УДК 330:519.237.5

АНАЛИЗ ДИАГНОСТИЧЕСКИХ ИНДИКАТОРОВ ОБЩЕЙ И ИНДИВИДУАЛЬНОЙ КОЛЛИНЕАРНОСТИ РЕГРЕССОРОВ

Орлова И.В.

Финансовый университет при Правительстве Российской Федерации» (Финансовый университет), Москва, e-mail: ivorlova@fa.ru

Статья посвящена анализу индикаторов общей и индивидуальной диагностики коллинеарности, направленных на решение проблемы мультиколлинеарности данных, возникающей по причине высокой информационной избыточности метрических данных. Индикаторы общей диагностики помогают получить представление о существовании мультиколлинеарности, но они не указывают, какой регрессор может быть причиной коллинеарности, в то время как индикаторы индивидуальной диагностики указывают на регрессоры, вызывающие коллинеарность. Исследования выполнялись в программной среде R, где для обнаружения коллинеарности среди регрессоров используют пакет mctest, в нем есть две функции: omcdiag и imcdiag, которые реализуют диагностику общей и индивидуальной проверки мультиколлинеарности. Рассмотрены две модификации фактора инфляции дисперсии – CVIF и MCVIF. Показано, что эти индикаторы имеют ограниченную сферу применения и могут показывать не интерпретируемые результаты. Проанализировано использование индикаторов общей и индивидуальной диагностики для тестирования мультиколлинеарности в задаче моделирования цены на бензин в Российской Федерации с января 2016 по сентябрь 2018. Решена задача устранения мультиколлинеарности. Получено уравнение регрессии, позволившее выявить наиболее важные факторы, оказывающие влияние на формирование цены на бензин. Комплексное применение диагностических индикаторов общей и индивидуальной коллинеарности, реализованных в пакете mctest в среде R, упрощает решение задачи выявления и устранения мультиколлинеарности. Хотя не все индикаторы, включенные в пакет, на данный момент одинаково полезны. В дальнейшем необходима модификация индикатора CVIF, разработка более обоснованных критических значений для красного индикатора.

Ключевые слова: мультиколлинеарность, многофакторная регрессионная модель, фактор инфляции дисперсии, избыточность данных

ANALYSIS OF THE DIAGNOSTIC INDICATORS GENERAL AND INDIVIDUAL COLLINEARITY OF REGRESSORS

Orlova I.V.

Financial University under the Government of the Russian Federation, Moscow, e-mail: ivorlova@fa.ru

The article is devoted to the analysis of indicators of general and individual diagnostics of collinearity, aimed at solving the problem of multicollinearity of data arising due to high information redundancy of metric data. General diagnostics indicators help to get an idea of the existence of multicollinearity, but they do not indicate which regressor can be the cause of collinearity, while the indicators of individual diagnostics indicate regressors that cause collinearity. The studies were performed in the R software environment, where the mctest package is used to detect collinearity among regressors; it has two functions: omcdiag and imcdiag, which implement the diagnostics of general and individual multicollinearity checks. Two modifications of the dispersion inflation factor are considered – CVIF and MCVIF. It is shown that these indicators have a limited scope and can show non-interpretable results. Analyzed the use of indicators of general and individual diagnostics for testing multicollinearity in the task of modeling the price of gasoline in the Russian Federation from January 2016 to September 2018. The problem of eliminating multicollinearity has been solved. A regression equation was obtained, which made it possible to identify the most important factors influencing the price of gasoline. Comprehensive application of diagnostic indicators of general and individual collinearity, implemented in the mctest package in the R environment, simplifies the solution of the problem of identifying and eliminating multicollinearity. Although not all indicators included in the package are equally useful at the moment. In the future, it is necessary to modify the CVIF indicator, to develop more reasonable critical values for the red indicator.

Keywords: multicollinearity, multivariate regression model, variance inflation factor, data redundancy

Мультиколлинеарность – это проблема, с которой можно столкнуться при построении регрессионных моделей. Распознавание мультиколлинеарности и выявление ее причин часто представляют серьезную задачу в эмпирических исследованиях, поскольку, с одной стороны, негативные последствия мультиколлинеарности не всегда происходят, а, с другой стороны, мультиколлинеарность может быть вызвана не только одной переменной, но и группой переменных. Задача усложняется, если есть сильная тенденция в объясняющих переменных или если доступный объем информации слишком мал для изучения влияния объясняющих переменных на зависимую переменную. Мультиколлинеарность увеличивает дисперсию оценок коэффициентов и делает оценки очень чувствительными к незначительным изменениям в модели. В результате оценки коэффициентов нестабильны и трудно поддаются интерпретации.

Цель исследования: анализ индикаторов общей и индивидуальной диагностики

коллинеарности, направленных на решение проблемы мультиколлинеарности данных, возникающей по причине высокой информационной избыточности метрических данных. Для обнаружения мультиколлинеарности среди регрессоров используются различные диагностические индикаторы. Во многих статистических программах присутствуют несколько процедур для оценки мультиколлинеарности.

Большинство индикаторов показывают, насколько исследуемые данные не идеальны, то есть в какой степени они отклоняются от «идеального случая», когда каждая объясняющая переменная линейно независима от других. Для некоторых индикаторов нет определенной границы для указания вредной степени отклонения. Интерпретация индикаторов мультиколлинеарности часто весьма субъективна. Результат методов, применяемых для уменьшения негативных эффектов мультиколлинеарности, напрямую связан со степенью распознавания мультиколлинеарности. Хотя использование большинства этих методов уменьшает или может уменьшить уровень негативных последствий мультиколлинеарности, это может сопровождаться другими отрицательными последствиями – например, вследствие значительной потери информации или неправильной интерпретируемости результатов.

Материалы и методы исследования

Для проведения исследования представляется полезным разбиение индикаторов мультиколлинеарности на две группы: реализующих общую диагностику всего массива переменных и индивидуальную диагностику [1]. Индикаторы общей диагностики помогают получить представление о существовании мультиколлинеарности, но они не указывают, какой регрессор может быть причиной коллинеарности, в то время как индикаторы индивидуальной диагностики указывают на регрессоры, вызывающие коллинеарность.

В R для обнаружения коллинеарности среди регрессоров используют пакет mctest, в нем есть две функции: omcdiag() и imcdiag(), которые реализуют диагностику общей и индивидуальной проверки мультиколлинеарности [2].

Рассмотрим использование индикаторов функции imcdiag() направленных на выявление влияния каждого регрессора на мультиколлинеарность.

Фактор инфляции дисперсии *VIF* (Variance Inflation Factor) и тест (TOL) являются широко используемыми мерами степени мультиколлинеарности *j*-й независимой переменной с другими независимыми переменными в регрессионной модели

$$VIF_j = VIF\left(\hat{\beta}_j, \hat{\beta}_{j0}\right) = \frac{var(\hat{\beta}_j)}{var(\hat{\beta}_{j0})} = \frac{1}{(1-R_j^2)},$$

где $\hat{\beta}_j$ – оценка коэффициента регрессии $\beta_j, \hat{\beta}_{j0}$ – соответствующая оценка по модели с *j*-м регрессором, ортогональным другим независимым переменным, R_j^2 – коэффициент детерминации регрессии для каждого *j*-го регрессора по всем остальным регрессорам.

$$TOL_j = \frac{1}{VIF_j} = 1 - R_j^2.$$

Коэффициенты дисперсии инфляции варьируются от 1 и выше. При ортогональности вектора значений признака остальным коэффициент дисперсии инфляции будет равен единице. То, насколько большим должен быть *VIF*, прежде чем он вызовет проблемы, является предметом обсуждения. Известно, что чем больше увеличивается *VIF*, тем менее достоверными будут результаты регрессии. В целом, если $VIF_j > 10$, то *j*-й регрессор может привести к мультиколлинеарности. Некоторые авторы [3] предлагают более консервативный уровень 5 или даже 2,5.

Иногда высокий *VIF* вообще не является поводом для беспокойства, например при использовании фиктивных переменных, представляющих номинальные переменные с тремя или более категориями.

Курто и Пинто [4] указали ситуации, когда традиционный *VIF* будет переоценивать реальное влияние корреляции между регрессорами на дисперсию и предложили индикатор, известный как исправленный *VIF* (*CVIF*):

$$CVIF_j = VIF_j \cdot \frac{1-R^2}{1-R_0^2} = VIF_j \cdot C,$$

где $R_0^2 = R_{yx_1}^2 + R_{yx_2}^2 + ... + R_{yx_k}^2$.

Однако Курто и Пинто не рассматривали ситуации, когда R_0^2 может быть больше 1. Следствием этого будет то, что $CVIF_j$ может принимать не интерпретируемые отрицательные значения. В работе [5] предложена модификация $CVIF_j$:

$$MCVIF_j = VIF_j \cdot |C|, j = 2,\ldots, p\ .$$

Несмотря на это изменение, может возникнуть другая проблема, когда C <1, $CVIF_j$ будет меньше 1, что не соответствует определению VIF_j, так как это будет означать, что $var\left(\hat{\beta}_j\right) < var(\hat{\beta}_{j0})$, а это невозможно

поскольку дисперсия ортогонального фактора должна быть наименьшей. Встречаются ситуации, когда применение $MCVIF_j$ и $CVIF_j$ могут дать более точное представление о мультиколлинеарности, но использование классического VIF_j возможно во всех случаях, поэтому можно пренебречь тем, что иногда он будет переоценивать реальное влияние корреляции между регрессорами на дисперсию.

Результаты исследования и их обсуждение

Проанализируем использование рассмотренных индикаторов для тестирования мультиколлинеарности в задаче моделирования цены на бензин в Российской Федерации с января 2016 по сентябрь 2018. Зависимая переменная Y – цена на бензин в РФ, (USD/lit) [6]; регрессоры: X1 – курс рубля к евро; X2 – курс доллара к евро [7]; X3 – цена на нефть Brent, (USD/barrel) [8]; X4 – цена бензина в Европе, (USD/lit) [9]; X5 – цена на бензин в США, (USD/lit); X6 – Мировое производство сырой нефти и жидкого топлива (миллионов баррелей в день) [10]. Данные получены из открытых источников.

Воспользовавшись функцией imcdiag() пакета mctest(), получим результаты индивидуальной диагностики (рис. 1).

Фактор инфляции дисперсии VIF больше 10 у X2, X3, X5, у X4 больше 5. Именно эти факторы приводят к мультиколлинеарности.

Значения индикатора *CVIF* для всех регрессоров отрицательные и меньше 1. Размещение индикатора *CVIF* в функции imcdiag() пакета mctest() не является обоснованным. Сообщение, что коэффициенты при факторах X3, X4, X5, X6 могут быть незначимы из-за мультиколлинеарности, подтверждается протоколом регрессионного анализа (рис. 2).

Построенное уравнение регрессии является значимым (критерий Фишера равен 66,89, p-value: 1.464e-14), коэффициент детерминации высокий 0,939, а коэффициенты при факторах X3, X4, X5, X6 незначимы (p-value больше 0,05). Такая ситуация характерна при частичной, или нестрогой, мультиколлинеарности в данных. Этот тип мультиколлинеарности обнаружить значительно сложнее, поскольку она не является ошибкой спецификации или моделирования, на самом деле это проявление избыточности данных.

```
> imcdiag(x = XX, y = Y, method = NULL, corr = FALSE, vif = 10, tol = 0.1,conf
= 0.95, cvif = 10, leamer = 0.1, all = FALSE)

Call:
imcdiag(x = XX, y = Y, method = NULL, corr = FALSE, vif = 10,
    tol = 0.1, conf = 0.95, cvif = 10, leamer = 0.1, all = FALSE)


All Individual Multicollinearity Diagnostics Result

       VIF    TOL      Wi       Fi Leamer    CVIF  Klein
X1  1.4934 0.6696  2.6643   3.4537 0.8183 -0.0659     0
X2 10.1981 0.0981 49.6698  64.3868 0.3131 -0.4499     0
X3 13.8558 0.0722 69.4213  89.9905 0.2686 -0.6113     0
X4  6.6975 0.1493 30.7664  39.8824 0.3864 -0.2955     0
X5 15.5068 0.0645 78.3365 101.5473 0.2539 -0.6842     0
X6  3.0007 0.3333 10.8040  14.0052 0.5773 -0.1324     0

All Individual Multicollinearity Diagnostics in 0 or 1

   VIF TOL Wi Fi Leamer CVIF Klein
X1   0   0  0  1      0    0     0
X2   1   1  1  1      0    0     0
X3   1   1  1  1      0    0     0
X4   0   0  1  1      0    0     0
X5   1   1  1  1      0    0     0
X6   0   0  1  1      0    0     0

1 --> COLLINEARITY is detected by the test
0 --> COLLINEARITY is not detected by the test

X3, X4, X5, X6, coefficient(s) are non-significant may be due to multicollinear
ity

R-square of y on all x: 0.9392

* use method argument to check which regressors may be the reason of collineari
ty
```

Рис. 1. Диагностика влияния каждого регрессора на мультиколлинеарность

```
fm<-lm(data=tab1,Y~X1+X2+X3+X4+X5+X6)
> summary(fm)

Call:
lm(formula = Y ~ X1 + X2 + X3 + X4 + X5 + X6, data = tab1)

Residuals:
      Min        1Q    Median        3Q       Max
-0.049109 -0.008286  0.005038  0.009108  0.028815

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.8374932  0.4433438   1.889   0.0701 .
X1          -0.0067172  0.0007017  -9.573 5.22e-10 ***
X2          -0.5575364  0.2586600  -2.155   0.0406 *
X3           0.0017382  0.0010055   1.729   0.0957 .
X4           0.0355540  0.1286239   0.276   0.7844
X5           0.1190930  0.1757627   0.678   0.5040
X6           0.0052044  0.0042419   1.227   0.2308
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.01838 on 26 degrees of freedom
Multiple R-squared:  0.9392,   Adjusted R-squared:  0.9251
F-statistic: 66.89 on 6 and 26 DF,  p-value: 1.464e-14
```

Рис. 2. Протокол регрессионного анализа

```
> omcdiag (x = XX, y = Y)

Call:
omcdiag(x = XX, y = Y)


Overall Multicollinearity Diagnostics

                       MC Results detection
Determinant |X'X|:         0.0015         1
Farrar Chi-Square:       190.2018         1
Red Indicator:             0.6061         1
Sum of Lambda Inverse:    49.3921         1
Theil's Method:           -0.2363         0
Condition Number:         10.1545         0

1 --> COLLINEARITY is detected by the test
0 --> COLLINEARITY is not detected by the test
```

Рис. 3. Индикаторы общей диагностики мультиколлинеарности

Значения *VIF* указывают, в какой степени каждый из факторов приводит к мультиколлинеарности:

VIF (X2 – курс доллара к евро) = 10,2;

VIF (X3 – цена на нефть Brent, (USD/barrel) = 13,86;

VIF (X4 – цена бензина в Европе, (USD/lit) = 6,7;

VIF (X5 – цена на бензин в США, (USD/lit) = 15,51.

Протестируем на избыточность анализируемые данные с помощью функции omcdiag() пакета mctest() (рис. 3).

В функции omcdiag() реализовано несколько тестов проверки мультиколлинеарности всего массива данных [11]: 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) и другие. В данной ситуации представляет интерес значение красного индикатора, свидетельствующее об избыточности анализируемых данных. Действительно, X3 – цена на нефть Brent, (USD/barrel) и X5 – цена на бензин в США, (USD/lit) дублируют, в какой-то степени, друг друга. В решаемой задаче исключение отдельных факторов из модели вполне обосновано.

Применяя пошаговую процедуру исключения факторов, получили трехфакторную модель регрессии (рис. 4):

$$\hat{Y}_i = 1{,}384 - 0{,}007X_1 - 0{,}541X_2 + 0{,}003\,X_3.$$

Анализ теста на мультиколлинеарность последней модели показал ее отсутствие.

```
> fm1<-lm(data=tab1,Y~X1+X2+X3)
>  summary(fm1)
Call:
lm(formula = Y ~ X1 + X2 + X3, data = tab1)
Residuals:
     Min       1Q   Median       3Q      Max
-0.051355 -0.008016  0.000111  0.011343  0.025496
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.3838240  0.1277549  10.832 1.04e-11 ***
X1          -0.0066567  0.0006147 -10.829 1.05e-11 ***
X2          -0.5408348  0.1052332  -5.139 1.72e-05 ***
X3           0.0029634  0.0003501   8.465 2.50e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.01824 on 29 degrees of freedom
Multiple R-squared:  0.9331,        Adjusted R-squared:  0.9262
F-statistic: 134.8 on 3 and 29 DF,  p-value: < 2.2e-16

> vif(fm1)
    X1     X2     X3
1.163 1.712 1.704
```

Рис. 4. Результаты последнего протокола регрессионного анализа

Значения факторов инфляции от 1,16 до 1,72 (рис. 4):

VIF (X1 – курс рубля к евро) = 1,163;

VIF (X2 – курс доллара к евро) = 1,712;

VIF (X3 – цена на нефть Brent, (USD/barrel) = 1,704.

После устранения мультиколлинеарности можно использовать полученное уравнение регрессии для ранжирования факторов по степени их влияния на зависимую переменную с помощью дельта-коэффициентов $\Delta(j)$ [12]:

$$\Delta_j = r_{y,x_j} \cdot \frac{\hat{\beta}_j}{R^2},$$

где r_{y,x_j} – коэффициент парной корреляции между фактором X_j и зависимой переменной, $\hat{\beta}_j$ – коэффициент при факторе X_j уравнения регрессии в стандартизованном виде, R^2 – коэффициент детерминации. Результаты представлены в таблице.

Дельта коэффициенты

Δ_1	Δ_2	Δ_3
0,379	0,184	0,477

Из этой таблицы можно сделать вывод, что наибольшее влияние на цену бензина в рассматриваемый период оказывает фактор X3 – цена на нефть Brent, затем X1 – курс рубля к евро и X2 – курс доллара к евро.

Заключение

Подводя итог, можно отметить, что комплексное применение диагностических индикаторов общей и индивидуальной коллинеарности, реализованных в пакете mctest() в среде *R*, упрощает решение задачи выявления и устранения мультиколлинеарности. Хотя не все индикаторы, включенные в пакет, на данный момент одинаково полезны. Так, требуется модификация индикатора *CVIF*, разработка более обоснованных критических значений для красного индикатора. Но в целом инструменты пакета mctest() вполне пригодны для использования.

Список литературы

1. Ullah M.I., Aslam M., Saima Altaf mctest: An R Package for Detection of Collinearity among Regressors. The R Journal. 2016. vol. 8:2. P. 495–505. DOI: 10.32614/RJ-2016-062.

2. Орлова И.В. Анализ инструментов языка R для решения проблемы мультиколлинеарности данных // Современные наукоемкие технологии. 2018. № 6. С. 129–137.

3. O'Brien R.M. A Caution Regarding Rules of Thumb for Variance Inflation Factors. Quality & Quantity. 2007. № 41. P. 673–690.

4. Curto J.D., Pinto J.C. The corrected vif (cvif). J. Appl. Stat. 2011. № 38 (7). P. 1499–1507. DOI: 10.1080/02664763.2010.505956.

5. Salmerón, Román & Pérez, Jose & Garcia, Catalina & López Martín, María. A note about the corrected VIF. Statistical Papers. 2015. 58. DOI: 10.1007/s00362-015-0732-9.

6. Розничные цены на бензин АИ-92 [Электронный ресурс]. URL: https://news.yandex.ru/quotes/1/20001.html (дата обращения: 25.12.2018).

7. Таблица валют: USD – Доллар США [Электронный ресурс]. URL: https://www.xe.com/currencytables/?from=USD&date=2018-09-01(дата обращения: 25.12.2018).

8. Short-Term Energy Outlook U.S. Energy Information Administration (EIA) [Электронный ресурс]. URL: https://www.eia.gov/outlooks/steo/data.php?type=figures (дата обращения: 25.12.2018).

9. Еженедельный бюллетень о нефти [Электронный ресурс]. URL: https://ec.europa.eu/energy/en/data-analysis/weekly-oil-bulletin (дата обращения: 25.12.2018).

10. Сырая нефть (petroleum) месячные цены [Электронный ресурс]. URL: https://www.indexmundi.com/commodities/?commodity=crude-oil&months=60 (дата обращения: 25.12.2018).

11. Орлова И.В. Анализ информационного контента метрических данных при построении моделей линейной регрессии // Системный анализ в экономике 2018: сборник трудов V Международной научно-практической конференции / Под общ. ред. Г.Б. Клейнера, С.Е. Щепетовой. М.: Прометей, 2018. С. 247–250.

12. 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.