УДК 330

МЕТОД ПОСТРОЕНИЯ СОДЕРЖАТЕЛЬНО ИНТЕРПРЕТИРУЕМЫХ РЕГРЕССИОННЫХ МОДЕЛЕЙ В УСЛОВИЯХ МУЛЬТИКОЛЛИНЕАРНОСТИ

Мокшина Светлана Ивановна, канд. экон. наук, доц.
Шуршикова Галина Владимировна, канд. техн. наук, доц.
Щекунских Светлана Станиславовна, канд. физ-мат. наук, доц.

Воронежский государственный университет, Университетская пл., 1, Воронеж, Россия, 394018; e-mail: mokshina@econ.vsu.ru; shurshikova@econ.vsu.ru; shchekunskikh@econ.vsu.ru

Цель: построение линейных регрессионных моделей без искажающих содержательный смысл эффектов мультиколлинеарности. *Обсуждение*: вопрос эконометрического моделирования в ситуации, когда имеет место мультиколлинеарность, продолжает оставаться актуальным, несмотря на достаточно большой арсенал рекомендаций. Практически все рекомендации ориентированы на решение проблем вычислительного характера, связанных с вырожденностью матрицы нормальных уравнений. Можно считать, что проблемы вычислительного характера в основном решены. Но проблема, связанная с содержательным смыслом окончательного варианта эконометрической модели, ни в одном из методов не рассматривается и по-прежнему остается не решенной. *Результаты*: предлагается оригинальный способ построения линейных регрессионных моделей для случаев, когда эффекты мультиколлинеарности проявляются искажениями содержательного смысла факторного влияния на моделируемый показатель. Проведенные эмпирические исследования подтвердили эффективность предлагаемого способа.

Ключевые слова: эконометрическое моделирование, главные компоненты, мультиколлинеарность.

DOI: 10.17308/meps.2017.5/1700

Введение

Эконометрическое моделирование является важным инструментом, используемым для анализа взаимодействия экономических процессов, обоснования нормативных величин, проведения прогнозных решений. Но все эти возможности имеют смысл только в тех случаях, когда результаты моделирования согласуются с экономической теорией, подтверждая ее количественными оценками. Однако не всегда результаты моделирования соответствуют нашим ожиданиям.

Чаще всего содержательный смысл регрессионных моделей

$$\mathbf{y} = \mathbf{Xb} + \boldsymbol{\varepsilon} \qquad (1)$$

искажают эффекты мультиколлинеарности, природа которых связана с вычислительными проблемами обращения матрицы $(\mathbf{X}'\mathbf{X})$ системы нормальных уравнений, определитель которой в этом случае близок к нулю. Фактически мультиколлинеарность зачеркивает возможность построения правильно специфицированной модели, что, по сути, ставит вопрос о целесообразности построения эконометрических моделей для их использования в экономическом анализе. Искаженные результаты моделирования приводят к ошибочным выводам анализа, которые не могут использоваться для обоснования принимаемых решений. Поэтому в настоящее время разработано несколько рекомендаций, обеспечивающих возможность построения регрессионных уравнений в условиях мультиколлинеарности. В основу этих рекомендаций положены различные подходы, основные из которых рассмотрены ниже.

Методы построения регрессионных уравнений в условиях мультиколлинеарности

Самый простой и чаще других используемый в практических работах, связанных с эконометрическим моделированием, является подход, предусматривающий исключение из модели одного или нескольких дублирующих друг друга, то есть линейно взаимосвязанных между собой факторов. Вопрос об идентификации исключаемых факторов решается с помощью матрицы парных коэффициентов корреляции, по элементам которой определяются тесно коррелирующие между собой факторы. Из каждой пары тесно коррелирующих факторов исключается тот, у которого теснота корреляционной связи с зависимой переменной ниже. Этот подход является эффективным в том смысле, что позволяет, как правило, устранить эффекты мультиколлинеарности. Однако он ограничивает возможности по достижению поставленной цели, предусматривающей изучение взаимосвязи моделируемого показателя именно с первоначальным набором факторов, а не с тем, который получен после удаления некоторых из них.

Сохранение первоначального набора факторов обеспечивает применение в этом случае главных компонент. Причем их применение предусматривает рассмотрение двух ситуаций: абсолютная мультиколлинеарность и частичная. При абсолютной мультиколлинеарности хотя бы одно из собственных значений равно нулю и, следовательно, число главных компонент меньше числа независимых переменных модели. Регрессия строится на главные компоненты с ненулевыми собственными значениями, а затем обратным преобразованием приводится к исходному виду.

При частичной мультиколлинеарности число главных компонент равно числу независимых переменных модели. Если все главные компоненты включить в модель, а затем вернуться к исходным переменным, то модель сохранит все нежелательные эффекты мультиколлинеарности. Поэтому в модель включаются только те главные компоненты, которые характеризуют основную долю вариации исходных переменных. Возможность использования главных компонент для целей устранения эффектов мультиколлинеарности объясняется следующим образом. Поскольку мультиколлинеарность связана с высокой степенью корреляции между факторными переменными, то возникает естественное желание заменить исходные факторы ортогональными переменными, которые представляют собой линейные комбинации исходных. В качестве коэффициентов в этих линейных комбинациях выступают компоненты собственных векторов ковариационной матрицы исходных переменных. Собственные вектора ортогональны, и построение регрессионной модели упрощается.

Оба подхода решают проблемы вычислительного характера, которые имеют место в случае мультиколлинеарности. Причем это осуществляется за счет целенаправленных информационных потерь (удаление факторов, использование не всех главных компонент). В то же время вопрос построения содержательно интерпретируемой модели в рамках этих подходов не обсуждается, так как не предусмотрен.

Есть подход, в котором реализуется процедура, основанная на получении смещенных оценок, которые в случае мультиколлинеарности обладают по сравнению с оценками МНК меньшими среднеквадратическими ошибками. В этом подходе используется ридж-оценивание, которое обеспечивает получение однопараметрического семейства оценок с помощью подправленной формулы МНК

$$\hat{\mathbf{b}} = \left(\mathbf{X}'\mathbf{X} + \alpha\mathbf{I}\right)^{-1}\mathbf{X}'\mathbf{y}, \qquad (2)$$

где $a \in [\underline{a}; \overline{a}]$ (как правило, $\underline{a} = 0{,}1$; $\overline{a} = 0{,}4$). Оптимальное значение этого параметра не определяется. Чем больше значение этого параметра, тем меньше стандартные ошибки получаемых оценок коэффициентов регрессии. Но одновременно увеличивается смещение этих оценок. В силу этого рекомендуется рациональный выбор этого параметра, который обеспечивает получение статистически значимых оценок, но со смещением, не превосходящим величину стандартных ошибок.

Этот подход, как и предыдущие подходы, в основном ориентирован на устранение только вычислительных проблем. Поэтому вопрос о получении содержательно интерпретируемых и правильно специфицированных регрессионных моделей остается открытым.

Идентификация эффектов мультиколлинеарности

Тестов для выявления мультиколлинеарности аналогичных тем, которые, например, используются при обнаружении гетероскедостичности или автокорреляции остатков, к сожалению, нет. Поэтому результаты построения регрессионной модели рекомендуется анализировать с точки зрения возможного их несоответствия сложившимся теоретическим представлениям о моделируемом процессе или логике взаимодействия факторных переменных. Как правило, мультиколлинеарность проявляется своими эффектами, которые обычно обнаруживаются по тем искажениям, которые присутствуют в построенной модели. Чаще всего искажениям подвергаются знаки коэффициентов при некоторых факторных переменных, а также статистическая значимость коэффициентов, определяемая с помощью t–статистики Стьюдента.

Если такие искажения обнаружены, то делается попытка поиска конкретных причин, вызвавших эти искажения. Причина общего характера известна – определитель системы нормальных уравнений имеет значение близкое к нулю. Но не известны причины частного характера, из-за которых значение определителя становится близким к нулю. На выяснение именно этих причин направлен анализ результатов моделирования.

Прежде всего имеет смысл выяснить, что происходить с моделью при удалении одной факторной переменной. Если это удаление приводит к слишком большим изменениям модели, то, скорее всего, удаленная переменная тесно связана с некоторыми переменными построенной модели и возможно она была причиной мультиколлинеарности. Эта взаимосвязь уточняется путем оценки тесноты линейной связи по парным коэффициентам корреляции.

Эффективным инструментом определения мультиколлинеарности является анализ факторов инфляции дисперсии (ФИД) [4]. ФИД *i*-го фактора записывается следующим образом

$$ФИД_i=(1–R_i^2)^{-1}, \qquad (3)$$

где R_i^2 – квадрат множественного коэффициента корреляции, характеризующего тесноту линейной связи *i*-й независимой переменной со всеми остальными независимыми переменными, включенными в регрессионное уравнение моделируемого показателя.

Из равенства ФИД единице следует, что тестируемый фактор не имеет линейных связей со всеми остальными факторами модели и поэтому не может быть причиной мультиколлинеарности. Последовательно вычисляя ФИД для всех факторов, можно определить фактор с максимальным значением ФИД. К сожалению, вероятностное распределение показателя ФИД неизвестно и, следовательно, критическое значение для ФИД не определено. Поэтому корректного однозначного суждения о причине мультиколлинеарности получить не удается.

В своей книге [1] Белсли предложил многоэтапную процедуру выявления мультиколлинеарности. В этой процедуре используются собственные вектора и собственные значения. Корни квадратные из собственных значений $\mu=\sqrt{\lambda}$ позволяют вычислить коэффициенты обусловленности

$$\eta_j = \mu_{\max} / \mu_j, \quad j = \overline{1,m}, \qquad (4)$$

по которым определяется число обусловленности в виде отношения $\mu_{\max} / \mu_{\min}$. Если число обусловленности очень большое, то данные в целом признаются мультиколлинеарными. Для определения факторов, участвующих в зависимостях, Белсли предлагает на основе собственных векторов и собственных значений строить специальные таблицы. Но критические значения для тестирования Белсли не указал, и поэтому проблема определения мультиколлинеарных факторов остается до конца не решенной.

Главные компоненты мультиколлинеарной пары факторов

Выше обсуждался метод устранения эффекта мультиколлинеарности, основанный на замене всех факторов главными компонентами. С помощью этого метода удается решить только проблему вырожденности матрицы нормальных уравнений. Вопрос же построения модели с правильно интерпретируемым содержательным смыслом, как правило, в полном объеме остается нерешенным. В данном разделе предлагается главными компонентами заменять не весь набор факторов регрессионной модели, а только пары тесно связанных между собой факторов. Что это дает? Рассмотрим несколько ситуаций. Будем предполагать, что оба фактора x_1 и x_2 достаточно тесно связаны с моделируемым показателем y, причем теснота этих связей положительна и подчиняется следующей логике: $r_{yx1} > r_{x1x2} > r_{yx2}$. Другими словами, теснота связи между самими факторами выше, чем теснота связи одного из них с моделируемым показателем. В рассматриваемом случае мы считаем, что теснота связи x_2 с y ниже, чем теснота связи между факторами. Как правило, в этом случае, несмотря на положительную взаимосвязь обеих факторов с моделируемым показателем, коэффициент регрессии при факторе x_2 оказывается в силу эффекта мультиколлинеарности отрицательным, то есть не соответствующим содержательной интерпретации.

В рассматриваемом случае эти факторы предлагается заменить линейной комбинацией, в качестве которой использовать главную компоненту, отвечающую максимальному собственному значению ковариационной матрицы. В силу наших предположений о положительности коэффициента корреляции между факторами, собственный вектор ковариационной матрицы

$$\mathbf{C} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}, \qquad (5)$$

соответствующий максимальному собственному значению, имеет положительные компоненты, а сама главная компонента записывается следующим образом

$$z_1 = \gamma_{11}(x_1 - \bar{x}_1) + \gamma_{12}(x_2 - \bar{x}_2). \qquad (6)$$

Следовательно, если, используя эту главную компоненту в качестве фактора, построить регрессионную зависимость, то получим

$$y = \hat{b}_g z = \hat{b}_g[\gamma_1(x_1 - \bar{x}_1) + \gamma_2(x_2 - \bar{x}_2)] = \hat{b}_0 + \hat{b}_1 x_1 + \hat{b}_2 x_2, \qquad (7)$$

где $\hat{b}_1 = \hat{b}_g \gamma_1$; $\hat{b}_2 = \hat{b}_g \gamma_2$.

В силу построения оба коэффициента $\hat{b}_1$ и $\hat{b}_2$ положительны и, следовательно, основная цель – содержательная интерпретируемость модели достигнута, но при этом нужно отметить, что реализация этого подхода связана с некоторой потерей информации. Какие потери и за счет чего они происходят удобно рассмотреть на примере построения регрессионной модели в стандартизованном масштабе. В этом случае вместо ковариационной матрицы следует рассматривать корреляционную матрицу

$$\mathbf{A} = \begin{pmatrix} 1 & r_{12} \\ r_{12} & 1 \end{pmatrix}. \qquad (8)$$

Собственные значения этой матрицы, которые являются дисперсиями главных компонент в стандартизованном масштабе, соответственно равны $\lambda_{\max} = 1 + r_{12}$ и $\lambda_{\min} = 1 - r_{12}$. Учитывая то, что след матрицы инвариантен, в соответствии с которой сумма дисперсий факторов равна сумме дисперсий главных компонент, можно определить процент вариационных потерь факторного пространства в виде отношения

$$\delta_x = (1 - r_{12}) / 2 \cdot 100\% . \qquad (9)$$

По аналогии можно оценить потери в объяснении вариации моделируемого показателя, которые происходят при замене двух мультиколлинеарных факторов главной компонентой. Величина этих потерь может быть рассчитана с помощью коэффициентов детерминации по следующей формуле

$$\delta_y = \left(1 - R_z^2 / R_{x_1 x_2}^2\right) \cdot 100\% . \qquad (10)$$

Рассмотренная ситуация является самым простым примером, хорошо иллюстрирующим основную идею устранения нежелательных эффектов мультиколлинеарности с помощью главной компоненты двух факторов. В том случае, когда один из двух факторов отрицательно коррелирует с моделируемым показателем и, естественно, между факторами отрицательная корреляция, то методика остается той же самой. Только в этом случае один из коэффициентов главной компоненты окажется отрицательным, что позволит в регрессионной модели получить коэффициенты с теми знаками, которые соответствуют экономической интерпретации.

Более сложные ситуации с эффектами мультиколлинеарности могут иметь место, когда в модель включается более двух факторов. В этом случае построение модели, свободной от искажающих содержательный смысл эффектов, осуществляется в несколько этапов. На первом этапе, при условии, что имеет место частичная мультиколлинеарность, строится регрессионная модель, вычисляется корреляционная матрица и на их основе определяются пары тесно коррелирующих между собой факторов. Каждая такая пара заменяется главной компонентой, в результате чего получаем множество регрессоров модели, состоящих из факторов и главных компонент. Например:

$$z_{12}, \quad x_3, \quad x_4, \quad z_{56}, \quad x_7, \quad z_{89},$$

где $z_{12} = \gamma_1 x_1 + \gamma_2 x_2; \quad z_{56} = \gamma_5 x_5 + \gamma_6 x_6; \quad z_{89} = \gamma_8 x_8 + \gamma_9 x_9$.

Оцененная регрессионная модель для этого примера может быть записана с помощью понятных обозначений следующим образом

$$y = \hat{b}_0 + \hat{b}_{12} z_{12} + \hat{b}_3 x_3 + \hat{b}_4 x_4 + \hat{b}_{56} z_{56} + \hat{b}_7 x_7 + \hat{b}_{89} z_{89} . \qquad (11)$$

Не уточняя пока детали этой модели, сделаем следующее предположение. В оцененной преобразованной модели продолжают наблюдаться эффекты мультиколлинеарности, искажающие содержательный смысл коэффициентов регрессии. Пусть, например, удалось выяснить, что причиной этих эффектов являются следующие пары регрессоров z_{12}, z_{56} и x_4, z_{89}. Тогда, следуя общему принципу, для данных регрессоров с учетом новой последовательной нумерации строятся главные компоненты второй очереди

$$u_{14} = \gamma_{12} z_{12} + \gamma_{56} z_{56}; \; u_{36} = \gamma_4 x_4 + \gamma_{89} z_{89} .$$

Таким образом, регрессионная модель строится на следующие регрессоры u_{14}, x_3, u_{36}, x_7. В результате получаем

$$y = \hat{b}_0 + \hat{d}_{14}u_{14} + \hat{b}_3 x_3 + \hat{d}_{36}u_{36} + \hat{b}_7 x_7 =$$
$$= \hat{b}_0 + \hat{d}_{14}\gamma_{12}z_{12} + \hat{d}_{14}\gamma_{56}z_{56} + \hat{b}_3 x_3 + \hat{d}_{36}\gamma_4 x_4 + \hat{d}_{36}\gamma_{89}z_{89} + \hat{b}_7 x_7 =$$
$$= \hat{b}_0 + \hat{d}_{12}z_{12} + \hat{d}_{56}z_{56} + \hat{b}_3 x_3 + \hat{b}_4 x_4 + \hat{d}_{89}z_{89} + \hat{b}_7 x_7, \quad (12)$$

где $\hat{d}_{12} = \hat{d}_{14}\gamma_{12}$; $\hat{d}_{56} = \hat{d}_{14}\gamma_{56}$; $\hat{b}_3 = \hat{b}_3$; $\hat{b}_4 = d_{36}\gamma_4$; $\hat{d}_{89} = d_{36}\gamma_{89}$; $\hat{b}_7 = \hat{b}_7$.

Подставив в полученную модель выражения для компонент z_{12}, z_{56} и z_{89} можем записать следующее уравнение

$$y = \hat{b}_0 + \hat{d}_{12}z_{12} + \hat{d}_{56}z_{56} + \hat{b}_3 x_3 + \hat{b}_4 x_4 + \hat{d}_{89}z_{89} + \hat{b}_7 x_7 =$$
$$= \hat{b}_0 + \hat{d}_{12}(\gamma_1 x_1 + \gamma_2 x_2) + \hat{b}_3 x_3 + \hat{b}_4 x_4 + \hat{d}_{56}(\gamma_5 x_5 + \gamma_6 x_6) + \hat{b}_7 x_7 + \hat{d}_{89}(\gamma_8 x_8 + \gamma_9 x_9),$$

из которого без труда получается финальный результат

$$y = \hat{b}_0 + \hat{b}_1 x_1 + \hat{b}_2 x_2 + \hat{b}_3 x_3 + \hat{b}_4 x_4 + \hat{b}_5 x_5 + \hat{b}_6 x_6 + \hat{b}_7 x_7 + \hat{b}_8 x_8 + \hat{b}_9 x_9, \quad (13)$$

где

$\hat{b}_1 = \hat{d}_{12}\gamma_1 = \hat{d}_{14}\gamma_{12}\gamma_1$; $\hat{b}_2 = \hat{d}_{12}\gamma_2 = \hat{d}_{14}\gamma_{12}\gamma_2$; $\hat{b}_3 = \hat{b}_3$; $\hat{b}_4 = \hat{d}_{36}\gamma_4$; $\hat{b}_5 = \hat{d}_{56}\gamma_5 = \hat{d}_{14}\gamma_{56}\gamma_5$;
$\hat{b}_6 = \hat{d}_{56}\gamma_6 = \hat{d}_{14}\gamma_{56}\gamma_6$; $\hat{b}_7 = \hat{b}_7$; $\hat{b}_8 = \hat{d}_{89}\gamma_8 = \hat{d}_{36}\gamma_{89}\gamma_8$; $\hat{b}_9 = \hat{d}_{89}\gamma_9 = \hat{d}_{36}\gamma_{89}\gamma_9$.

Процедура попарной замены высоко коррелированных факторов главными компонентами обеспечивает получение содержательно интерпретируемой регрессионной модели. Нужно только оценить потери точности, которые на каждом этапе построения модели имели место. Для этой цели можно использовать критерий (9).

Эмпирические исследования

Апробацию предложенной методики рассмотрим на конкретном примере построения регрессионной зависимости объема валового регионального продукта Воронежской области y (млн руб.) от шести факторов: x_1, x_2, x_3, x_4, x_5, x_6, отражающих изменения показателей: среднегодовая численность занятых в экономике (тыс. чел.), среднегодовые денежные доходы в месяц (руб.), потребительские расходы в среднем на душу населения в месяц (руб.), среднемесячная номинальная заработная плата работников организаций (руб.), основные фонды в экономике на конец года (млн руб.), объем отгруженных товаров собственного производства, выполненных работ и услуг собственными силами по видам экономической деятельности (млн руб.) за период 2004-2016 гг. (см. табл. 1).

Таблица 1

Значения показателя y и факторов, влияющих на его изменение, за период 2004-2016 гг.

№ п/п	y	x_1	x_2	x_3	x_4	x_5	x_6
1	63217,2	1091,8	2574,3	2517,9	2579,7	314552	54374
2	88151,6	1065,9	3391,2	3351,6	3549,1	375512	66563
3	105053,4	1059,6	4050,1	4062	4340,6	395404	76631
4	116975,9	1055,5	5456,8	3790,8	5382,2	428766	94013
5	136152,7	1057,2	7020,2	4463,6	6750,3	469878	112240
6	163246,3	1062	8530,3	5596,9	8730,9	562523	132807

Окончание табл. 1

№ п/п	y	x_1	x_2	x_3	x_4	x_5	x_6
7	228666,4	1064,7	10304,8	7168,6	11490	674657	169702
8	289322,3	1055,3	11727,9	8171,2	12786,1	738634	169819
9	302510,1	1054,3	13580	9822	14337,3	788059	207002
10	328770,8	1054,9	15870,9	12190,4	16054,7	1019463	273998
11	447155,4	1057,9	18885,1	14809,8	19538,1	1078383	302902
12	568613	1057	22056	17006	21825	1158136	322179
13	606667,7	1055,3	25505,3	19327,9	24001	1233524	373850

В табл. 2 представлена корреляционная матрица, содержащая коэффициенты взаимосвязи каждого из шести факторов с зависимой переменной y и взаимосвязи факторов между собой.

Таблица 2

	y	x_1	x_2	x_3	x_4	x_5	x_6
y	1						
x_1	-0,47107	1					
x_2	0,991193	-0,50489	1				
x_3	0,992375	-0,46728	0,993339	1			
x_4	0,986565	-0,51612	0,995911	0,987053	1		
x_5	0,975409	-0,50899	0,988764	0,986914	0,992505	1	
x_6	0,978076	-0,49804	0,993908	0,99182	0,991851	0,995831	1

Уравнение регрессии, построенное с помощью пакета «Анализ данных», имеет следующий вид:

$$y= -186273+169{,}0574\,x_1+17{,}37926\,x_2+31{,}29262\,x_3+$$
$$+8{,}675576\,x_4+0{,}090798\,x_5-2{,}00316\,x_6 \qquad (14)$$

Анализ коэффициентов регрессионного уравнения и коэффициентов корреляции показывает, что фактор x_1 отрицательно коррелирует с моделируемым показателем, в то же время коэффициент регрессии при этом факторе является положительным, что свидетельствует об эффекте мультиколлениарности фактора x_1 с другими факторами. Наибольшая взаимосвязь фактора x_1 наблюдается с фактором x_4 (r_{x1x4}=–0,51612); это самое большое по модулю значение коэффициента корреляции в сравнении с

$$r_{x1x2},\ r_{x1x3},\ r_{x1x5},\ r_{x1x6}.$$

Аналогичное искажение наблюдается и при сопоставлении коэффициента регрессии при переменной x_6 (значение его отрицательно, $\hat{b}_6=-2{,}003$) с коэффициентом корреляции этого фактора с зависимым показателем (r_{yx6}= 0.978076).

Переменная x_6 очень тесно коррелирует с переменной x_5, причем наблюдается следующее соотношение: $r_{yx6} < r_{x6x5}$, что также говорит об эффекте мультиколлениарности, который внес искажение в значение коэффициента регрессии $\hat{b}_6$. $\hat{b}_6$ = – 2,003 не поддается содержательной интерпретации.

Следующим шагом будет замена факторов x_1 и x_4, x_5 и x_6 главными компонентами:

$$z_1=\eta_1 x_1+\eta_2 x_4,\ z_2 = \mu_1 x_5+\mu_2 x_6.$$

Коэффициенты главных компонент (η_1, η_2) и (μ_1, μ_2) соответствуют собственным векторам, отвечающим максимальным собственным значениям ковариационных матриц

$$\begin{pmatrix} \sigma^2_{x_1} & \sigma_{x_1x_4} \\ \sigma_{x_4x_1} & \sigma^2_{x_4} \end{pmatrix} \text{ и } \begin{pmatrix} \sigma^2_{x_5} & \sigma_{x_5x_6} \\ \sigma_{x_6x_5} & \sigma^2_{x_6} \end{pmatrix},$$

а именно:

$$\begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} = \begin{pmatrix} -0,00072 \\ 1 \end{pmatrix}; \quad \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} = \begin{pmatrix} 0,949551 \\ 0,313612 \end{pmatrix}.$$

Таким образом, главная компонента $z_1=-0,00072x_2+1x_4$; главная компонента $z_2=0,949551x_5+0,313612x_6$.

Следующим шагом является построение регрессионной модели, в которой в качестве регрессоров выступают: z_1, x_2, x_3, z_2. Табл. 3 содержит данные, по которым строится регрессия.

Таблица 3

Значения показателя y и факторов, представленных главными компонентами z_1, z_2 и переменными x_2, x_3.

№ п/п	y	z_1	x_2	x_3	z_2
1	63217,2	2578,913	2574,3	2517,9	315735,6
2	88151,6	3548,331	3391,2	3351,6	377442,9
3	105053,4	4339,836	4050,1	4062	399488,8
4	116975,9	5381,438	5456,8	3790,8	436618,9
5	136152,7	6749,537	7020,2	4463,6	481373,1
6	163246,3	8730,133	8530,3	5596,9	575794,3
7	228666,4	11489,23	10304,8	7168,6	693842
8	289322,3	12785,34	11727,9	8171,2	754628,1
9	302510,1	14336,54	13580	9822	813220,7
10	328770,8	16053,94	15870,9	12190,4	1053961
11	447155,4	19537,33	18885,1	14809,8	1118974
12	568613	21824,23	22056	17006	1200749
13	606667,7	24000,23	25505,3	19327,9	1288538

Результат построения регрессионной модели по этим данным представлен следующей формулой:

$$y=61889,73+22,76845z_1 - 3,40185x_2+33x_3 - 0,42108z_2 \qquad (15)$$

В табл. 4 приводятся коэффициенты корреляции между моделируемым показателем y и регрессорами z_1, x_2, x_3, z_2, а также коэффициенты корреляции между самими регрессорами.

Таблица 4

	y	z_1	x_2	x_3	z_2
y	1				
z_1	0,9865651	1			
x_2	0,9911931	0,995911	1		
x_3	0,9923746	0,987053	0,993339	1	
z_2	0,9760341	0,992809	0,989639	0,987764	1

Анализ коэффициентов регрессии, среди которых есть отрицательные ($\hat{b}_{x_2}$ и $\hat{b}_{z_2}$), и анализ корреляционной матрицы ($r_{yx_2} < r_{z_1x_2}$; $r_{yz_2} < r_{z_2x_3}$) снова свидетельствуют об эффекте мультиколлениарности, для устранения которого будет осуществлена замена факторов z_1 и x_2, x_3 и z_2, главными компонентами: $u_1 = \alpha_1 z_1 + \alpha_2 x_2$; $u_2 = \beta_1 x_3 + \beta_2 z_2$.

Коэффициенты главных компонент $(\alpha_1, \alpha_2) u (\beta_1, \beta_2)$ – это координаты собственных векторов, отвечающих максимальным собственным значениям, ковариационных матриц

$$\begin{pmatrix} \sigma_{z_1}^2 & \sigma_{z_1 x_2} \\ \sigma_{x_2 z_1} & \sigma_{x_2}^2 \end{pmatrix} \text{ и } \begin{pmatrix} \sigma_{x_3}^2 & \sigma_{x_3 z_2} \\ \sigma_{z_2 x_3} & \sigma_{z_2}^2 \end{pmatrix},$$

$$\begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} = \begin{pmatrix} 0,69757 \\ 0,716517 \end{pmatrix}; \quad \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} 0,016314 \\ 0,999867 \end{pmatrix}.$$

Таким образом, главная компонента $u_1 = 0,69757 z_1 + 0,716517 x_2$, главная компонента $u_2 = 0,016314 x_3 + 0,999867 z_2$.

Далее строится регрессивная модель между показателем у и регрессорами u_1 и u_2.

Данные, используемые для построения регрессионного уравнения, отражены в табл. 5.

Таблица 5

Значения показателя у и факторов, представленных главными компонентами u_1, u_2

№ п/п	y	u_1	u_2
1	63217,2	3643,501	315734,6
2	88151,6	4905,061	377447,3
3	105053,4	5929,303	399501,9
4	116975,9	7663,818	436622,6
5	136152,7	9738,364	481381,8
6	163246,3	12201,98	575809
7	228666,4	15398,1	693866,6
8	289322,3	17321,9	754661
9	302510,1	19731,03	813272,7
10	328770,8	22570,51	1054020
11	447155,4	27160,15	1119066
12	568613	31027,42	1200866
13	606667,7	35016,82	1288682

Результат построения регрессионной модели представлен формулой:

$$y=21865,66+24,45231u_1 - 0,21356u_2 \tag{16}$$

В табл. 6 – корреляционная матрица, в ячейках которой коэффициенты корреляции между u_1 и u_2, а также между y и u_1, y и u_2.

Таблица 6

	y	u_1	u_2
y	1		
u_1	0,989953	1	
u_2	0,976042	0,992198	1

Наблюдается искажение коэффициента регрессии $\hat{b}_{u_2}$ (он имеет отрицательный знак, в то время как $r_{yu_2} > 0$). Кроме того, выполняется соотношение $r_{yu_2} < r_{u_1 u_2}$, т.е. имеет место эффект мультиколлениарности, который может быть устранен через построение главной компоненты между u_1 и u_2:

$$v = \gamma_1 u_1 + \gamma_2 u_2 ,$$

где $(\gamma_1\ ,\gamma_2)$ – координаты собственного вектора, отвечающего максимальному собственному значению ковариационной матрицы

$$\begin{pmatrix} \sigma_{u_1}^2 & \sigma_{u_1u_2} \\ \sigma_{u_2u_1} & \sigma_{u_2}^2 \end{pmatrix}; \begin{pmatrix} \gamma_1 \\ \gamma_2 \end{pmatrix} = \begin{pmatrix} 0,030133 \\ 0,999546 \end{pmatrix}.$$

Таким образом, $v=0,030133u_1+0,999546u_2$.

На завершающем шаге строится регрессионная модель между моделируемым показателем *y* и главной компонентой *v*, значения которых отражены в табл. 7.

Таблица 7

Значения показателя *y* и главной компоненты *v*

№ п/п	y	v
1	63217,2	315701,1
2	88151,6	377423,7
3	105053,4	399499,1
4	116975,9	436655,3
5	136152,7	481456,7
6	163246,3	575915,2
7	228666,4	694015,5
8	289322,3	754840,3
9	302510,1	813498
10	328770,8	1054222
11	447155,4	1119377
12	568613	1201256
13	606667,7	1289152

Результат построения регрессионной модели представлен уравнением:

$$y = -118023 + 0,523368v. \qquad (17)$$

На заключительном этапе остается перейти к первоначальным переменным, подставив в полученную модель выражение для главной компоненты v, затем последовательно выражения для главных компонент u_1 и u_2, и наконец, выражения для главных компонент z_1 и z_2:

$$y = -118023+ 0,523368v = -118023+0,523368\cdot(0,030133u_1 + 0,999546u_2)=$$
$$= -118023+ 0,523368\cdot(0,030133\cdot(0,69757z_1 + 0,716517x_2) +$$
$$+ 0,999546(0,16314x_3 + 0,999867z_2)) = -118023+ 0,523368\cdot$$
$$\cdot(0,030133\cdot(0,69757\cdot(-0,00072x_1 + 1\cdot x_4) + 0,716517x_2) +$$
$$0,999546(0,016314x_3+ 0,999867\cdot(0,949551x_5+ 0,313612x_6))).$$

Итоговый результат имеет следующий вид:

$$y= -118023 - 0,00000792x_1 + 0,011299937x_2 + 0,008534349x_3 +$$
$$+ 0,01100113x_4 + 0,496672919x_5 + 0,164038148x_6. \qquad (18)$$

Анализ построенной модели свидетельствует о том, что нам удалось получить содержательно интерпретируемые оценки коэффициентов регрессии в условиях мультиколлинеарности, сохранив при этом все объясняющие переменные.

Заключение

Эмпирические исследования подтвердили эффективность применения предлагаемой методики для построения линейных регрессионных моделей в условиях мультиколлинеарности. В отличие от известных методик с помощью этой удается решить все проблемы частичной мультиколлинеарности, сохраняя первоначальный состав факторов и обеспечивая полную их согласованность с теоретическими представлениями о моделируемых процессах. Оригинальность и полезность методики очевидна, но ее реализация достаточно сложная процедура. Поэтому возникает вопрос о возможностях программной реализации этого подхода. Такая возможность есть. Ее реализация требует формализованных критериев. Проведенные эмпирические исследования продемонстрировали логику вычислений, основанную на формализованных критериях.

Список источников

1. Belsley D.A. *Conditioning Diagnostics, Collinearity and Weak Data Regression*. New York, Wiley, 1991.

2. Быстров О.Ф., Тарасов Д.Э. Организация и содержательная интерпретация вычислительного эксперимента с моделью логистической системы // *Транспортное строительство*, 2016, no. 8, с. 25-27.

3. Дрейпер Н., Смит Г. *Прикладной регрессионный анализ*, 3-е изд. Москва, Издательский дом «Вильямс», 2007.

4. Кузнецова И.С., Белозерских А.В. Проблемы мультиколлинеарности в регрессионных моделях // *Актуальные направления научных исследований XXI века: теория и практика*, 2015, т. 3, no. 8-3 (19-3), с. 308-312.

5. Луценко Е.В. Принципы и перспективы корректной содержательной интерпретации субъективных (виртуальных) моделей физической и социальной реальности, формируемых сознанием человека // *Политематический сетевой электронный научный журнал Кубанского государственного аграрного университета*, 2016, no. 115, с. 22-75.

6. Марон А.Е., Монахова Л.Ю. Структурно-содержательная модель программно-технологического сопровождения образования взрослых // *Человек и образование*, 2015, no. 1 (42), с. 34-37.

7. Орлова И.В., Филонова Е.С. Выбор экзогенных факторов в модель регрессии при мультиколлинеарности данных // *Международный журнал прикладных и фундаментальных исследований*, 2015, no. 5-1, с. 108-116.

8. Потанина М.В. Свойства среднеквадратической ошибки прогноза ридж-регрессии для идентификации модели сложного объекта управления при мультиколлинеарности факторов // *Вестник СевНТУ*, 2014, no. 146, с. 113-117.

9. Роберт И.В., Власов В.В. Аналитический отчет «теоретическая модель информационно-методического обеспечения процесса освоения методов решения задач (содержательный блок базы знаний)» // *Хроники объединенного фонда электронных ресурсов. Наука и образование*, 2015, no. 8 (75), с. 159.

10. Утемисова Ж.Ж., Гончарова К.В., Подповетная Ю.В. Мультиколлинеарность экономических показателей и методы ее устранения // *Современное бизнес-пространство: актуальные проблемы и перспективы*, 2014, no. 2 (3), с. 30-32.

THE CONSTRUCTION METHOD OF MEANINGFUL INTERPRETED REGRESSION MODELS IN CONDITIONS OF MULTICOLLINEARITY

Mokshina Svetlana Ivanovna, Cand. Sc. (Econ.), Assoc. Prof.
Shurshikova Galina Vladimirovna, Cand. Sc. (Eng.), Assoc. Prof.
Shchekunskih Svetlana Stanislavovna, Cand. Sc. (Phys.-Math.), Assoc. Prof.

Voronezh State University, University sq., 1, Voronezh, Russia, 394018; e-mail: mokshina@econ.vsu.ru; shurshikova@econ.vsu.ru; shchekunskikh@econ.vsu.ru

Purpose: the authors build a linear regression models without the effects multicollinearity, because they distort the meaning. *Discussion*: the issue of econometric modeling in a situation of multicollinearity continues to be relevant despite the set of recommendations. Almost of all recommendations solve problems of computational nature. The matrix degeneracy of normal equations is the reason of these problems. The authors notice that, the scientists decided the computational problem, but did not decide the meaningful problem of econometric model. *Results*: the authors offer an original method for constructing linear regression models with effects of multicollinearity. The writers conducted the empirical studies. They confirmed the effectiveness of the proposed method.

Keywords: econometric modeling, principal components, multicollinearity.

References

1. Belsley D.A. *Conditioning Diagnostics, Collinearity and Weak Data Regression*. New York, Wiley, 1991.

2. Bystrov O.F., Tarasov D.E. Organizatsiia i soderzhatel'naia interpretatsiia vychislitel'nogo eksperimenta s model'iu logisticheskoi sistemy organization and meaningful interpretation of computer simulation model of the logistics system [Organization and meaningful interpretation of computer experiment with the model of the logistics system]. *Transportnoe stroitel'stvo*, 2016, no. 8, pp. 25-27. (In Russ.)

3. Dreiper N., Smit G. *Prikladnoi regressionnyi analiz* [Applied regression analysis], 3-e izd. Moscow, Izdatel'skii dom «Vil'iams», 2007. (In Russ.)

4. Kuznetsova I.S., Belozerskikh A.V. Problemy mul'tikollinearnosti v regressionnykh modeliakh [Multicollinearity problems in regression models]. *Aktual'nye napravleniia nauchnykh issledovanii XXI veka: teoriia i praktika,* 2015, vol. 3, no. 8-3 (19-3), pp. 308-312. (In Russ.)

5. Lutsenko E.V. Printsipy i perspektivy korrektnoi soderzhatel'noi interpretatsii sub"ektivnykh (virtual'nykh) modelei fizicheskoi i sotsial'noi real'nosti, formiruemykh soznaniem cheloveka. [Principles and perspectives of correct meaningful interpretation of the subjective (virtual) models in the physical and social reality formed by the human consciousness] *Politematicheskii setevoi elektronnyi nauchnyi zhurnal Kubanskogo gosudarstvennogo agrarnogo universiteta*, 2016, no. 115, pp. 22-75. (In Russ.)

6. Maron A.E., Monakhova L.Iu. Strukturno-soderzhatel'naia model' program-

mno-tekhnologicheskogo soprovozhdeniia obrazovaniia vzroslykh [Structurally-substantial model of software and technological support for the adults education]. *Chelovek i obrazovanie*, 2015, no. 1 (42), pp. 34-37. (In Russ.)

7. Orlova I.V., Filonova E.S. Vybor ekzogennykh faktorov v model' regressii pri mul'tikollinearnosti dannykh [The choice of exogenous factors in the regression model in case of multicollinearity data]. *Mezhdunarodnyi zhurnal prikladnykh i fundamental'nykh issledovanii*, 2015, no. 5-1, pp. 108-116. (In Russ.)

8. Potanina M.V. Svoistva srednekvadraticheskoi oshibki prognoza ridzh-regressii dlia identifikatsii modeli slozhnogo ob"ekta upravleniia pri mul'tikollinearnosti faktorov [Properties of predict square error in ridge regression for identify the model of complex object management with multicollinearity of the factors]. *Vestnik SevNTU*, 2014, no. 146, pp. 113-117. (In Russ.)

9. Robert I.V., Vlasov V.V. Analiticheskii otchet «teoreticheskaia model' informatsionno-metodicheskogo obespecheniia protsessa osvoeniia metodov resheniia zadach (soderzhatel'nyi blok bazy znanii)» [Analytical report «theoretical model of information-methodical maintenance of development process in methods for solving problems (the meaningful unit of knowledge)»]. *Khroniki ob"edinennogo fonda elektronnykh resursov Nauka i obrazovanie*, 2015, no. 8 (75), pp. 159. (In Russ.)

10. Utemisova Zh.Zh., Goncharova K.V., Podpovetnaia Iu.V. Mul'tikollinearnost' ekonomicheskikh pokazatelei i metody ee ustraneniia [Multicollinearity of the economic indicators and the methods of its elimination]. *Sovremennoe biznes-prostranstvo: aktual'nye problemy i perspektivy,* 2014, no. 2 (3), pp. 30-32. (In Russ.)