УДК 004.43:519.6

АНАЛИЗ ИНСТРУМЕНТОВ ЯЗЫКА R ДЛЯ РЕШЕНИЯ ПРОБЛЕМЫ МУЛЬТИКОЛЛИНЕАРНОСТИ ДАННЫХ

Орлова И.В.

*Финансовый университет при Правительстве Российской Федерации
(Финансовый университет), Москва, e-mail: ivorlova@gmail.com*

Данная статья посвящена анализу инструментов языка программирования R, направленных в первую очередь на решение проблемы выявления мультиколлинеарности данных. Для моделей линейной регрессии характерна проблема мультиколлинеарности, когда два или более регрессора сильно коррелированы. Мультиколлинеарность не только приводит к неустойчивым оценкам регрессионных коэффициентов, но и вызывает некоторые серьезные проблемы в валидации и интерпретации модели. В статье приводится анализ наиболее используемых процедур обнаружения мультиколлинеарности среди регрессоров, которые реализованы в среде R: анализ корреляционной матрицы, VIF (фактор инфляции дисперсии), собственные значения корреляционной матрицы, графические процедуры визуализации мультиколлинеарности. Сравнительный анализ выполнен на смоделированных данных из более ранней работы автора. Большое внимание уделено недавно разработанному R-пакету *mctest*, который вычисляет популярные и широко используемые диагностические тесты мультиколлинеарности. Пакет также указывает, какие регрессоры могут быть причиной коллинеарности среди регрессоров. В пакете реализовано несколько функций: *omcdiag*, *imcdiag*, *eigprop* и *mc.plot* для графического представления значений VIF и собственных значений. Объединение в одном пакете различных диагностических мер общей и индивидуальной диагностики мультиколлинеарности, простой и удобный интерфейс позволяет сделать вывод, что пакет *mctest* является полезным дополнением к «Comprehensive R Archive Network».

Ключевые слова: мультиколлинеарность, язык программирования R, многофакторная регрессионная модель, VIF (фактор инфляции дисперсии), собственные значения корреляционной матрицы

ANALYSIS TOOLS LANGUAGE OF R TO SOLVE THE PROBLEM OF MULTICOLLINEARITY OF DATA

Orlova I.V.

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

This article is devoted to the analysis of R programming language tools aimed primarily at solving the problem of multicollinearity of data. Linear regression models are characterized by a multicollinearity problem where two or more regressors are strongly correlated. Multicollinearity not only leads to unstable estimates of regression coefficients, but also causes some serious problems in the validation and interpretation of the model. The article presents the analysis of the most used procedures for detection of multicollinearity among regressors, which are implemented in the environment R: analysis of the correlation matrix, VIF (variance inflation factor), eigenvalues of the correlation matrix, graphical procedures for visualization of multicollinearity. The comparative analysis is made on the simulated data from the earlier work of the author. Much attention is paid to the newly developed R-package *mctest*, which calculates the popular and widely used diagnostic tests for multicollinearity. The package also indicates which regressors can be the cause of collinearity among the regressors. The package includes several features: *omcdiag*, *imcdiag*, *eigprop*, and *mc.plot* a graphical representation of the values of VIF and eigenvalues. Combining in one package various diagnostic measures of General and individual diagnostics of multicollinearity, a simple and convenient interface allows you to conclude that the *mctest* package is a useful addition to the «Comprehensive R Archive Network».

Keywords: multicollinearity, programming language R, multi-factor regression model, VIF (variance inflation factor), eigenvalues of the correlation matrix

В современных условиях количество данных, на основании которых принимаются важные решения, непрерывно растет, однако увеличение количества данных не сопровождается соответствующим увеличением объема количества полезной информации. Огромные объемы данных часто имеют лишь небольшой информационный контент, а это означает, что высокая информационная избыточность. В случае с моделью линейной регрессии мультиколлинеарность можно интерпретировать как тип избыточности. К мультиколлинеарности приводит нарушение одной из пред-

посылок метода наименьших квадратов (МНК), а именно предпосылки на матрицу регрессоров:

$$\text{rank}(X) = \text{rank}(X^T X) = k,$$

где k – число столбцов матрицы регрессоров. Негативные последствия, к которым приводит мультиколлинеарность, хорошо известны [1, 2]. Мультиколлинеарность не только приводит к неустойчивым оценкам регрессионных коэффициентов, но и вызывает некоторые серьезные проблемы в валидации и интерпретации модели.

Существуют различные методы тестирования уровня мультиколлинеарности и ее устранения [1, 3]. Успех методов, используемых для уменьшения или устранения отрицательных эффектов мультиколлинеарности, во многом зависит от степени распознавания мультиколлинеарности. Выбор метода основывается на содержательной постановке задачи и на возможностях используемых компьютерных программ. В задачах, где целью моделирования является выявление раздельного влияния факторов на зависимую переменную, устранение мультиколлинеарности выполняют методами, в основе которых лежит сокращение числа регрессоров. В задачах прогнозирования устранение мультиколлинеарности возможно также за счет их преобразования (компонентный анализ, ридж регрессия).

Материалы и методы исследования

Свободно распространяемая статистическая среда R [4] объединяет язык программирования высокого уровня и библиотеки программных модулей (пакетов) для обработки данных. Постоянно происходит увеличение количества библиотек. Это привело к возникновению распределенной системы хранения и распространения пакетов R, то есть «CRAN» (от «Comprehensive R Archive Network») [5]. Для обнаружения мультиколлинеарности среди регрессоров R-пакеты предоставляют несколько диагностических мер для оценки мультиколлинеарности. В этой работе будет представлен анализ инструментов языка программирования R, направленных на решение проблемы выявления мультиколлинеарности данных. Для сравнительного анализа используются смоделированные данные. Описание данных [1]: «Объем продаж – это зависимая переменная Y (тыс. руб.). В качестве независимых (объясняющих) переменных выбраны: время – X1 (мес.), расходы на рекламу X2 (тыс. руб.), цена товара X3 (руб.), средняя цена товара у конкурентов X4 (руб.), индекс потребительских расходов X5 (%), количество наблюдений $n = 16$, количество факторов $k = 5$ ».

Результаты исследования и их обсуждение

Выявление мультиколлинеарности с помощью анализа матрицы коэффициентов парной корреляции

Наличие коэффициентов парной корреляции больших $|0,8|$ может свидетельствовать о мультиколлинеарности данных. В среде R построение матрицы коэффициентов парной корреляции выполняется при помощи функции `cor(x)`, применяя которую получаем матрицу коэффициентов парной корреляции (табл. 1). Анализ матрицы коэффициентов парной корреляции показал высокое значение коэффициента корреляции между факторами X1 и X5 ($r_{X_1X_5} = 0,960$), что указывает на наличие мультиколлинеарности.

Таблица 1
Корреляционная матрица

	X1	X2	X3	X4	X5
X1	1,000	0,106	0,174	–0,051	0,960
X2	0,106	1,000	–0,003	0,204	0,273
X3	0,174	–0,003	1,000	0,698	0,235
X4	–0,051	0,204	0,698	1,000	0,031
X5	0,960	0,273	0,235	0,031	1,000

При большом числе регрессоров визуальное исследование корреляционной матрицы становится сложным. В среде R есть возможность облегчить этот процесс с использованием графических функций `plotcorr` из пакета `ellipse` и `corrplot` из одноименного пакета. В пакете `ellipse` значения коэффициентов корреляции изображены в виде эллипсов. Чем ближе значение коэффициента корреляции к $|1|$ – тем более вытянутым становится эллипс (рис. 1). Наклон эллипса отражает знак.

Рис. 1. Изображение матрицы коэффициентов парной корреляции в пакете `ellipse`

Более информативным представляется применение пакета `corrplot`, при использовании которого ячейки отображаемой матрицы коэффициентов корреляции залиты цветом в соответствии с уровнем корреляции. Для выявления групп высоко коррелирующих переменных ячейки на таких диаграммах часто упорядочивают в соответствии с результатами параллельно выполняемого кластерного анализа. На рис. 2 выделяются группы тесно связанных факторов: X1, X5 и X3, X4.

Рис. 2. Матрица коэффициентов парной корреляции в цвете

В состав пакета `caret` входит функция `findCorrelation`, которая возвращает вектор с номерами переменных, чей уровень корреляции с другими переменными превышает порог (аргумент `cutoff`), заданный пользователем и удаляет одну из них.

```
findCorrelation(R1,cutoff = 0.8,verbose = TRUE,names = TRUE,exact = ncol(R1)< 100)
Compare row 5 and column 1 with corr 0.96
Means: 0.375 vs 0.261 so flagging column 5
All correlations <= 0.8
[1] «X5»
```

Эту функцию можно рекомендовать для применения только в том случае, когда нет предпочтений по содержательному составу факторов.

*Метод дополнительных регрессий.
Выявление мультиколлинеарности
с помощью исследования факторов
инфляции дисперсии*

Метод дополнительных регрессий заключается в том, что строятся регрессии для каждого j -го регрессора X_j ($j = 1, \dots, k$, k – число факторов модели), по всем остальным регрессорам, вычисляются коэффициенты детерминации R_j^2 и проверяется их значимость по F -критерию. Если коэффициент детерминации R_j^2 значимо отличается

от нуля, то считаем, что фактор X_j приводит к мультиколлинеарности. В R представлен близкий к описанному метод, основанный на вычислении и последующем исследовании так называемого фактора инфляции дисперсии VIF (Variance Inflation Factor).

$$VIF_j = \frac{1}{(1 - R_j^2)},$$

Если $VIF_{X_j} > 10$, то данный регрессор может привести к мультиколлинеарности.

В среде R для оценки факторов инфляции дисперсии применяется функция `vif(fm,..)`, где `fm` – объект, который содержит информацию о построенной модели из пакета `car`.

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

X1	X2	X3	X4	X5
21.112067	1.889292	2.474376	2.331028	23.389356

Анализируя полученные значения факторов инфляции для всех переменных, можно сделать вывод, что факторы X1 и X5 приводят к мультиколлинеарности.

Методов тестирования мультиколлинеарности разработано достаточно много, но нет единого критерия, применяя который можно сделать однозначный вывод о присутствии мультиколлинеарности. Более того, различные методы обнаружения мультиколлинеарности плохо сопоставимы друг с другом. Поэтому в практических задачах применяют несколько методов, полученные результаты обобщаются для получения обоснованного вывода. Например, тест на мультиколлинеарность Фарра – Глоубера [1] содержит три вида статистических критериев проверки: наличия мультиколлинеарности в общем, критерий («хи-квадрат»); выявление регрессоров, связанных с другими регрессорами (F-критерий); выявление пар тесно связанных переменных (t-тест) дополняют VIF тестом. Хорошо известный метод Бэлси [6], который реализован в Gretl и называется диагностика Belsley-Kuh-Welsch (BKW) основан на вычислении собственных значений и собственных векторов матрицы. Этот метод позволяет не только выявить мультиколлинеарность среди исходных переменных, но и определить регрессоры, которые показывают: наибольший вклад в зависимости полезно сочетать с VIF тестом и анализом матрицы коэффициентов парной корреляции [7].

Постоянно ведется разработка новых методов исследования мультиколлинеарности. Интересные индикаторы мультиколлинеарности в моделях линейной регрессии представили Д. Курто и Х.К. Пинто [8]. Показатель, который еще не нашел широкого применения, для диагностики мультиколлинеарности с использованием собственных значений или количественной оценки средней корреляции данных «Red Indicator» разработал П. Ковач [9], но этот тест уже включен в рассматриваемый ниже пакет R mctest.

Пакет R mctest для обнаружения мультиколлинеарности

В 2016 г. вышла работа «mctest: пакет R для обнаружения коллинеарности среди регрессоров» [10], а в конце февраля 2018 г. документация пакета была размещена в «CRAN» [11]. В пакете mctest есть две функции: omcdiag и imcdiag, которые обеспечивают общую и индивидуальную диагностическую проверку мультиколлинеарности соответственно, функция mc.plot, которая может использоваться для графического

отображения VIF и собственных значений, и функция eigprop. Эта функция вычисляет собственные значения, индексы обусловленности и пропорции дисперсионного разложения коэффициентов регрессии. Фактически в этой функции реализован метод Belsley-Kuh-Welsch [6]. Чтобы проверить линейные зависимости, связанные с соответствующим собственным значением, eigprop сравнивает пропорцию дисперсии с пороговым значением (по умолчанию 0,5) и отображает пропорции, превышающие заданный порог из каждой строки и столбца, если таковые имеются.

Анализ диагностики мультиколлинеарности по методу Бэлси с помощью функции eigprop, которой приведен на рис. 3, показал, что наибольшему значению индекса обусловленности (CI) равному 376,49 в шестой строке соответствуют факторы X1 и X5, значит, между этими факторами существует тесная зависимость. Заметим, что, по сравнению с предыдущими тестами, появилась новая информация – в 5 строке зафиксирована тесная связь между переменными X3 и X4.

Выявив с помощью функции eigprop переменные, участвующие в зависимости, удаляют из регрессионной модели одну из этих переменных.

В пакете mctest тесты выявления мультиколлинеарности классифицированы по двум признакам: проверка мультиколлинеарности всего массива данных – общие тесты проверки мультиколлинеарности и выявление регрессоров, приводящих к мультиколлинеарности – индивидуальные тесты.

Общие тесты проверки мультиколлинеарности

Функция omcdiag() реализует общую диагностическую проверку мультиколлинеарности. В эту функцию включены следующие тесты:

- Проверка равенства нулю определителя. Оценки коэффициентов регрессии являются решением системы нормальных уравнений $(X^T X)\hat{\beta} = X^T Y$. Если регрессоры мультиколлинеарны, то определитель нормированной корреляционной матрицы $|X^T X|$ близок к нулю. По умолчанию порог равен 0,1 (detr = 0.01).

- Тест Фарра – Глоубера (1 часть) (При рассмотрении командных кодов функции omcdiag при вычислении статистики Фарра – Глоубера выявлена допущенная небольшая ошибка)

Проверка наличия мультиколлинеарности всего массива переменных (критерий


```
library ("mctest")
```

```
eigprop(ТТ, Inter = TRUE, prop = 0.7)
```

```
Call:
```

```
eigprop(x = ТТ, Inter = TRUE, prop = 0.7)
```

	Eigenvalues	CI	Intercept	x1	x2	x3	x4	x5
1	5.6388	1.0000	0.0000	0.0003	0.0029	0.0001	0.0001	0.0000
2	0.1963	5.3592	0.0000	0.0358	0.1861	0.0000	0.0001	0.0000
3	0.1605	5.9279	0.0000	0.0129	0.3531	0.0009	0.0009	0.0000
4	0.0033	41.1701	0.0082	0.0042	0.0001	0.1418	0.1144	0.0046
5	0.0011	72.8276	0.0000	0.0078	0.0561	0.7865	0.8736	0.0001
6	0.0000	376.4908	0.9917	0.9390	0.4016	0.0706	0.0110	0.9952

```
=====  
Row 6==> x1, proportion 0.938953 >= 0.70  
Row 5==> x3, proportion 0.786549 >= 0.70  
Row 5==> x4, proportion 0.873616 >= 0.70  
Row 6==> x5, proportion 0.995243 >= 0.70
```

Рис. 3. Результат диагностики коллинеарности по методу Белли

«хи-квадрат»). Значение статистики Фарра-Глоубера вычисляется по формуле

$$FG_{\text{набл}} = - \left[n - 1 - \frac{1}{6}(2k + 5) \right] \ln(\det[R]),$$

где n – количество наблюдений, k – количество факторов, $\det[R]$ – определитель корреляционной матрицы. Эта статистика имеет распределение χ^2 (хи-квадрат). Вычисленное значение этого критерия сравнивается с табличным значением χ^2 с $0,5k(k-1)$ степенями свободы и уровне значимости α . Если $FG_{\text{набл}}$ больше табличного, то считаем, что в массиве объясняющих переменных присутствует мультиколлинеарность.

Red Indicator (Красный индикатор)

Красный индикатор П. Ковача [12] можно использовать для количественной оценки степени избыточности. Определить его можно следующим образом:

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

где k – количество регрессоров,

$$V_{\lambda} = \sqrt{\frac{\sum_{j=1}^m (\lambda_j - 1)^2}{k}} = \sigma_{\lambda}, \quad \lambda_j - \text{собственные}$$

числа матрицы $(X^T X)$. Если значение Красного индикатора равно нулю (Красный = 0), то это свидетельствует об отсутствии избыточности, а значения, близкие к 1 (Красный ~ 1), указывают максимальную избыточность (мультиколлинеарность). По умолчанию порог равен 0,5 (red = 0,5).

• Сумма обратных значений собственных чисел

В ортогональной системе $\sum_{j=1}^k \frac{1}{\lambda_j} = k$, следовательно, для матрицы корреляции R с собственными числами λ_j $\sum_{j=1}^k \frac{1}{\lambda_j}$ может использоваться для выявления коллинеарности. Если $\sum_{j=1}^k \frac{1}{\lambda_j}$ больше, чем число регрессоров, используемых в модели, то регрессоры коллинеарны.

• Theil индикатор

Theil предложил меру коллинеарности на основе вклада вспомогательных регрессий $(R^2 - R_j^2)$ в коэффициент детерминации, где R_j^2 является R^2 вспомогательной регрессии. $k = R^2 - \sum_{i=1}^k (R^2 - R_i^2)$. Если

$k = 0$, то все X взаимно некоррелированы (нет избыточности), если $k \sim 1$, то коллинеарность существует среди регрессоров. По умолчанию порог равен 0,5 (theil = 0,5).

• Индекс обусловленности (CI)

Число обусловленности (индекс, параметр обусловленности) $CI = \text{cond}(X^T X)$ показывает, во сколько раз может возрасти относительная погрешность решения системы нормальных уравнений по сравнению с относительной погрешностью исходных данных. Для симметричной положительно определённой матрицы $(X^T X)$ индекс обусловленности равен $CI_j = \sqrt{\frac{\lambda_{\max}}{\lambda_j}}$. По умолчанию порог равен 30 (cn = 30).

```
library("mctest")
omcdiag(x = XX, y = Y, na.rm = TRUE, Inter = TRUE, detr = 0.01, red = 0.5,
+       conf = 0.95, theil = 0.5, cn = 30)
```

```
Call:
omcdiag(x = XX, y = Y, na.rm = TRUE, Inter = TRUE, detr = 0.01,
       red = 0.5, conf = 0.95, theil = 0.5, cn = 30)
```

overall Multicollinearity Diagnostics

	MC Results	detection
Determinant $ X'X $:	0.0177	0
Farrar Chi-Square:	48.4333	1
Red Indicator:	0.4032	0
Sum of Lambda Inverse:	51.1961	1
Theil's Method:	-0.0154	0
Condition Number:	9.8309	0

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

Рис. 4. Результат общей диагностической проверки мультиколлинеарности функции *omcdiag* в пакете *mctest*

Результаты тестирования мультиколлинеарности с помощью функции *omcdiag* приведены на рис. 4. При заданных пороговых значениях только два теста показали наличие мультиколлинеарности: Фаррара – Глоубера и Sum of Lambda Inverse (сумма обратных значений собственных чисел).

Диагностические показатели коллинеарности для каждого регрессора

Функция *imcdiag()* реализует диагностическую проверку мультиколлинеарности, выявляя влияние каждого регрессора. Эта функция включает в себя семь различных тестов.

- VIF тест, который входит также в пакет *car*;
- тест TOL

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

Если $VIF > 10$ или значение $TOL_j \sim 0$, то это указывает на существование коллинеарности среди регрессоров.

- Farrar wi (тест Фаррара – Глоубера, 2 часть)

Проверка наличия мультиколлинеарности каждой переменной с другими переменными (F-критерий):

$$w_j = \frac{R_j^2}{1 - R_j^2} \left(\frac{n-k}{k-1} \right) \sim F_{(n-k, k-1)}.$$

Если $w_j > F_{(\alpha, n-k, k-1)}^*$, то есть, если вычисленный F-критерий превышает критиче-

ский $F_{(\alpha, n-k, k-1)}^*$ на выбранном уровне значимости, то следует понимать, что конкретный X_j коллинеарен другим регрессорам.

- Fi [13]. Тест практически дублирует предыдущий тест.

Связь F-критерия и R^2 модели регрессии X_j на остальные регрессоры может быть использована для обнаружения мультиколлинеарности. Связь описана следующим образом:

$$F_j = \frac{R_{X_j, X_1 \dots X_2}^2 / k - 2}{1 - R_{X_j, X_1 \dots X_2}^2 / n - k - 1} \sim F_{(k-2, n-k+1)}.$$

Если вычисленный F-критерий превышает критический $F_{(\alpha, n-k, k-1)}^*$ на выбранном уровне значимости, то следует понимать, что конкретный X_j коллинеарен другим регрессорам и его следует исключить из модели.

- Leamer's test

Лимер [14] предложил меру измерения влияния мультиколлинеарности для j -го регрессора

$$C_j = \left\{ \frac{\left(\sum (X_{ij} - \bar{X}_j)^2 \right)^{-1}}{(X^T X)_{jj}^{-1}} \right\}^{0.5}.$$

Если X_j не коррелирует с другими регрессорами, C_j будет равен 1. В противном случае будет равен $(1 - R_j^2)^{0.5}$ т.е., $C_j \sim 0$ указывает на существование коллинеарности среди регрессоров.

```
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, na.rm = TRUE, 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	21.1121	0.0474	55.3082	80.4483	0.2176	-3.5596	1
X2	1.8893	0.5293	2.4456	3.5572	0.7275	-0.3185	0
X3	2.4744	0.4041	4.0545	5.8975	0.6357	-0.4172	0
X4	2.3310	0.4290	3.6603	5.3241	0.6550	-0.3930	0
X5	23.3894	0.0428	61.5707	89.5574	0.2068	-3.9435	1

1 --> COLLINEARITY is detected by the test

0 --> COLLINEARITY is not detected by the test

Рис. 5. Результат диагностической проверки индивидуального влияния факторов

```
> mctest(XX, Y, type="b", all = TRUE)
```

```
Call:
omcdiag(x = x, y = y, Inter = Inter, detr = detr, red = red,
       conf = conf, theil = theil, cn = cn)
```

Overall Multicollinearity Diagnostics

	MC Results	detection
Determinant X'X :	0.0177	0
Farrar Chi-Square:	48.4333	1
Red Indicator:	0.4032	0
Sum of Lambda Inverse:	51.1961	1
Theil's Method:	-0.0154	0
Condition Number:	9.8309	0

1 --> COLLINEARITY is detected by the test

0 --> COLLINEARITY is not detected by the test

=====

```
Call:
imcdiag(x = x, y = y, method = method, corr = FALSE, vif = vif,
       tol = tol, conf = conf, cvif = cvif, leamer = leamer, all = all)
```

All Individual Multicollinearity Diagnostics in 0 or 1

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

1 --> COLLINEARITY is detected by the test

0 --> COLLINEARITY is not detected by the test

X1 , X2 , X3 , X4 , coefficient(s) are non-significant may be due to multicollinearity

Рис. 6. Результат работы функции mctest

- CVIF

Curto и Pinto [15] предложили меру мультиколлинеарности для оценки влияния корреляции между регрессорами в дисперсию

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

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

Если $|CVIF_j| \geq 10$ то можно считать, что j -й регрессор приводит к мультиколлинеарности.

- Klein's rule [16]

Если коэффициент детерминации дополнительной регрессии (регрессии, объясняющей переменной j по всем остальным регрессорам) больше, чем коэффициент детерминации основной регрессии (Y на все регрессоры X), то мультиколлинеарность можно считать установленной.

На рис. 5 приведены результаты тестирования мультиколлинеарности с помощью функции `imcdiag`.

Большинство тестов указывают, что X_1 и X_5 являются причиной мультиколлинеарности.

Надо отметить, что в пакете `mctest` есть функция `mctest`, в которой реализованы различные режимы, в том числе возможность выполнения функций `omcdiag` и `imcdiag` (рис. 6).

Диагностические графики коллинеарности: график VIF и собственных значений

Функция `mc.plot` может использоваться для отображения графиков значений VIF и собственных значений, по которым можно судить о существовании коллинеарности среди регрессоров. Аргумент `vif = 10` и `ev = 0,1` являются пороговыми значениями для VIF и собственных значений, соответственно им будет отображаться горизонтальная красная пунктирная линия. Из графика VIF значения VIF каждого регрессора больше 10 указывают на существование мультиколлинеарности среди регрессоров (рис. 7).

Рис. 7. Графики VIF и собственных значений

Заключение

Для решения проблемы мультиколлинеарности разработано много программных продуктов [17]. В пакете `mctest`, созданном в среде R, реализована попытка объединить наиболее широко используемые и обсуждаемые диагностики мультиколлинеарности. При реализации допущены ошибки в некоторых тестах (тест Фаррара – Глоубера, `Red test`) функции `omcdiag`, дублирование тестов в функции `imcdiag`. К сожалению, в этом пакете отсутствует тест, основанный на анализе частных коэффициентов корреляции (третий тест Фаррара – Глоубера). При этом необходимо отметить, что функции `eiqrpr`, `omcdiag` и `imcdiag` облегчают пользователю формирование выводов о существовании мультиколлинеарности среди регрессоров, а также дают представление о том, какой регрессор может быть причиной мультиколлинеарности.

Список литературы

1. Эконометрика и эконометрическое моделирование: учебник / Л.О. Бабешко, М.Г. Бич, И.В. Орлова. – М.: Вузовский учебник: ИНФРА-М, 2018. – 384 с.
2. Орлова И.В., Турундаевский В.Б. Многомерный статистический анализ при исследовании экономических процессов: монография. – М.: МЭСИ, 2014. – 190 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. Проект R для статистических вычислений. URL: <http://www.r-project.org/> (дата обращения 12.03.2018).
5. Comprehensive R Archive Network. URL: <http://cran.r-project.org> (дата обращения: 12.04.2018).
6. Belsley D.A., Kuh E., Welsch R.E. Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. John Wiley & Sons; New York: 1980. P. 297.
7. 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.
8. Jose Dias Curto and Jose Castro Pinto New Multicollinearity Indicators in Linear Regression Models International Statistical Review / Revue Internationale de Statistique Vol. 75, No. 1 (April 2007), P. 114–121.
9. Kovacs P., Petres T., Toth L.A. A new measure of multicollinearity in linear regression // International Statistical Review / Revue Internationale de Statistique, 2005. No. 73(3), P. 405–412.
10. 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.
11. Muhammad Imdad Ullah, Muhammad Aslam Multicollinearity Diagnostic Measures. Package ‘`mctest`’. URL: <https://cran.r-project.org/web/packages/mctest/mctest.pdf> (дата обращения: 12.04.2018).
12. Peter Kovacs Examination of Multicollinearity in Linear Regression Models Examination of PETRES’ Red. Thesis of PhD Dissertation Szeged, 2008. URL: <http://docplayer.hu/3607097-Examination-of-multicollinearity-in-linear-regression-models-examination-of-petres-red.html> (дата обращения: 12.04.2018).
13. Gujarati D.N., Porter D.C. Basic Econometrics. McGraw Hill, 5 edition, 2008. [p. 495, 496, 497, 498, 503].
14. Leamer, Edward E. Multicollinearity: A Bayesian Interpretation. The Review of Economics and Statistics, MIT Press, 1973. vol. 55(3), P. 371–380.
15. Curto J.D., Pinto J.C. The corrected vif (cvif). J. Appl. Stat. 2011, 38(7), P. 1499–1507.
16. Green W.H. Econometric Analysis (7th ed.) / W.H. Green. – N.Y., 2012. 1024 p.
17. Шитиков В.К., Мاستицкий С.Э. Классификация, регрессия и другие алгоритмы Data Mining с использованием R. – 2017. – 351 с. URL: <https://github.com/ranalytics/data-mining> (дата обращения: 12.04.2018).