DOI: 10.33278/SAE-2018.eng.196-198

JEL C15

ANALYSIS OF INFORMATION CONTENT OF METRIC DATA WHEN CONSTRUCTING MODELS OF LINEAR REGRESSION

ORLOVA IRINA (ORCID 0000-0001-5397-2450)[1],
IOUDINA VERA[2]

[1]Financial University under the Government of the Russian Federation,
[2] Texas State University

Abstract. This article is devoted to the analysis of methods aimed at solving the problem of multicollinearity of data arising due to the high information redundancy of metric data. Decision making in modern conditions is based on the analysis of huge amounts of data, which often have only a small amount of informational content, which means that information redundancy is high.

In the case of a linear regression model, multicollinearity can be interpreted as a type of redundancy. The possibility of using the red indicator PETRES Red ¾ red indicator to measure the proportion of useful content when evaluating linear regression parameters, that is, to quantify the degree of redundancy, is considered. The article provides a comparative analysis of the PETRES Red test with the most used multicollinearity detection procedures among the regressors, which are implemented in the mctest R package: Farrar-Glober test, VIF (dispersion inflation factor) and others. Comparative analysis was performed on data from the author's earlier work on 186 enterprises related to the crude oil production activity for 2016. It was concluded that the use of an integrated approach to testing multicollinearity with the help of the r-package mctest, which calculates general and individual diagnostic multicollinearity tests.

Keywords: multicollinearity, information redundancy, red indicator, R-package mctest.

INTRODUCTION

The accumulation of large amounts of data is accompanied by the problem of data redundancy, that is, the accumulation of data that do not convey new or noteworthy information in terms of the problem being solved. Therefore, the task of analyzing the information content of metric data is important in econometric modeling. Multicollinearity, which occurs quite often in the construction of linear regression models, can be interpreted as a type of redundancy.

Estimates of the coefficients of the regression equation $\hat{\beta}$ we get by solving the system of normal equations: $X^T X \cdot \hat{\beta} = X^T Y$. The system of linear equations has a unique solution if the matrix of the system ($X^T X$) has full rank, that is, if all columns of the matrix ($X^T X$) linearly independent. If the columns of the matrix are collinear, then they speak of a strict (full) multicollinearity between the columns. The case of full multicollinearity is extremely rare in the practice of econometric modeling. Strict multicollinearity, as a rule, arises due to errors in the specification of the model, it is rather simply diagnosed and corrected.

If there is a close correlation between the columns of the matrix, then they say that there is a partial (lax) multicollinearity. It is this type of multicollinearity that is much more difficult to detect because it is not an error in specification or modeling, in fact it is a manifestation of data redundancy [1].

Multicollinearity and identification of its causes are often a serious problem in economic research, because, on the one hand, negative consequences of multicollinearity do not always occur, and on the other hand, multicollinearity can be caused not only by one variable, but also by a group of variables [2]. To avoid incorrect conclusions from the model about the effect of regressors on the endogenous variable, the existence of multicollinearity should always be checked when analyzing a data set as an initial step in multiple regression analysis [3].

MULTICOLLINEARITY TESTING

The red indicator ¾ PETRES Red, proposed by P. Kovacs [4] is proposed to be used to measure the proportion of useful content in relation to the estimate $\hat{\beta} = (X^T X)^{-1} X^T Y$, i.e. to quantify the degree of redundancy and, consequently, multicollinearity.

The red indicator is based on the following assumption. If the database used as the source of the explanatory variables is redundant with respect to the $\hat{\beta}_j$ estimate, that is, if the covariance of the data is significant, not all the data will have useful content. The smaller the proportion of data with useful content, the greater the redundancy. The greater the variance of eigenvalues, the greater the covariance of the explanatory variables. There are two extreme cases: all eigenvalues are equal to each other, or all eigenvalues except one are zero. The degree of dispersion can be determined quantitatively by the relative dispersion of the eigenvalues of the correlation matrix R of exogenous variables:

$$V_\lambda = \frac{\sigma_\lambda}{\bar{\lambda}} = \frac{\sqrt{\frac{\sum_{j=1}^{k}(\lambda_j - \bar{\lambda})^2}{k}}}{\sum_j^k \lambda_j} = \frac{\sqrt{\frac{\sum_{j=1}^{k}(\lambda_j - \bar{\lambda})^2}{k}}}{\frac{k}{k}} = \sqrt{\frac{\sum_{j=1}^{k}(\lambda_j - \bar{\lambda})^2}{k}} = \sigma_\lambda,$$

where k is the number of regressors, λ_j — the eigenvalues of the matrix.

To make the redundancy of different databases comparable, the above indicator should be normalized. Since the eigenvalues are non-negative, the normalization is performed with the value $\sqrt{k-1}$ because $0 \leq V_\lambda \leq \sqrt{k-1}$.

Red Indicator (red indicator) is defined by the formula:

$$\boldsymbol{Red} = \frac{\boldsymbol{V_\lambda}}{\sqrt{\boldsymbol{k-1}}}$$

If the value of the red indicator is zero, it indicates the absence of redundancy, and the value close to 1 indicates the maximum redundancy (multicollinearity).

The red indicator can be expressed without calculating the eigenvalues of the correlation matrix of independent variables, simply as the mean square of the correlation coefficients [5]

$$\boldsymbol{Red} = \sqrt{\frac{\sum_{i=1}^{k}\sum_{j=1}^{k} r_{i,j}^2}{k(k-1)}},\ j \neq i.$$

testing functions are included. The red indicator is included in the mctest package R [7]: to detect collinearity among regressors, in the **omcdiag** function, which implements the general diagnostic test of multicollinearity [6].

The **omcdiag** function implements several tests for checking the multicollinearity of the entire data array [7]:

- checking the equality to zero of the determinants of the correlation matrix;
- Farrar-Glober test (the first part, checking the presence of multicollinearity of the entire array of variables using the chi-square test);
- Red Indicator (red indicator);
- test ¾ Sum of Lambda Inverse (sum of inverse values of eigenvalues);
- Theil indicator;
- Condition Index (CI).

EMPIRICAL RESULTS

We test for redundancy data prepared in the SPARK system for enterprises belonging to the activity "Extraction of crude oil" to solve the problem of analyzing the impact of twenty-two indicators on the variable profit (loss) before taxation of a number of enterprises [9]. A sample was made of data representing the financial performance of 186 firms. The indicators for 2016 were used as regressors of the model - the average number of employees, return on assets, the cost of fixed production assets and equipment, the value of total assets, tangible assets, etc. There was a close correlation between many variables ($r_{i,j} > 0{,}9).r_{i,j} > 0{,}9)$.

As a result of the **omcdiag** function at given thresholds, only two Red Indicator and Theil's Method tests did not reveal the presence of multicollinearity.

In the **mctest** package, there is an imcdiag () function that includes seven

tests to test the effect of each regressor on multicollinearity. Among them are the most well-known and widely used *VIF test*, which is also included in the package **car** and *Farrar wi* (Farrar-Glober test, the second part is the identification of regressors leading to multicollinearity). According to the VIF test, only 8 of the 22 variables were not involved in multicollinearity, and according to the Farrar-Glober test, only two.

CONCLUSION

The results of applying the two functions of the **mctest** package indicate the presence of the strongest multicollinearity in the analyzed data and confirm the conclusions made from these data in [10] using the Belsley method.

In conclusion, it can be noted that combining various tests in the **mctest** package allows us to analyze the multicollinearity problem and the data redundancy problem associated with it, from different angles, changing the test suite, threshold values, and the form of results output.

REFERENCES

1. *Péter Kovács*. The elliptical model of multicollinearity and the Petres'Red indicator //Hetesi, E. – Kürtösi, Zs. (eds) 2011: The diversity of research at the Szeged Institute of Business Studies. JATEPress, Szeged, pp. 145-154

2. *L. O. Babeshko, M. G. Beach, I. V. Orlova* Econometrics and econometric modeling: textbook - Moscow: Vuzovskij uchebnik: INFRA-M, 2018. - 384 p.

3. *Orlova I. V., Filonova, E. S.* The Choice of exogenous factors in the regression model when multicollinearity data // Mezhdunarodnyj zhurnal prikladnyh i fundamental'nyh issledovanij. — 2015. — № 5—1. pp. 108—116.

4. *Péter Kovács*. Examination of Multicollinearity in Linear Regression Models Examination of PETRES' Red. Theses of PhD Dissertation Szeged 2008 http://docplayer.hu/3607097-Examintion-of-multicollinearity-in-linear-regression-models-examination-of-petres-red.html (request date 12.03.2018)

5. *Kovacs, P., Petres, T., and Toth, L. A.* A new measure of multicollinearity in linear regression// International Statistical Review / Revue Internationale de Statistique, 2005.73(3): pp. 405-412

6. *M. I. Ullah, M. Aslam*. Saima Altaf mctest: An R Package for Detection of Collinearity among Regressors // The R Journal (2016) volume 8:2, pages 495-505.

7. *Muhammad Imdad Ullah, Muhammad Aslam*. Multicollinearity Diagnostic Measures. Package 'mctest' https: //cran.r-prject.org/web/packages/mctest/mctest.pdf (request date 30.07.2018)

8. Project R for statistical calculations http://www.r-project.org/ (request date 12.03.2018)

9. Network edition « SPARK Information resource (System of Professional Analysis of Markets and Companies) » http://www.spark-interfax.ru (request date 13.12.2017).

10. *Orlova I. V.* Approach to solving the problem of multicollinearity in the analysis of the influence of factors on the resulting variable in regression models // Fundamental research - 2018. - № 3. pp. 58-63.