision makers have to face today is not the lack, but the abundance of information. This massive amount of present data frequently has little informational content, which means that redundancy is high. Redundancy means "superfluous" data which does not convey new or noteworthy information in terms of the examination. For this reason, the information content of metric data is an essential issue in empirical analyses. This is particularly true for the application of linear regression models. In the case of linear regression models, multicollinearity can be interpreted as a type of redundancy. With matrix algebraic notation this can be written in the form of $\tilde{\mathbf{y}} = \tilde{\mathbf{X}}\tilde{\boldsymbol{\beta}} + \tilde{\boldsymbol{\varepsilon}}$, where $\tilde{\mathbf{y}}$ is the n component column vector of the dependent variable;

$\tilde{\mathbf{X}}$ is the matrix of explanatory variables consisting of row n and column (m+1), where the first column is always an $\tilde{\mathbf{x}}_0$ sum vector; $\tilde{\boldsymbol{\beta}}$ is the (m+1) component column vector of the model parameters unknown to us; m is the number of explanatory variables (explanatory variables); $\tilde{\boldsymbol{\varepsilon}}$ is the n component column vector of the error term (Hajdu 2003).

The problems of multicollinearity are almost always encountered in the course of economic analyses. The concept of multicollinearity is apparently uniform in literature. Definitions usually differ from each other in one word, but this entails significant changes in content.

1.1. Multicollinearity

Multicollinearity as an expression was first used by Ragnar Frisch. He used it for the description of cases in which one variable was present in several relations. In his examinations he did not distinguish dependent variables from explanatory variables. He assumed that the measurement of all variables was erroneous; the correlation between the actual values of the variables had to be estimated on this basis (Kovács 2008).

It is considered superficial when multicollinearity is defined as the absence of the independence of explanatory variables. This definition is problematic because it is defined ambiguously without the independent meaning of the explanatory variables clarified. Does it mean their linear independence or possibly their independence in the statistical sense?

One of the primary conditions of the standard linear regression model is the linear independence of the explanatory variables (Kennedy 2003). Therefore in certain sources, multicollinearity is interpreted as the absence of the linear independence of explanatory variables. This approach can be regarded as a special case of multicollinearity, which is called extreme multicollinearity. This case does not pose special problems in practice as it is easily manageable.

In the course of empirical analyses, cases close to extreme multicollinearity are frequently encountered; when the variances of individual estimated parameters are considerably increased as compared to the variance of the error term. The great majority of literature on multicollinearity deals with this case. However, it is best to note that multicollinearity could mean a much more general phenomenon, namely the correlation of explanatory variables. Naturally, the special cases of this definition would convey the content meant by multicollinearity to everybody.

1.2. Red indicator

Petres' Red is one possibility for measuring the proportion of data with a useful content in respect of the estimator $\hat{\tilde{\boldsymbol{\beta}}} = (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{y}}$. Petres' Red is a new possible indicator of redundancy and thus of multicollinearity. The Red indicator is defined by using the eigenvalues λ_j (j=1,2,…,m) of the correlation matrix R of the explanatory variables. The Red indicator is based on the following train of thought. If the database serving as the source of the explanatory variables is redundant in respect of estimator $\tilde{\boldsymbol{\beta}}$, that is if the correlation of the data is considerable, not all the data will have a useful content. The smaller the proportion of the data with a useful content is, the greater the extent of redundancy will be. The greater the dispersion of the eigenvalues is, the greater the correlation of the explanatory variables in the database will be. There are two extreme cases, either all the eigenvalues are equal to each other (that is their value is one) or all the eigenvalues with the exception of one equal zero. The extent of dispersion can be quantified with the relative dispersion of the eigenvalues or with their dispersion (being equal in this case).

$$\mathrm{v}_\lambda = \frac{\sigma_\lambda}{\bar{\lambda}} = \frac{\sqrt{\dfrac{\sum_{j=1}^{m}(\lambda_j - \bar{\lambda})^2}{m}}}{\dfrac{\sum_{j=1}^{m}\lambda_j}{m}} = \frac{\sqrt{\dfrac{\sum_{j=1}^{m}(\lambda_j - 1)^2}{m}}}{\dfrac{m}{m}} = \sigma_\lambda \tag{1}$$

In order to make the redundancy of various databases comparable, the above indicator has to be normalized. As the eigenvalues are nonnegative, normalization is carried out with value $\sqrt{m-1}$ because of the relationship $0 \le \mathrm{v}_\lambda \le \sqrt{m-1}$ concerning relative dispersion.

The indicator obtained in this way can be used to quantify the extent of redundancy, and the Red indicator can be defined with its help as follows.

$$\mathrm{Re}\,d = \frac{\mathrm{v}_\lambda}{\sqrt{m-1}} \tag{2}$$

In the case of the absence of redundancy, the value of the above indicator is zero or zero percent, while in the case of maximum redundancy, it is one or one hundred percent.

The Red indicator measures the redundancy of the examined database of the given size. When the redundancies of two or more databases of different sizes are compared, the Red indicators can only be used to determine how redundant individual databases are, but one cannot make a direct statement as to which of these has more useful data.

The Red indicator can be expressed without knowing the eigenvalues of the correlation matrix of the explanatory variables, merely as the quadratic mean of the correlation coefficients (Kovács et al 2005).

$$\operatorname{Re} d = \sqrt{\frac{\sum_{i=1}^{m}\sum_{\substack{j=1 \\ j\neq i}}^{m} r_{ij}^2}{m(m-1)}} \,. \tag{3}$$

This means that this indicator shows not only the proportion of the data with a useful content in respect of the estimator $\tilde{\boldsymbol{\beta}}$, but also the mean correlation of the explanatory variables. It ensues from the definition of the indicator and from formula (3) that, as compared to other indicators based on eigenvalues, the advantage of this indicator is that it considers all the eigenvalues in such a way that its value is influenced by all the eigenvalues with the same weight. It also considers all the pair correlation of the explanatory variables, thus the Red indicator definitely represents an advance compared to the research of multicollinearity to date. Various cases of extreme multicollinearity can also be distinguished with the help of the indicator, as it can also be used when one of the eigenvalues is zero.

The correlation of the variable pairs and the correlation of the variable groups may pose a problem during the examination of multicollinearity. However, no detailed methodology has been worked out for this yet. A possible solution to the problem could be the use of canonical correlation analysis in conjunction with the redundancy index. It has been established that one special case of this can be measured with the Red indicator, while another special case with the help of the harmonic mean of the VIFj values.

2. New modelling possibilities of multicollinearity

The question may arise how multicollinearity can be modelled. By plotting the explanatory variables as vectors, conjectures can be drawn up concerning the presence of multicollinearity.

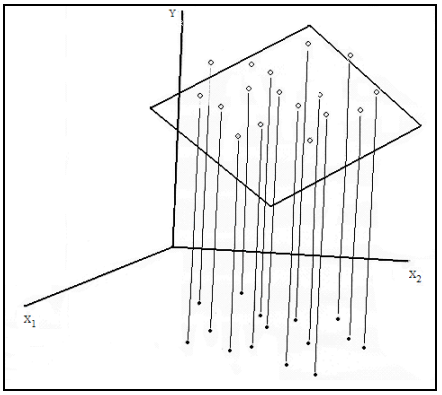
2.1. Orthogonality of variables

One of the most frequently mentioned possibilities for modelling is to examine the orthogonality of explanatory variables. If the vectors plotted are orthogonal, meaning that the space of explanatory variables is stretched maximally, there is no multicollinearity in the model. The smaller the stretching of the space, the greater the extent of multicollinearity there will be. The question rightly arises whether multicollinearity can be modelled in a different way.

2.2. Projection

Another possibility is to examine the projections of the regression plane, hyper plane, in each x_i-x_j plane projection. For instance, with two explanatory variables Figure 1 shows that – in the case of the statistically insignificant correlation of explanatory variables – the variance of the estimated parameters is considerably smaller compared to the variance calculated in the case of significant correlation. This is because, in the first case, the "cloud of points" of the data base is dispersed in the x_1-x_2 plane projection in every dimension, and thus the fitted regression plane is stable (Tričković 1976).

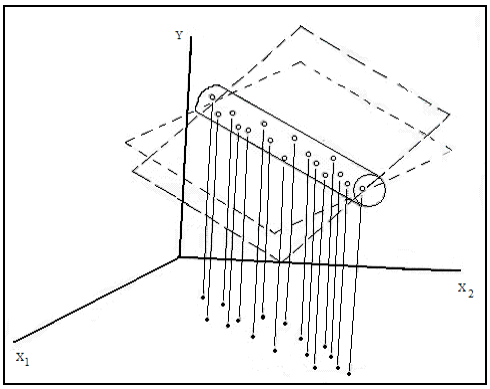
Figure 1. Stable regression plane in the case of the non-significant correlation of explanatory variables (m=2)



Source: Tričković (1976)

At any rate, the "cloud of points" in Figure 2 is not dispersed in the x_1-x_2 plane projection in every dimension, thus the fitted plane is tilted easily and fitting becomes instable. This way of plotting is very work-intensive and only the pair correlation of the explanatory variables can be illustrated.

Figure 2. Instable regression plane in the case of significant multicollinearity (m=2)



Source: Tričković (1976)

2.3. The elliptical model of multicollinearity

Starting from the definition of the Red indicator, a different type of model for multicollinearity can also be given. The following relationship is obtained by rearranging formula (2) of the Red indicator.

$$\sum_{i=1}^{m}(\lambda_i - 1)^2 = (\sqrt{m(m-1)}\,\mathrm{Re}\,d)^2 \tag{4}$$

Equation (4) is the equation of a sphere the radius of which is $\sqrt{m(m-1)}\,\mathrm{Re}\,d$, and every coordinate of its centre point is one. If the mean correlation of the variables is zero, that is there is no correlation between the explanatory variables, then the sphere is reduced to the single point each coordinate of which is one. The greater the extent of the mean correlation of the variables is, the greater the radius of the sphere will be, and specifically the greater the "inflation" of the sphere will be.

If the mean correlation of the variables is one, that is the absolute value of the correlation coefficient between each explanatory variable pair is one, the radius of the sphere is $\sqrt{m(m-1)}$.

Naturally, not each point of the spheres represents an existing correlation structure. By definition, combinations of eigenvalues can also be found on the spheres which are not possible in the case of correlation matrixes. The question is which points of the spheres represent an existing correlation structure. In the following example, these eigenvalue combinations are going to be called "possible eigenvalues" for clarity purposes. In order to examine "possible eigenvalues", the properties of the eigenvalues of the correlation matrix need to be considered. As the sum of eigenvalues equals the number of explanatory variables, or the dimension of the sphere, it is certain that "possible eigenvalues" are located on the intersections of equation (4) and of (5).

$$\sum_{i=1}^{m} \lambda_i = m \tag{5}$$

In the following, without restriction of generality, it can be assumed that:

$$\lambda_{\max} = \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m = \lambda_{\min}.$$

By calculating the smallest eigenvalue from formula (5) and by substituting it into equation (4), the following equation is obtained:

By rearranging the equation the following equation is obtained:

$$\sum_{i=1}^{m-1} \lambda_i^2 - m\sum_{i=1}^{m-1} \lambda_i + \sum_{i=1}^{m} \sum_{\substack{j=1 \\ j>i}}^{m-1} \lambda_i \lambda_j + \frac{m(m-1)}{2} = \frac{m(m-1)}{2} \operatorname{Re} d^2 \tag{6}$$

Equation (6) means that the "possible eigenvalues" – with a given Red indicator – are contained in an (m–1)-dimensional ellipsoid. In the special case of three explanatory variables, some points of the ellipses mean the "possible eigenvalues". The elliptical name of the model ensues from the nature of the curves. It can be seen that on the basis of equation (6) the representation of the eigenvalues is obtained in a dimension lower by one compared to the number of the eigenvalues.

If the number of explanatory variables is three, equation (6) can be written in the following form:

$$\lambda_1^2 + \lambda_2^2 - 3\lambda_1 - 3\lambda_2 + \lambda_1\lambda_2 + 3 = 3\,\mathrm{Re}\,d^2 \tag{7}$$

In the case of three explanatory variables, the range of "possible eigenvalues" can be delimited – in addition to formula (5) – by giving three more conditions.

- With the consideration of the relation between the eigenvalues: .
- With the consideration of the relation between the eigenvalues: $\lambda_2 \geq \lambda_3 = 3 - \lambda_1 - \lambda_2$, therefore $\lambda_2 \geq \frac{3-\lambda_1}{2}$.
- Moreover: $\lambda_1 + \lambda_2 \leq 3$. This condition already includes conditions $\lambda_1 + \lambda_3 \leq 3$ and $\lambda_2 + \lambda_3 \leq 3$.

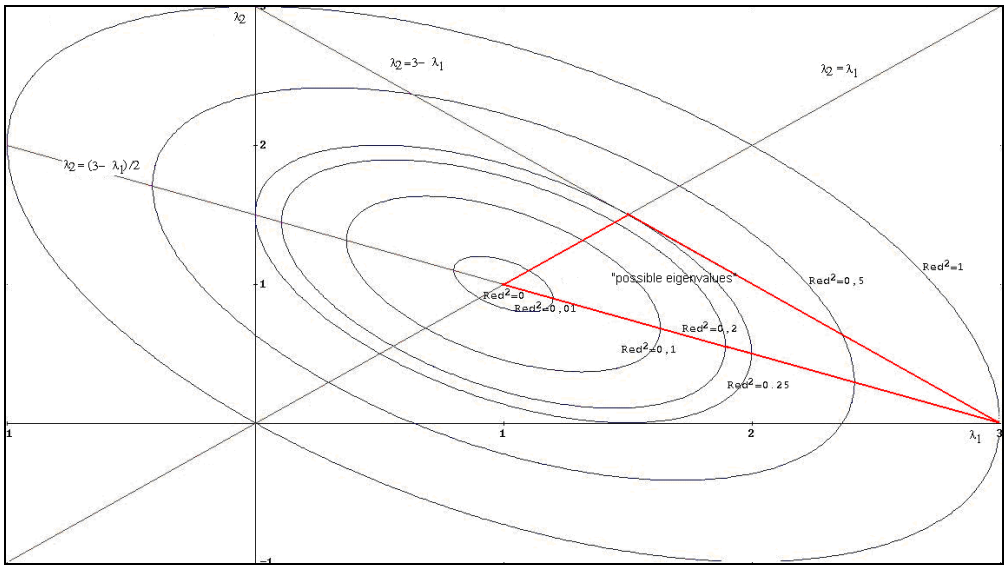
Some of the level lines with different Red values are illustrated in Figure 3. Plotting is made as the function of the two largest eigenvalues.

Thus, in the case of three dimensions, "possible eigenvalues" can be found in the triangle of Figure 3. The cases of extreme multicollinearity are given by the interceptions of the ellipses and line $\lambda_2 = 3 - \lambda_1$. This also shows that various cases of multicollinearity can also be distinguished with the help of the Red indicator.

In the case of higher dimensions – in line with the above train of thought – the great number of conditions makes it difficult to plot the "possible eigenvalues". Therefore, in higher dimensions, all we can state for certain is that the radius of the examined m-dimensional sphere will increase with the increase of the mean correlation of the variables. Furthermore, with a fixed value of the Red indicator, the "possible eigenvalues" are located on the surface part of a (m–1)-dimensional ellipsoid.

A similar plotting exists in literature for linear correlation coefficients. These form an elliptope (Bolla–Krámli 2005). In higher dimensions such an approach to plotting is unhandy.

Figure 3. The elliptic model of multicollinearity in the case of three explanatory variables



Source: own construction

In the following I am going to present some characteristics of the ellipses in the case of three explanatory variables.

1. If the extent of the correlation of the explanatory variables is greater, the section of the ellipses falling into the "possible range" is shifted to the right.
2. Empirical experience shows that, with a given Red value, the increase of eigenvalue λ_1 is accompanied by a greater decrease of eigenvalue λ_2, therefore the smallest eigenvalue will also increase as the sum of eigenvalues is three.
3. The correlation matrixes in which all the elements outside the diagonal are the same – in this case $\mathrm{Re}\, d = r = \mathbf{R}_{ij(i\neq j)}$ – are located on the lower boundary of the possible range. Then the determinant of the correlation matrix equals the value of the $1 - 3\,\mathrm{Re}\, d^2 + 2\,\mathrm{Re}\, d^3$.
4. Empirical experience shows that the product of eigenvalues decreases when moving upwards on a given ellipse, that is the determinant of the correlation matrix is becoming smaller and smaller. Thus, with a given

Red value, the determinant of the correlation matrix falls into the range of $[\max(1-3\,\mathrm{Re}\,d^2-2\,\mathrm{Re}\,d^3;0)\,;\,1-3\,\mathrm{Re}\,d^2+2\,\mathrm{Re}\,d^3]$ on a fixed ellipse.

3. Conclusions

As a new approach, the elliptical model of multicollinearity has been formulated. Parallel with the increase in the extent of the mean correlation of the variables, the "possible eigenvalues" are situated on an m-dimensional sphere with a greater radius. The "possible eigenvalues" are situated on a segment of the m-dimensional sphere in such a way that with a fixed Red value they are located on an (m–1)-dimensional ellipsoid. Unfortunately, the higher the dimension number of the model is, the more conditions have to be given for determining and studying the range of "possible eigenvalues". Therefore, the detailed examination of this range and of the elliptical curves was carried out only for three explanatory variables.

References

Bolla, M. – Krámli, A. 2005: *Statisztikai következtetések elmélete*. Typotex Kiadó, Budapest.

Hajdu, O. 2003: *Többváltozós statisztikai számítások*, Központi Statisztikai Hivatal Budapest.

Kennedy, P. 2003: *A Guide to econometrics*. 5. Edition, MIT, Cambridge.

Kovács, P. – Petres, T. – Tóth, L. 2005: A new measure of multicollinearity in linear regression models. *International Statistical Review (ISR)*, Vol. 73., No. 3., Voorburg, The Netherlands, pp. 405–412.

Kovács, P. 2008: A multikollinearitás vizsgálata lineáris regressziós modellekben. *Statisztikai szemle*, 86. évf., 1. sz., 38–67. o.

Tričković, V. 1976: *Teorijski modeli i metodi kvantitativnog istraživanja tržišta*. Institut za ekonomiku industrije, Beograd.