DOI 10.33278/SAE-2018.rus.247-250
УДК 330.4:519.237
JEL C23

АНАЛИЗ ИНФОРМАЦИОННОГО КОНТЕНТА МЕТРИЧЕСКИХ ДАННЫХ ПРИ ПОСТРОЕНИИ МОДЕЛЕЙ ЛИНЕЙНОЙ РЕГРЕССИИ

Ирина Владленовна Орлова (ORCID 0000-0001-5397-2450)

ФГОБУ ВО «Финансовый Университет при Правительстве Российской Федерации»

Аннотация. Данная статья посвящена анализу методов, направленных на решение проблемы мультиколлинеарности данных, возникающей по причине высокой информационной избыточности метрических данных. Принятие решений в современных условиях базируется на анализе огромных объемов данных, часто имеющих лишь небольшой информационный контент, а это означает, что высока информационная избыточность. В случае с моделью линейной регрессии мультиколлинеарность можно интерпретировать как тип избыточности. Рассматривается возможность применения теста PETRES Red — красного индикатора для измерения доли полезного контента при оценке параметров линейной регрессии, т.е. для количественной оценки степени избыточности. В статье приводится сравнительный анализ теста PETRES Red с наиболее используемыми процедурами обнаружения мультиколлинеарности среди регрессоров, которые реализованы в R-пакете mctest: тест Фаррара-Глоубера, VIF (фактор инфляции дисперсии) и другими. Сравнительный анализ выполнен на данных из более ранней работы автора по 186 предприятиям, относящихся к виду деятельности добыча сырой нефти за 2016 г. Сделан вывод о целесообразности применения комплексного подхода к тестированию мультиколлинеарности с помощью R-пакета mctest, который вычисляет общие и индивидуальные диагностические тесты мультиколлинеарности.

Ключевые слова: мультиколлинеарность, информационная избыточность, красный индикатор, R-пакет mctest.

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 — 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.

Введение

Накоплению больших объемов данных сопутствует проблема избыточности данных, т.е. накопление данных, которые не передают новую или заслуживающую внимания информацию с точки зрения решаемой проблемы. Поэтому задача анализа информационного содержания метрических данных является важной при эконометрическом моделировании. Мультиколлинеарность, возникающая достаточно часто при построении моделей линейной регрессии, можно интерпретировать как тип избыточности.

Оценки коэффициентов уравнения регрессии $\hat{\beta}$ мы получаем, решая систему

нормальных уравнений: $X^T X \cdot \hat{\beta} = X^T Y$. Система линейных уравнений имеет единственное решение, если матрица системы ($X^T X$) имеет полный ранг, то есть если все столбцы матрицы ($X^T X$) линейно независимы. Для этого необходимо, чтобы столбцы матрицы исходных данных X были линейно независимы. Если столбцы матрицы X коллинеарны, то говорят о *строгой (полной) мультиколлинеарности* между столбцами. Случай полной мультиколлинеарности встречается крайне редко в практике эконометрического моделирования. Строгая мультиколлинеарность, как правило, возникает из-за ошибок спецификации модели, достаточно просто диагностируется и корректируется.

Если между столбцами матрицы существует тесная корреляционная зависимость, то говорят о наличии частичной (нестрогой) мультиколлинеарности. Именно этот тип мультиколлинеарности обнаружить значительно сложнее, поскольку она не является ошибкой спецификации или моделирования, на самом деле это проявление избыточности данных [1].

Мультиколлинеарность и выявление ее причины часто представляют собой серьезную проблему в экономических исследованиях, поскольку, с одной стороны, не всегда возникают отрицательные последствия мультиколлинеарности, а с другой стороны, мультиколлинеарность может быть вызвана не только одной переменной, но и группой переменных [2]. Чтобы избежать неверных выводов из модели о влиянии регрессоров на эндогенную переменную, существование мультиколлинеарности должно всегда проверяться при анализе набора данных в качестве начального шага при множественном регрессионном анализе [3].

Тестирование мультиколлинеарности

Предложенный П. Ковачем [4] красный индикатор — PETRES Red — предлагается использовать для измерения доли полезного контента в отношении оценки $\hat{\beta} = (X^T X)^{-1} X^T Y$, т.е. для количественной оценки степени избыточности и, следовательно, мультиколлинеарности.

В основе красного индикатора лежит следующее предположение. Если база данных, используемая в качестве источника независимых переменных, является избыточной в отношении оценки $\hat{\beta}_j$, то есть если ковариация данных значительна, не все данные будут иметь полезное содержание. Чем меньше доля данных с полезным содержимым, тем больше будет избыточность. Чем больше дисперсия собственных значений, тем больше ковариация независимых переменных. Есть два крайних случая: все собственные значения равны друг другу или все собственные значения, за исключением одного, равны нулю. Степень разброса может быть определена количественно с помощью относительной дисперсии собственных значений корреляционной матрицы R экзогенных переменных:

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

где k — количество регрессоров, λ_j — собственные числа матрицы R.

Чтобы сделать избыточность различных баз данных сопоставимыми, приведенный выше индикатор должен быть нормирован. Так как собственные значения неотрицательны, нормировка выполняется со значением $\sqrt{k-1}$, поскольку $0 \leq V_\lambda \leq \sqrt{k-1}$.

Red Indicator (красный индикатор)

определяется по формуле: $\boldsymbol{Red}=\frac{V_\lambda}{\sqrt{k-1}}$.

Если значение красного индикатора равно нулю, то это свидетельствует об отсутствии избыточности, а значения, близкие к 1, указывают максимальную избыточность (мультиколлинеарность).

Красный индикатор может быть выражен без вычисления собственных значений корреляционной матрицы независимых переменных просто как среднее квадратичное коэффициентов корреляции [5].

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

В среде R реализовано несколько пакетов, в которые включены функции тестирования мультиколлинеарности. Красный индикатор включен в **mctest** пакета R [7]: для обнаружения коллинеарности среди регрессоров в функцию **omcdiag(),** которая реализует общую диагностическую проверку мультиколлинеарности [6].

В функции **omcdiag()** реализовано несколько тестов проверки мультиколлинеарности всего массива данных [7]:

- проверка равенства нулю определителя корреляционной матрицы;
- тест Фаррара-Глоубера (первая часть, проверка наличия мультиколлинеарности всего массива переменных по критерию «хи-квадрат»);
- Red Indicator (красный индикатор);
- тест — Sum of Lambda Inverse (сумма обратных значений собственных чисел);
- Theil индикатор;
- Индекс обусловленности (CI).

Протестируем на избыточность данные, подготовленные в системе СПАРК по предприятиям, относящимся к виду деятельности «Добыча сырой нефти» для решения задачи анализа влияния двадцати двух показателей на переменную прибыль (убыток) до налогообложения ряда предприятий [9]. Была сделана выборка данных, представляющих финансовые показатели 186 фирм. В качестве регрессоров модели использованы показатели за 2016 год — среднесписочная численность работников, рентабельность активов, стоимость основных производственных средств и оборудования, стоимость совокупных активов, материальные активы и др. Между многими переменными была выявлена тесная корреляционная связь ($r_{i,j} > 0{,}9$).

В результате работы функции **omcdiag** при заданных пороговых значениях только два теста Red Indicator и Theil's Method не выявили наличие мультиколлинеарности.

В пакете **mctest** есть функция **imcdiag()**, включающая семь тестов проверки влияния каждого регрессора на мультиколлинеарность. Среди них наиболее известные и широко применяемые — VIF тест, который входит также в пакет **car** и Farrar wi (тест Фаррара-Глоубера, вторая часть — выявление регрессоров, приводящих к мультиколлинеарности). По тесту VIF только 8 из 22 переменных не оказались вовлеченными в мультиколлинеарность, а по тесту Фаррара-Глоубера всего две.

Выводы

Результаты применения двух функций пакета **mctest** свидетельствуют о наличии сильнейшей мультиколлинеарности в анализируемых данных и подтверждают выводы, сделанные по этим данным в работе [10] с помощью метода Белсли.

В заключение можно отметить, что объединение в пакете **mctest** различных тестов позволяет проанализировать проблему мультиколлинеарности и связанную с ней проблему избыточности данных с разных сторон, изменяя набор тестов, пороговые значения, форму выдачи результатов.

Список источников

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. Эконометрика и эконометрическое моделирование: учебник /Л.О. Бабешко, М.Г. Бич, И.В. Орлова. – М.: Вузовский учебник: ИНФРА-М, 2018. – 384 с.
3. Орлова И.В., Филонова Е.С. Выбор экзогенных факторов в модель регрессии при мультиколлинеарности данных // Международный журнал прикладных и фундаментальных исследований. — 2015. — № 5—1. С. 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-Examination-of-multicollinearity-in-linear-regression-models-examination-of-petres-red.html (дата обращения 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-project.org/web/packages/mctest/mctest.pdf (дата обращения 30.07.2018)
8. Проект R для статистических вычислений http://www.r-project.org/ (дата обращения 12.03.2018)
9. СПАРК -Режим доступа: http://www.spark-interfax.ru (дата обращения 13.12.2017).
10. Орлова И.В. Подход к решению проблемы мультиколлинеарности при анализе влияния факторов на результирующую переменную в моделях регрессии // Фундаментальные исследования — 2018. — № 3. С. 58—63.