УДК 330.4:519.237.5

ПОДХОД К РЕШЕНИЮ ПРОБЛЕМЫ МУЛЬТИКОЛЛИНЕАРНОСТИ ПРИ АНАЛИЗЕ ВЛИЯНИЯ ФАКТОРОВ НА РЕЗУЛЬТИРУЮЩУЮ ПЕРЕМЕННУЮ В МОДЕЛЯХ РЕГРЕССИИ

Орлова И.В.

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

Данная статья посвящена рассмотрению методики отбора и ранжирования информативных факторов по степени их влияния на результирующую переменную в моделях регрессионного анализа в условиях мультиколлинеарности данных. Важную роль при оценке влияния факторов играют коэффициенты регрессионной модели. Однако непосредственно с их помощью нельзя сопоставить факторы по степени их влияния на зависимую переменную из-за различия единиц измерения и разной степени колеблемости. Для устранения таких различий при интерпретации применяются средние частные коэффициенты эластичности, бета-коэффициенты и дельта-коэффициенты, которые нельзя использовать при мультиколлинеарности данных. Для решения задачи уменьшения мультиколлинеарности предлагается подход, основанный на методе Белсли, позволяющий не только выявить мультиколлинеарность среди исходных регрессоров, но и определить переменные, которые показывают наибольшую вовлеченность в зависимости. Проиллюстрировано применение метода инфляционных факторов и метода Белсли для выявления и устранения мультиколлинеарности при решении задачи анализа значимости и степени влияния ряда традиционных показателей на переменную прибыль (убыток) до налогообложения ряда предприятий, относящихся к виду деятельности «Добыча сырой нефти за 2016 г.» Применение этих методов позволило устранить мультиколлинеарность и решить задачу ранжирования факторов по степени их влияния на результирующий показатель.

Ключевые слова: многофакторная регрессионная модель, коэффициенты эластичности, дельта-коэффициенты, мультиколлинеарность, коэффициент (индекс) обусловленности

APPROACH TO THE SOLUTION OF THE MULTICOLLINEARITY PROBLEM AT THE ANALYSIS OF THE INFLUENCE OF THE FACTORS ON THE RESULTING VARIABLE IN MODELS OF REGRESSION

Orlova I.V.

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

This article is devoted to the consideration of the method of selection and ranking of informative factors by the degree of their influence on the resulting variable in regression analysis models in conditions of multicollinear data. The coefficients of the regression model play an important role in assessing the influence of factors. However, directly with their help, it is impossible to compare the factors by the degree of their influence on the dependent variable because of the difference in units of measure and the varying degree of variability. To eliminate such differences, the interpretation uses the average partial elasticity coefficients, beta coefficients and delta coefficients, which cannot be used for multicollinear data. To solve the problem of reducing multi-collinearity, an approach based on the Balsley method is proposed, which allows not only to show multicollinearity among the initial regressors, but also to determine the variables that show the greatest involvement in the dependence. Application of a method of inflationary factors and the Balsley method for identification and elimination of multicollinearity is illustrated at the solution of a task of the analysis of the importance and extent of influence of a number of traditional indicators on variable profit (loss) before the taxation of a number of the enterprises relating to a kind of activity Extraction of crude oil for 2016. Application of these methods has allowed to eliminate multicollinearity and to solve a problem of ranging of factors of extent of their influence on a resultant indicator.

Keywords: multifactorial regression model, elasticity coefficients, delta coefficients, multicollinearity, conditionation coefficient (index)

Эконометрическое моделирование используется для решения различных прикладных задач. Но по конечным целям С.А. Айвазян выделяет две основные: «(а) прогноз экономических и социально-экономических показателей (переменных); (б) имитация различных возможных сценариев социально-экономического развития анализируемой системы, когда статистически выявленные взаимосвязи между характеристиками производства, потребления, социальной и финансовой политики и т.п. используются для прослеживания того, как планируемые (возможные) изменения тех или иных поддающихся управлению параметров производства или распределения скажутся на значениях интересующих нас «выходных» характеристик» [1]. При решении задач второго типа возможно использование многофакторных регрессионных моделей для выявления и ранжирования факторов по степени их влияния на результирующую переменную. В первую очередь при оценке влияния факторов на эндогенную переменную учитывают значения коэффициентов регрессионной моде-

ли. Однако непосредственно с их помощью нельзя сопоставить факторы по степени их влияния на зависимую переменную из-за различия единиц измерения и разной степени колеблемости [2].

Для устранения таких различий при интерпретации применяются средние частные коэффициенты эластичности, бета-коэффициенты и дельта-коэффициенты [2].

Коэффициенты эластичности $Э_j = \hat{a}_j \cdot \frac{\overline{x}_j}{\overline{y}}$ позволят сравнивать факторы в процентах и могут быть вычислены не только для линейной функции. Коэффициенты эластичности показывают, на сколько процентов в среднем изменится зависимая переменная при изменении объясняющей переменной на один процент при фиксированных значениях других объясняющих переменных. Коэффициент эластичности не учитывает степень колеблемости факторов.

Для решения вопроса сравнения силы влияния факторов, имеющих разную степень колеблемости, используют бета-коэффициенты, или коэффициенты регрессии в стандартизованном виде: $\beta_j = \hat{a}_j \cdot \frac{S_{x_j}}{S_y}$. Стандартизованные коэффициенты сравнимы между собой, поэтому с их помощью можно ранжировать факторы по силе воздействия на результирующую переменную. Бета-коэффициент показывает, на какую часть величины среднего квадратического отклонения S_y изменится эндогенная переменная с изменением соответствующей экзогенной переменной x_j на величину своего среднеквадратического отклонения при фиксированном значении остальных независимых переменных.

Перечисленные коэффициенты позволяют упорядочить факторы по степени влияния факторов на эндогенную переменную. Однако еще раз обратим внимание на важный момент – коэффициенты эластичности, бета-коэффициенты, как и коэффициенты регрессии, могут интерпретироваться только при условии, что остальные переменные в модели регрессии неизменны, когда изменение одной переменной не приводит к изменению других переменных [2].

При наличии мультиколлинеарности переменных по коэффициентам регрессии нельзя судить о влиянии этих переменных на функцию [3]. Существуют различные методы, направленные на выявление мультиколлинеарности: анализ матриц коэффициентов парной и частной корреляции, метод Фаррара – Глоубера [3], метод дополнительных регрессий, тесно связанный с методом инфляционных факторов (VIF) и другие. Выявление мультиколлинеарности с помощью VIF реализовано во многих программах, в том числе в R и Gretl [4].

Фактор инфляции дисперсии VIF (Variance Inflation Factor) показывает, во сколько раз увеличивается (вздувается) дисперсия коэффициента регрессии за счёт коррелированности регрессоров X1,…,Xk по сравнению с дисперсией этого коэффициента, если бы регрессоры были некоррелированы. Фактор инфляции дисперсии вычисляется по формуле

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

где R_j^2 – коэффициент детерминации j-го регрессора X_j, ($j = 1,\ldots,k$, k – число факторов модели), по всем остальным регрессорам. Если фактор инфляции дисперсии равен единице, то это свидетельствует об ортогональности вектора значений признака остальным. Высокие значения фактора инфляции дисперсии соответствуют почти линейной зависимости j-го столбца от остальных, происходящей из-за высокой корреляции данных. Принято считать, что значение от 1 до 2 (R_j^2 от 0 до 0,5) означает, что включение X_j в модель не приводит к мультиколлинеарности.

Если $VIF_j = 16$, то стандартная ошибка оценки параметра βj в 4 раза ($\sqrt{VIF_j} = \sqrt{16} = 4$) превышает эту оценку, полученную при полном отсутствии мультиколлинеарности. Считается, что если $VIF_{x_j} > 10$, то данный регрессор приводит к мультиколлинеарности. Недостатками этого критерия мультиколлинеарности является то, что он может принимать большие значения сразу для нескольких признаков, что мешает определить, какой из признаков необходимо удалить.

В программе Gretl представлен ещё один метод, основанный на вычислении собственных значений и собственных векторов матрицы X^TX, позволяющий не только выявить мультиколлинеарность среди исходных регрессоров, но и определить переменные, которые показывают наибольшую вовлеченность в зависимости. Этот метод известен как метод Белсли [5]. В Gretl метод представлен как диагностика коллинеарности Belsley-Kuh-Welsch (BKW).

Если матрица X^TX обратима, то

$$(X^TX)^{-1} = U\Lambda^{-1}U^T,$$

где Λ^{-1} – матрица, обратная к диагональной матрице собственных значений матрицы X^TX, расположенных в порядке убывания; U – матрица, столбцами которой являются нормированные собственные векторы, то

есть сумма квадратов координат каждого вектора равна 1.

Дисперсия коэффициента регрессии $\hat{\alpha}_j$ равна

$$Var(\hat{\alpha}_j) = \hat{\sigma}^2 (\mathrm{X}^T \mathrm{X})^{-1}_{jj},$$

где $\hat{\sigma}^2$ – выборочная дисперсия остатков.

Вычислив диагональные элементы матрицы $U\Lambda^{-1}U^T$, получаем представление оценки дисперсии параметров регрессии в виде суммы p слагаемых

$$\hat{\sigma}^{-2} Var(\hat{\alpha}_i) = (\mathrm{X}^T \mathrm{X})^{-1}_{ii} = (U\Lambda^{-1}U^T)_{ii} =$$

$$= \left(\frac{u_{i1}^2}{\lambda_1} + \frac{u_{i2}^2}{\lambda_2} + ... + \frac{u_{ip}^2}{\lambda_p} \right) =$$

$$= (q_{i1} + q_{i2} + ... + q_{ip}) \sum\nolimits_{k=1}^{p} \frac{u_{ik}^2}{\lambda_k},$$

где λ_k – собственное число матрицы X^TX; p – количество параметров в модели регрессии, $\lambda_1 > \lambda_2 > \ldots > \lambda_p$,

q_{ij} – доля j-го собственного вектора в дисперсии i-го коэффициента регрессии $\hat{\alpha}_i$:

$$q_{ij} = \frac{u_{ij}^2}{\lambda_j} / \sum\nolimits_{k=1}^{p} \frac{u_{ik}^2}{\lambda_k}.$$

Коэффициент (индекс) обусловленности η_j вычисляется по формуле

$$\eta_i = \sqrt{\lambda_{\max}} / \sqrt{\lambda_i}, \; i = 1,2,...,p,$$

где $\lambda_{\max} = \lambda_1$ – максимальное собственное число.

Для обнаружения мультиколлинеарности признаков в программе Gretl выдаются значения факторов инфляции дисперсии (Метод инфляционных факторов) и таблица диагностики коллинеарности Belsley-Kuh-Welsch [6] (табл. 1), в которой каждая строка соответствует своему индексу обусловленности η_j, *а* элементы строки – значения $q_{ij.}$. Сумма элементов по столбцам равна 1.

Большие величины η_j означают, что, возможно, есть зависимость между регрессорами. Большие значения q_{ij} в строках, соответствующих большим величинам η_j, относятся к регрессорам, между которыми эта зависимость существует.

Вопрос о том, какое значение коэффициента обусловленности считать большим, решается в каждом конкретном случае индивидуально, в зависимости от ценности информации, её объёма, целей и задач исследования. Часто значения коэффициента обусловленности считаются большими, если они больше 10 [7].

Относительно близкое к нулю λ_j приводит к большим коэффициентам обусловленности. Нулевое значение λ_j означает, что существует строгая мультиколлинеарность. Большие значения коэффициентов обусловленности свидетельствуют о наличии зависимостей; большие значения долевых коэффициентов q_{ij} внутри соответствующих строк указывают столбцы матрицы X, участвующие в зависимостях.

Отметим также, что если больших коэффициентов обусловленности больше одного, то в зависимостях могут участвовать все переменные, которые имеют большие суммарные значения коэффициентов в последних строках таблицы с большими коэффициентами обусловленности.

Проиллюстрируем применение метода инфляционных факторов и метода Белсли для выявления и устранения мультиколлинеарности при решении задачи анализа значимости и степени влияния ряда традиционных показателей на переменную прибыль (убыток) до налогообложения ряда предприятий, относящихся к виду деятельности «Добыча сырой нефти» (Система СПАРК 13.12.2017 [8]). Была сделана выборка данных, представляющих финансовые показатели 186 фирм за 2016 г. В качестве регрессоров модели использованы более 20 переменных – среднесписочная численность работников, рентабельность активов, стоимость основных производственных средств и оборудования, стоимость совокупных активов, материальные активы и др.

Таблица 1

Разложение дисперсии коэффициентов регрессии

Собственные числа матрицы X^TX (lambda)	Индекс обусловленности (cond)	Доли дисперсии			
		const	$\mathbf{x_1}$	**…**	$\mathbf{x_p}$
λ_1	η_1	q_{11}	q_{21}	…	q_{p1}
λ_2	η_2	q_{12}	q_{22}	…	q_{p2}
…	…	…	…	…	…
λ_p	η_p	q_{1p}	q_{2p}		q_{pp}

На основании собранных данных была построена эконометрическая линейная модель множественной регрессии. Анализ мультиколлинеарности, выполненный средствами пакета Gretl, показал наличие сильной мультиколлинеарности (табл. 2 и рис. 1). Из табл. 2 видно, что факторы X2, X7, X15 имеют самые большие значения VIF, скорее всего именно эти факторы приводят к мультиколлинеарности [7].

Анализ диагностики мультиколлинеарности по методу Белсли, фрагмент которой приведен на рис. 1, показал, что наибольшему значению индекса обусловленности равному 6317,33 соответствуют факторы X2, X7, X12, X15 с большими значениями q_{ij}, между этими факторами существует тесная зависимость. Выявив с помощью метода Белсли переменные, участвующие в зависимости, удаляем из регрессионной модели одну из этих переменных. На первом шаге удаляем X2. Затем параметры модели оцениваются заново. Если обнаружена другая зависимость, исключаем из модели одну из переменных второй группы и далее повторяем исследование на мультиколлинеарность. Через несколько шагов была получена модель, не содержащая коллинеарных факторов, все коэффициенты которой значимы (рис. 2 и 3):

$$\hat{Y}_i = -316364246{,}83 + 1{,}76X_5 + \\ +0{,}026X_9 + 0{,}10X_{15} + 2{,}88X_{17} + \\ +0{,}16X_{18} + 33787786{,}71X_{20}.$$

Belsley-Kuh-Welsch collinearity diagnostics:

		--- variance proportions ---							
lambda	cond	const	x1	x2	x3	x4	x5	x6	x7
10,035	1,000	0,001	0,001	0,000	0,000	0,000	0,000	0,001	0,000
2,773	1,902	0,000	0,000	0,000	0,000	0,001	0,001	0,001	0,000
2,156	2,158	0,033	0,000	0,000	0,008	0,000	0,011	0,000	0,000
1,667	2,454	0,001	0,022	0,000	0,007	0,000	0,022	0,000	0,000
1,281	2,799	0,002	0,000	0,000	0,070	0,000	0,011	0,023	0,000
1,082	3,045	0,014	0,085	0,000	0,051	0,002	0,021	0,047	0,000
1,017	3,141	0,005	0,009	0,000	0,016	0,000	0,119	0,026	0,000
0,946	3,258	0,018	0,000	0,000	0,544	0,000	0,123	0,008	0,000
0,704	3,775	0,151	0,029	0,000	0,131	0,000	0,000	0,060	0,000
0,683	3,833	0,006	0,096	0,000	0,000	0,000	0,076	0,017	0,000
0,661	3,895	0,004	0,073	0,000	0,070	0,000	0,068	0,022	0,000
0,596	4,102	0,004	0,002	0,000	0,003	0,001	0,022	0,068	0,000
0,498	4,489	0,489	0,043	0,000	0,021	0,000	0,006	0,001	0,000
0,434	4,806	0,002	0,193	0,000	0,005	0,002	0,023	0,046	0,000
0,371	5,199	0,019	0,066	0,000	0,002	0,002	0,004	0,036	0,000
0,315	5,648	0,031	0,005	0,000	0,012	0,001	0,214	0,076	0,000
0,285	5,938	0,021	0,030	0,000	0,002	0,001	0,000	0,026	0,000
0,178	7,503	0,015	0,000	0,000	0,000	0,000	0,000	0,034	0,000
0,143	8,387	0,000	0,130	0,000	0,005	0,000	0,000	0,097	0,000
0,078	11,372	0,138	0,012	0,000	0,001	0,008	0,068	0,105	0,000
0,057	13,254	0,004	0,005	0,000	0,013	0,003	0,019	0,009	0,000
0,025	20,034	0,000	0,122	0,000	0,000	0,254	0,078	0,244	0,000
0,013	28,120	0,018	0,052	0,000	0,000	0,598	0,008	0,009	0,000
0,002	69,787	0,003	0,012	0,000	0,009	0,002	0,000	0,039	0,000
0,000	151,894	0,000	0,007	0,000	0,000	0,057	0,093	0,003	0,003
0,000	6317,330	0,019	0,003	1,000	0,030	0,067	0,014	0,002	0,997

lambda = eigenvalues of X'X, largest to smallest
cond = condition index
note: variance proportions columns sum to 1.0

Рис. 1. Фрагмент диагностики коллинеарности по методу Белсли

Таблица 2

Значения инфляционных факторов

VIF(X1)	VIF(X2)	VIF(X3)	VIF(X4)	VIF(X5)	VIF(X6)	VIF(X7)
2,05	**2132100**	1,11	40,81	1,78	2,77	**265088,9**

VIF(X8)	VIF(X9)	VIF(X10)	VIF(X11)	VIF(X12)	VIF(X13)	VIF(X14)
13,66	985,55	4,66	4,46	**1057523**	6,73	7,05

VIF(X15)	VIF(X16)	VIF(X17)	VIF(X18)	VIF(X19)	VIF(X20)	VIF(X21)
92723,61	7,24	3,99	**258,42**	**248,41**	1,2	1,14

VIF(X22)	VIF(X23)	VIF(X24)	VIF(X25)
18,75	7,45	3,78	1,41

Модель 7: МНК, использованы наблюдения 1-186

Зависимая переменная: Y

	Коэффициент	*Ст. ошибка*	*t-статистика*	*P-значение*	
const	−3,16876e+08	3,38695e+08	−0,9356	0,3508	
x5	1,76225	0,137719	12,7960	<0,0001	***
x9	0,0259503	0,0124928	2,0772	0,0392	**
x15	0,0980736	0,0241297	4,0644	<0,0001	***
x17	2,88453	0,536359	5,3780	<0,0001	***
x18	0,15851	0,0126406	12,5398	<0,0001	***
x20	3,37813e+07	1,36184e+07	2,4806	0,0140	**

Среднее зав. перемен	3,69e+09	Ст. откл. зав. перемен	1,13e+10
Сумма кв. остатков	3,20e+21	Ст. ошибка модели	4,23e+09
R-квадрат	0,864876	Испр. R-квадрат	0,860346
F(6, 179)	190,9510	P-значение (F)	4,93e-75
Лог. правдоподобие	−4383,098	Крит. Акаике	8780,197
Крит. Шварца	8802,777	Крит. Хеннана – Куинна	8789,347

Рис. 2. Результат оценки параметров модели МНК

gretl: мультиколлинеарность

```
Метод инфляционных факторов
Минимальное возможное значение = 1.0
Значения > 10.0 могут указывать на наличие мультиколлинеарности

          x5    1,064
          x9    1,120
         x15    1,803
         x17    1,895
         x18    2,432
         x20    1,043

VIF(j) = 1/(1 - R(j)^2), где R(j) - это коэффициент множественной корреляции
между переменной j и другими независимыми переменными

Belsley-Kuh-Welsch collinearity diagnostics:

                     --- variance proportions ---
    lambda      cond     const        x5        x9       x15       x17       x18       x20
     2,637     1,000     0,036     0,016     0,020     0,049     0,040     0,039     0,014
     1,201     1,482     0,066     0,245     0,114     0,004     0,070     0,030     0,109
     0,998     1,625     0,003     0,020     0,365     0,013     0,016     0,000     0,449
     0,778     1,841     0,401     0,591     0,009     0,004     0,003     0,007     0,102
     0,708     1,930     0,493     0,118     0,259     0,015     0,002     0,001     0,325
     0,425     2,490     0,000     0,008     0,233     0,622     0,409     0,000     0,000
     0,253     3,226     0,001     0,001     0,000     0,293     0,460     0,923     0,001

  lambda = eigenvalues of X'X, largest to smallest
  cond   = condition index
  note: variance proportions columns sum to 1.0
```

Рис. 3. Результат теста на мультиколлинеарность

В последнюю модель вошли следующие факторы: X5 – Денежные средства, ед. RUB; X9 – Займы и кредиты (долгосрочные), ед. RUB; X15 – Краткосрочные обязательства, ед. RUB; X17 – Нематериальные активы, ед. RUB; X18 – Оборотные активы, ед. RUB; X20 – Рентабельность активов (ROA), %. Коэффициент детерминации 0,865 и скорректированный коэффициент детерминации свидетельствуют о хорошем приближении модели исходным данным.

Анализ теста на мультиколлинеарность последней модели показал ее отсутствие. Значение факторов инфляции от 1 до 2,4; наибольшее значение индекса обусловленности 3,2 (рис. 3).

Избавившись от мультиколлинеарности, можно использовать полученное урав-

нение регрессии для оценки влияния факторов на зависимую переменную с помощью бета и дельта-коэффициентов Δ (j):

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

где r_{y,x_j} – коэффициент парной корреляции между фактором X_j и зависимой переменной, R^2 – коэффициент детерминации.

Дельта-коэффициент показывает долю влияния фактора в суммарном влиянии всех факторов [3]. Полученные результаты приведены в табл. 3, из которой можно сделать вывод, что наибольшее влияние на переменную прибыль (убыток) до налогообложения предприятий, относящихся к виду деятельности «Добыча сырой нефти» оказывает фактор X18 – Оборотные активы (50,8 %), затем X5 – Денежные средства (19,3 %) и X17 – Нематериальные активы (15,4 %).

Таблица 3

Бета и дельта-коэффициенты

$\hat{\beta}_5$	$\hat{\beta}_9$	$\hat{\beta}_{15}$	$\hat{\beta}_{17}$	$\hat{\beta}_{18}$	$\hat{\beta}_{20}$
0,363	0,060	0,149	0,203	0,538	0,070
Δ_5	Δ_9	Δ_{15}	Δ_{17}	Δ_{18}	Δ_{20}
0,193	0,015	0,113	0,154	0,508	0,017

Таким образом, была решена задача ранжирования факторов по степени их влияния на результирующий показатель.

Список литературы

1. Айвазян С.А. Методы эконометрики. – М.: Магистр: ИНФРА-М, 2010. – 512 с.

2. Орлова И.В., Половников В.А. Экономико-математические методы и модели: компьютерное моделирование: учебное пособие, 3-е изд., перераб. и доп. – М.: Вузовский учебник: ИНФРА-М, 2012. – 389 с.

3. Орлова И.В., Турундаевский В.Б. Многомерный статистический анализ при исследовании экономических процессов: монография. – М.: МЭСИ, 2014. —190 с.

4. Куфель Т. Эконометрика. Решение задач с применением пакета программ Gretl. [Текст] – М.: Горячая линия – Телеком, 2007. – 200 с.

5. Regression Diagnostics – Identifying Influential Data and Sources of Collinearity / David A. Belsley, Edwin Kuh, Roy E. Welsch // John Wiley & Sons. – N.Y., 1980. – P. 297.

6. Дрейпер Норман, Смит Гарри. Прикладной регрессионный анализ, 3-е изд.: Пер. с англ. – М.: Издательский дом «Вильямс», 2007. – 912 с.

7. 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.

8. СПАРК – Проверка контрагента [Электронный ресурс]. – Режим доступа: http://www.spark-interfax.ru (дата обращения 13.12.2017).