DOI:

УМЕНЬШЕНИЕ МУЛЬТИКОЛЛИНЕАРНОСТИ С ПОМОЩЬЮ ПРЕОБРАЗОВАНИЯ ПЕРЕМЕННЫХ

Орлова И.В.

Финансовый университет при Правительстве РФ, Россия, г. Москва, Ленинградский пр.,49

ivorlova@fa.ru

Аннотация:В работе рассматривается метод частичной ортогонализации регрессоров, основанный на линейном преобразовании исходных переменных, внешне схожем с первым шагом алгоритма ортогонализации Грама-Шмидта. Процесс ортогонализации организуется таким образом, чтобы коэффициенты регрессии имели содержательную интерпретацию, адекватно отражающую реальные зависимости между переменными. Свойства рассматриваемого метода иллюстрируются на примере.

Ключевые слова: регрессия, мультиколлинеарность, ортогонализация переменных, программная среда R

Введение

Одним из условий корректного применения регрессионного анализа для выявления зависимостей в эмпирических данных является отсутствие мультиколлинеарности регрессоров. Хорошо известны негативные последствия, к которым приводит мультиколлинеарность [1, 2, 3]. Поэтому диагностирование мультиколлинеарности является неотъемлемой составляющей частью регрессионного анализа. Для принятия обоснованного решения о мультиколлинеарности регрессоров приходится пользоваться набором тестов. Достаточно полно тесты на наличие мультиколлинеарности переменных представлены в пакете программ R. В случае наличия мультиколлинеарности в данных возникает необходимость либо избавиться от нее, либо хотя бы ослабить степень мультиколлинеарности регрессоров [2, 3]. Существует множество методов ослабления степени или преодоления негативных последствий мультиколлинеарности. Все они обладают теми или иными недостатками и поэтому разработка методов ослабления степени мультиколлинеарности является актуальной. В работе [4] предлагается подход к уменьшению степени мультиколлинеарности с помощью преобразования переменных, направленного на то, чтобы коэффициенты регрессии либо были равны коэффициентам регрессии по исходным переменным, либо равны коэффициентам регрессии по новым переменным, имеющим содержательную интерпретацию. Рассматриваемый метод неполной ортогонализации переменных позволяет существенно снизить степень мультиколлинеарности и получить при этом уравнение регрессии, адекватно отражающее анализируемый процесс.

В случае, когда исходные регрессоры являются пространственными переменными, множество регрессоров разбивается на непересекающиеся группы сильно коррелированных между собой предикторов. Предикторы, не имеющие высокой корреляционной связи ни с одним регрессором, образуют группу из одного элемента. Далее в каждой группе выбирается один предиктор X_k, называемый далее “выбранным“, и затем строятся вспомогательные уравнения регрессии остальных регрессоров X_j группы на выбранный регрессор X_k. Коэффициенты вспомогательных регрессий будем вычислять с помощью метода наименьших квадратов. Никаких допущений относительно распределения остатков от регрессии и их взаимосвязей с другими остатками и регрессорами не делается. Остатки от регрессии X_j на “выбранный“ регрессор обозначаем через U_j. В дальнейшем в регрессии эндогенной переменной Y на объясняющие переменные вместо X_j участвуют новые переменные U_j. Переменные U_j равны разности между заменяемой переменной и предсказанными по уравнению регрессии значениями этой переменной. Таким образом, U_j являются линейной комбинацией исходных регрессоров X_j. Пусть $x_j^{(i)} = a_0^j + a_k^j x_k^{(i)} + u_j^{(i)}$, тогда $u_j^{(i)} = x_j^{(i)} - a_0^j - a_k^j x_k^{(i)}$, где i – номер наблюдения. Переменные X_k и U_j не коррелированы, в то время как X_k и X_j могут быть как угодно коррелированы. Таким образом, заменяя часть регрессоров X_j на U_j, мы избавляемся от некоторых корреляционных зависимостей и тем самым уменьшаем уровень мультиколлинеарности участвующих в регрессии переменных. Подставляя представление X_j через X_k и U_j в уравнение регрессии по исходным регрессорам $Y = b_0 + \mathrm{b}_1 X_1 + \cdots + \mathrm{b}_m X_m + e,$ замечаем, что коэффициенты регрессии $\mathrm{b_j}$ при всех X_j , кроме тех j, для которых X_j являются “выбранными“, равны коэффициентам регрессии g_j по новым регрессорам U_j. Коэффициенты регрессии g_k при таких k, для которых X_k являются “выбранными“, равны $\mathrm{g_k}$=$b_k + \sum_j b_j a_k^j$. Суммирование ведётся по всем таким j, для которых

X_j является зависимой переменной, а X_k - независимой в дополнительных регрессиях. Из последней формулы получаем, что коэффициенты g_k регрессии Y по новым переменным имеют содержательную интерпретацию. При увеличении X_k на единицу Y меняется на b_k за счёт изменения X_k, но, кроме этого, при неизменных корреляционных связях, регрессоры X_j, входящие в одну группу с X_k, изменятся в среднем на a_k^j . Это влечёт за собой изменение Y на $\sum_j b_j a_k^j$. Как видим, изменение Y равно g_k. Таким образом, коэффициент регрессии g_k можно трактовать как приращение Y при изменении X_k на единицу, учитывающее соответствующие изменения тех X_j, которые участвовали в дополнительных регрессиях X_j на X_k, то есть учитывающее корреляционные связи X_k с другими регрессорами.

Объясняющие переменные X_j, которые не выступали в роли зависимых переменных во вспомогательных регрессиях и не были заменены на U_j, для удобства записи будем обозначать как U_j . Обозначим через **X** и **U** матрицы значений исходных регрессоров $X_1, X_2, \dots, X_m$ и новых регрессоров $U_1, U_2, \dots, U_m$. Матрицы **X** и **U** имеют размерность $(m+1) \times n$, где n – число наблюдений, первый столбец матриц **X** и **U** состоит из единиц. Через $\boldsymbol{Y}$ обозначим вектор значений объясняющих переменных, через **b, g** - векторы коэффициентов регрессии Y на $X_1, X_2, \dots, X_m$ и $U_1, U_2, \dots, U_m$ соответственно. Обозначим через $\boldsymbol{B}$ матрицу преобразования переменных X_j. Тогда $\boldsymbol{U=X \cdot B.}$ Диагональные элементы матрицы $\boldsymbol{B}$ равны 1, остальные элементы равны 0, кроме тех столбцов j, для которых X_j выступают в роли независимых переменных в дополнительных регрессиях. Эти столбцы заполняются в соответствии с формулой перехода от X_j к U_j. Матрица $\boldsymbol{B}$ является невырожденной, следовательно, линейное преобразование, задаваемое матрицей $\boldsymbol{B,}$ является взаимно-однозначным. Отсюда получаем формулу, отражающую взаимосвязь коэффициентов регрессии $\boldsymbol{b= B \cdot g.}$ Из последней формулы вытекает, что ковариационные матрицы оценок коэффициентов регрессии по исходным и новым переменным связаны соотношением: $cov(\boldsymbol{b}) = \boldsymbol{B} \cdot cov(\boldsymbol{g}) \cdot \boldsymbol{B}'$.

Проиллюстрируем рассматриваемый метод моделирования пространственных переменных на примере построения регрессионной модели объема инновационных товаров, работ, услуг по субъектам Российской Федерации, Y (млн. руб.). В качестве объясняющих переменных выбраны: X_1−валовой региональный продукт по субъектам Российской Федерации, млн руб.; X_2−численность рабочей силы в возрасте 15 лет и старше, тыс. чел.; X_3−среднемесячная заработная плата по субъектам Российской Федерации за 2018 год, рублей; X_4−капитальные затраты на научно-исследовательские и опытно-конструкторские работы, млн. руб.; X_5− инновационная активность организаций (удельный вес организаций, осуществлявших технологические, организационные, маркетинговые инновации в отчетном году, в общем числе обследованных организаций), %; X_6− численность исследователей, имеющих ученую степень, чел., [5], [6], [7].

Все расчеты выполнялись в свободно распространяемой среде R.

Коэффициенты корреляции Y со всеми регрессорами, кроме X_3 (заработная плата) значимо отличаются от нуля.

Тестирование мультиколлинеарности можно выполнить с помощью пакета `mctest` [8, 9] для диагностики общей и индивидуальной мультиколлинеарности данных. Четыре показателя – X_1, X_2, X_4, X_6 образуют группу сильно коррелированных между собой показателей, все коэффициенты корреляции между ними больше 0,8. Это свидетельствует о мультиколлинеарности регрессоров.

Группа тестов функции `omcdiag()` пакета `mctest` направлены на проверку присутствия мультиколлинеарности во всем наборе данных. Пять тестов из шести показали, что данные избыточны, в них присутствует мультиколлинеарность.

В процессе применения процедуры пошаговой регрессию к модели зависимости Y от $X_1, \dots, X_6$ из модели удаляются факторы X_3 и X_2 Уравнение регрессии имеет вид: $\hat{Y} = -32565 + 0{,}036\, X_1 + 24{,}9\, X_4 + 7679{,}8\, X_5 - 20{,}9\, X_6$. Все коэффициенты регрессии значимы, P-значения всех коэффициентов при переменных не превосходят 0,0002. По уравнению регрессии связь Y с X_6 отрицательная, так как коэффициент регрессии при X_6 (численность исследователей, имеющих ученую степень) отрицателен, в то время как коэффициент корреляции Y с X_6 положителен и равен 0,41. Это является проявлением мультиколлинеарности. Наличие мультиколлинеарности подтверждают результаты тестирования с помощью функции **vif()** из пакета **car.** Факторы инфляции дисперсии VIF_j равны 4,9; 5,0; 1,1; 8,2, что свидетельствует о мультиколлинеарности регрессоров в построенной модели.

Для снижения уровня мультиколлинеарности воспользуемся рассматриваемым в данной работе подходом преобразования переменных. Один из четырёх регрессоров – X_6 назовём “выбранным” и заменим в регрессионной модели X_1, X_2, X_4 на U_1, U_2, U_4 - остатки от регрессии X_1, X_2, X_4 на X_6.

Построим уравнение регрессии Y по U_1 , U_2 , U_4 , X_3 , X_5 , X_6 . Факторы инфляции дисперсии VIF_j равны 1,9; 1,7; 1,1; 1,5; 1,1; 1,2, следовательно, есть основания считать, что мультиколлинеарность отсутствует. После исключения незначимых факторов X_3 и U_2 получаем уравнение регрессии: $\widehat{Y} = -10862 + 0{,}036\,U_1 + 24{,}9\,U_4 + 7679{,}8\,X_5 + 7{,}71\,X_6$. Стандартная ошибка и коэффициент детерминации двух построенных моделей совпадают. Коэффициенты регрессии при U_1 , U_4, X_5 уравнения по новым переменным равны коэффициентам регрессии при X_1, X_4, X_5 в уравнении по исходным переменным. Коэффициенты регрессии при "выбранной" переменной. X_6 в уравнениях отличаются не только величиной, но и знаком. Все коэффициенты регрессии в уравнении по новым переменным значимы, коэффициент регрессии при U_1 имеет Р-значение, равное 0,0002, остальные коэффициенты имеют Р-значения, меньшие 0,0001.

Проверка мультиколлинеарности в новом наборе данных (U_1, U_4, X_5 , X_6) с помощью функции `imcdiag()` пакета **`mctest`** показала полное ее отсутствие. В функции `imcdiag()` реализованы тесты индивидуальной проверки регрессоров на мультиколлинеарность. Результаты выполнения функции `imcdiag()` показали, что ни один из регрессоров U_1, U_4, X_5 , X_6 не может быть причиной мультиколлинеарности. Факторы инфляции дисперсии VIF_j равны 1,01; 1,01; 1,07; 1,06, следовательно, регрессоры модели почти ортогональны и, следовательно, коэффициенты регрессии имеют близкую к минимальной дисперсию.

Изменение X_6 на единицу при неизменных корреляционных связях приводит в среднем к изменению X_1 на величину a_1^6=384 и X_4 на величину a_4^6 = 0,59, а это, в свою очередь, приводит к изменению Y на величину g_6 , равную $g_6 = \mathrm{b}_6 + \mathrm{b}_1 a_1^6 + \mathrm{b}_4 a_4^6$=- 20,9+ 0,036·384+24,9 ·0,59=7,71. Таким образом, в уравнении по новым переменным коэффициент регрессии g_6=7,71 при "выбранной" переменной X_6 учитывает корреляционную связь X_6 с переменными, входящими в одну с ней группу высоко коррелированных переменных. Непосредственной проверкой убеждаемся, что векторы оценок коэффициентов регрессии $\boldsymbol{b}$ и $\boldsymbol{g}$ удовлетворяют соотношению $\boldsymbol{b = B \cdot g}$.

Таким образом, рассматриваемый в работе метод неполной ортогонализации пространственных исходных переменных, предложенный в [4], позволяет получать поддающиеся содержательной интерпретации результаты моделирования. Предложенные в [4] методы предполагается реализовать в программной среде R.

Литература

1. *Дрейпер, Норман; Смит, Гарри* Прикладной регрессионный анализ, [Текст], 3-е изд.: Пер. с англ. – М.: Издательский дом «Вильямс», 2007. — 912 с.
2. *Orlova I., Ioudina V.* Analysis of information content of metric data when constructing models of linear regression //: System analysis in economics - 2018 Proceedings of the V International research and practice conference-biennale. 2018. С. 196-198. DOI: 10.33278/SAE-2018.eng.196-198
3. *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. DOI 10.17513/fr.42103.
4. *Орлова И.В.* Корректировка спецификации модели множественной регрессии при наличии мультиколлинеарности исходных регрессоров В книге: УПРАВЛЕНИЕ РАЗВИТИЕМ КРУПНОМАСШТАБНЫХ СИСТЕМ MLSD'2019 Материалы двенадцатой международной конференции Научное электронное издание. Под общей ред. С.Н. Васильева, А.Д. Цвиркуна. 2019. С. 993-995
5. http://old.gks.ru/wps/wcm/connect/rosstat_main/rosstat/ru/statistics/science_and_innovations/science/ (дата обращения 3.02.2019)
6. https://www.gks.ru/free_doc/new_site/population/trud/tab_trud1.htm (дата обращения 3.02.2019)
7. https://www.gks.ru/labor_market_employment_salaries?print=1 (дата обращения 3.02.2019)
8. *M. I. Ullah, M. Aslam, Saima Altaf* mctest: An R Package for Detection of Collinearity among Regressors // The R Journal (2016) volume 8:2, pages 495-505.
9. *Muhammad Imdad Ullah, Muhammad Aslam* Multicollinearity Diagnostic Measures. Package 'mctest' https: //cran.r-project.org/web/packages/mctest/mctest.pdf (дата обращения 30.11.2019)