УДК 51-77:519.237.5

ПОДХОД К РЕШЕНИЮ ПРОБЛЕМЫ МУЛЬТИКОЛЛИНЕАРНОСТИ С ПОМОЩЬЮ ПРЕОБРАЗОВАНИЯ ПЕРЕМЕННЫХ

Орлова И.В.

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

Построение эконометрических моделей, их анализ и прогнозирование эндогенной переменной Y по значениям экзогенных переменных X_1, X_2, …X_m зачастую затруднено наличием мультиколлинеарности переменных X_1, X_2, … X_m. Мультиколлинеарность приводит к ряду негативных последствий при построении и анализе уравнения регрессии. Это приводит к необходимости тем или иным способом избавиться от неё или ослабить степень мультиколлинеарности. В работе предлагается метод неполной ортогонализации исходных переменных путём замены переменных, приводящий к поддающимся содержательной интерпретации результатам и позволяющий, в силу взаимно-однозначного соответствия исходных и новых переменных, получать прогнозные оценки значения Y, переходить при необходимости от коэффициентов регрессии по новым переменным к коэффициентам регрессии по исходным регрессорам. При этом получены соответствующие формулы перехода для коэффициентов регрессии, ковариационных матриц коэффициентов регрессии, вычисления прогнозных значений новых переменных по прогнозным значениям исходных переменных. Предлагаемый метод замены переменных позволяет существенно уменьшить степень мультиколлинеарности регрессоров, получить интерпретируемые коэффициенты уравнения регрессии и оценить вклад каждого фактора. Применение метода иллюстрируется на примере из более ранней работы автора, продолжением которой можно считать данную работу.

Ключевые слова: мультиколлинеарность, регрессия, индекс обусловленности, ортогонализация переменных

THE APPROACH TO SOLVING THE PROBLEM OF MULTICOLLINEARITY BY USING THE TRANSFORMATION OF VARIABLES

Orlova I.V.

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

The construction of econometric models, their analysis and prediction of the endogenous variable Y from the values of exogenous variables X1, X2, ... Xm, is often made difficult by the presence of multicollinearity of the variables X1, X2, ... Xm. Multicollinearity leads to a number of negative consequences in the construction and analysis of the regression equation. This leads to the need in one way or another to get rid of it or to weaken the degree of multicollinearity. The paper proposes a method of incomplete orthogonalization of the initial variables by changing the variables, leading to meaningful interpretation of the results and allowing, by virtue of the one-to-one correspondence between the initial and new variables, to obtain predictive estimates of the Y value, to switch from the regression coefficients for the new variables to the regression coefficients by source regressors. The corresponding transition formulas for the regression coefficients, the covariance matrices of the regression coefficients, and the calculation of the predicted values of the new variables from the predicted values of the initial variables were obtained. The proposed variable replacement method allows one to significantly reduce the degree of multicollinearity of the regressors, to obtain interpretable coefficients of the regression equation, and to estimate the contribution of each factor. The application of the method is illustrated by an example from an earlier work by the author, the continuation of which can be considered this work.

Keywords: multicollinearity, regression, conditionality index, orthogonalization of variables

Построение эконометрических моделей, их анализ и прогнозирование эндогенной переменной Y по значениям экзогенных переменных X_1, X_2, …, X_m, зачастую затруднено наличием мультиколлинеарности исходных переменных. Переменные называют мультиколлинеарными, если они связаны корреляционной связью почти линейно [1]. В практических исследованиях экзогенные переменные, как правило, в той или иной степени коррелированы. Методы обнаружения мультиколлинеарности реализованы практически во всех пакетах программ эконометрического моделирования [2–4]. Нежелательные последствия мультиколлинеарности приводят к необходимости тем или иным способом избавиться от нее или ослабить степень мультиколлинеарности [5]. Самым распространённым приёмом ослабления мультиколлинеарности является удаление из модели «виновных в мультиколлинеарности» исходных переменных. Однако это приводит к обеднению модели, к невозможности исследовать достаточно полно влияние экзогенных переменных на эндогенную переменную. Другой приём состоит в линейном преобразовании переменных, приводящем к новым, ортогональным, переменным. Обычно речь идёт о преобразовании к главным компонентам. Однако главные компоненты, являясь линейными комбинациями всех экзогенных переменных, плохо поддаются содержа-

тельной интерпретации и потому значения коэффициентов регрессии по главным компонентам мало что говорят исследователю о влиянии исходных переменных на Y. К тому же последние главные компоненты, соответствующие близким к нулю и фактически незначимым собственным числам ковариационной матрицы исходных переменных, неустойчивы к незначительным колебаниям исходных переменных и их обычно удаляют из модели. Третьим способом борьбы с мультиколлинеарностью является применение ридж-регрессии для оценки коэффициентов регрессии. Однако при этом оценки получаются смещёнными, и пользоваться этим методом надо с осторожностью.

В данной работе предлагается метод неполной ортогонализации исходных переменных путём замены переменных, приводящий к поддающимся содержательной интерпретации результатам и позволяющий, в силу взаимно-однозначного соответствия исходных и новых переменных, получать прогнозные оценки значения Y, переходить при необходимости от коэффициентов регрессии по новым переменным к коэффициентам регрессии по исходным регрессорам. При этом получены соответствующие формулы перехода для коэффициентов регрессии, ковариационных матриц коэффициентов регрессии, вычисления прогнозных значений новых переменных по прогнозным значениям исходных переменных.

Материалы и методы исследования

Суть предлагаемого метода состоит в том, что некоторые переменные, коррелированные с другими, заменяются на остатки от регрессии этих переменных на другие регрессоры, имеющие с ними корреляционную связь. При этом никаких допущений относительно этих остатков не делается. Полученные коэффициенты уравнений регрессии используются только для вычисления остатков, которые равны разности между самой заменяемой переменной и вычисленными по уравнению регрессии значениями этой переменной. Таким образом, новые переменные являются линейными комбинациями исходных переменных и константы. Предлагаемый метод замены переменных позволяет существенно уменьшить степень мультиколлинеарности.

Результаты исследования и их обсуждение

Рассмотрим линейную регрессию эндогенной переменной Y на экзогенные переменные X_1, X_2…., X_m. Количество наблюдений равно n. Сведём наблюдения значений независимых переменных X_j в матрицу X размерности n×(m + 1), первый столбец матрицы X состоит из единиц.

Спецификация модели линейной регрессии имеет вид

$$y_i = \beta_0 + \beta_1 x_1^{(i)} + \beta_2 x_2^{(i)} + \ldots + \\ + \beta_m x_m^{(i)} + \varepsilon_i \, i = 1, \ldots n, \qquad (1)$$

где $x_j^{(i)}$ – значение X_j в i-м наблюдении, ε_i – остаточный член регрессии, удовлетворяющий условиям Гаусса – Маркова.

Допустим, что регрессоры мультиколлинеарны и среди них есть группы тесно связанных между собой переменных. Выберем последовательно в каждой группе одну или несколько переменных и с помощью метода наименьших квадратов (МНК) определим коэффициенты дополнительных регрессий выбранных переменных на остальные переменные группы. Остатки этих регрессий обозначим через U_j. Далее будем называть выбранные регрессоры X_j, которые выступали в роли зависимых переменных в дополнительных регрессиях, «выбранными» переменными, а остальные регрессоры X_j – «не выбранными» переменными.

Уравнение регрессии «выбранных» X_j на «не выбранные» регрессоры группы имеет вид

$$x_j^{(i)} = \alpha_0^j + \alpha_1^j x_1^{(i)} + \ldots + \\ + \alpha_k^j x_k^{(i)} + \ldots + \alpha_{j-1}^j x_{j-1}^{(i)} + \\ + \alpha_{j+1}^j x_{j+1}^{(i)} + \ldots + \alpha_m^j x_m^{(i)} + u_j^{(i)}, \qquad (2)$$

где $u_j^{(i)}$ – остатки U_j от регрессии X_j на остальные регрессоры в i-м наблюдении, α_k^j – коэффициент регрессии X_j при переменной X_k. Коэффициенты регрессии в (2), как отмечалось, будем находить с помощью МНК, при этом никаких допущений относительно остатков U_j не делается.

Из (2) получаем

$$u_j^{(i)} = x_j^{(i)} - \alpha_0^j - \alpha_1^j x_1^{(i)} - \ldots - \alpha_{j-1}^j x_{j-1}^{(i)} - \\ - \alpha_{j+1}^j x_{j+1}^{(i)} - \ldots - \alpha_m^j x_m^{(i)}. \qquad (3)$$

Как видим, U_j являются линейной комбинацией исходных регрессоров $X_1, X_2 \ldots, X_m$ и константы.

Обозначим через U матрицу значений новых регрессоров размерности n×(m + 1). По столбцам матрицы, соответствующим «выбранным» X_j, находятся $u_j^{(i)}$ – значения остатков U_j для наблюдения i, столбцы, соответствующие «не выбранным» X_j, равны соответствующим столбцам матрицы X; элементы первого столбца $u_0^{(i)}$ равны единице.

Рассмотрим матрицу A преобразования переменных X_j, представленного формулой (3). Для «выбранных» X_j элементы матрицы определяются равенством (3), для

«не выбранных» X_j столбцы матрицы A равны единичному вектору с единицей на (j + 1)-ом месте, поскольку $U_j = X_j$. Матрица преобразования переменных A, если бы все X_j заменялись бы на остатки регрессий X_j на все остальные регрессоры, имела бы вид

$$A = \begin{vmatrix} 1 & -\alpha_0^1 & \alpha_0^2 & \dots & -\alpha_0^{m-1} & -\alpha_0^m \\ 0 & 1 & -\alpha_1^2 & \dots & -\alpha_1^{m-1} & -\alpha_1^m \\ 0 & -\alpha_2^1 & 1 & \dots & -\alpha_2^{m-1} & -\alpha_2^m \\ 0 & -\alpha_3^1 & -\alpha_3^2 & \dots & -\alpha_3^{m-1} & -\alpha_3^m \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & -\alpha_m^1 & -\alpha_m^2 & \dots & -\alpha_m^{m-1} & 1 \end{vmatrix}$$

В соответствии с (3) матрицы X и U удовлетворяют равенству

$$U = X \cdot A. \quad (4)$$

Остатки U_j коррелируют между собой гораздо меньше, чем X_j между собой. Поэтому, если исходные переменные X_j даже мультиколлинеарны, U_j могут быть почти ортогональны и потому вычисление коэффициентов регрессии и их интерпретация по новым переменным не связано с преодолением мультиколлинеарности регрессоров.

Матрица A является невырожденной (если только X_j не являются строго мультиколлинеарными). Отсюда получаем, что преобразование переменных, определяемое матрицей A, является взаимно-однозначным.

Обозначим через $\boldsymbol{\gamma}$ и $\boldsymbol{\beta}$ векторы коэффициентов регрессии Y на новые переменные и на X_j соответственно, $\boldsymbol{\gamma}^{\mathrm{T}} = (\gamma_0, \gamma_1, \dots \gamma_m)$, $\boldsymbol{\beta}^{\mathrm{T}} = (\beta_0, \beta_1, \dots \beta_m)$. Поскольку преобразование, определяемое матрицей A, является взаимно-однозначным, то минимумы суммы квадратов остатков обеих регрессий, т.е. $Y\text{-}X\cdot\boldsymbol{\beta}$ и $Y\text{-}U\cdot\boldsymbol{\gamma}$, совпадают и вектор остатков регрессии Y на новые переменные совпадает с вектором остатков от регрессии Y на X_j. Очевидно, совпадают и коэффициенты детерминации обеих регрессий.

Получим формулу, описывающую взаимосвязь $\boldsymbol{\gamma}$ и $\boldsymbol{\beta}$.

Уравнения регрессий Y на U_j и X_j можно записать в виде

$$\mathbf{Y} = \mathbf{X}\cdot\beta + \boldsymbol{\varepsilon}, \quad (5)$$

$$\mathbf{Y} = \mathbf{U}\cdot\boldsymbol{\gamma} + \boldsymbol{\varepsilon}. \quad (6)$$

Подставим (4) в (6):

$$\mathbf{Y} = \mathrm{X}\cdot\mathrm{A}\cdot\boldsymbol{\gamma} + \boldsymbol{\varepsilon}. \quad (7)$$

Сравнивая (7) с (5), получаем формулу, показывающую взаимосвязь коэффициентов регрессии:

$$\boldsymbol{\beta} = \mathrm{A}\cdot\boldsymbol{\gamma}. \quad (8)$$

Ковариационную матрицу cov(**β**) вектора коэффициентов регрессии $\boldsymbol{\beta}$ вычисляем исходя из того, что коэффициенты регрессии β_j являются, в соответствии с (8), линейными комбинациями коэффициентов γ_k (k = 0,1,…,m),

$$\mathrm{cov}(\boldsymbol{\beta}) = \mathrm{A}\cdot\mathrm{cov}(\boldsymbol{\gamma})\cdot\mathrm{A}^{\mathrm{T}}.$$

Коэффициенты $\boldsymbol{\gamma}_j$ регрессии Y на новые переменные можно трактовать как приращение Y при изменении X_j на единицу, учитывающее соответствующие изменения остальных регрессоров, то есть при условии, что корреляционная матрица регрессоров при этом не изменяется. В этом случае мы трактуем каждую выбранную переменную как состоящую из двух частей. Одна часть формируется под влиянием корреляционных связей с другими регрессорами, а другая часть – U_j представляет собой «специфическую компоненту», не связанную или слабо связанную с остальными регрессорами X_j. Если меняем на единицу только сам этот фактор, не затрагивая его корреляционных связей с другими факторами, то меняем только его «специфическую» часть, а это и есть остаток U_j.

Для прогнозирования значения Y при значениях регрессоров, равных $\mathrm{Xпр} = (x_1^{\mathrm{пр}}, x_2^{\mathrm{пр}}, \dots, x_m^{\mathrm{пр}})$, можно воспользоваться уравнением регрессии по исходным переменным (5). Однако можно и не переходить к регрессии по X_j, а воспользоваться уравнением регрессии по U_j. Для этого надо вычислить прогнозное значение $\mathrm{Uпр} = (u_1^{\mathrm{пр}}, u_2^{\mathrm{пр}}, \dots, u_m^{\mathrm{пр}})$ по формуле (4): $U_{\mathrm{пр}} = X_{\mathrm{пр}}\cdot A$ и далее воспользоваться оценками коэффициентов уравнения регрессии (6).

Предложенный метод уменьшения мультиколлинеарности проиллюстрируем на примере данных, описанных в работе [3, с. 130], так как данная статья является логическим ее продолжением.

Наличие межфакторных связей проверим с помощью матрицы коэффициентов парной корреляции. На рис. 1 представлена визуализация корреляционной матрицы, полученная в среде R.

На рис. 2 приведены результаты построения регрессии Y на X_1, X_2, X_3, X_4, X_5 в среде R. Коэффициент детерминации достаточно высокий – 0,89; уравнение регрессии значимо, а все его коэффициенты незначимы (p = 0,01), что является признаком частичной (нестрогой) мультиколлинеарности.

Тестирование мультиколлинеарности с помощью наиболее популярного метода факторов инфляции дисперсии (VIF) в среде R [6] не только подтвердило наличие мультиколлинеарности (VIF>5), но и выявило факторы, приводящие к ней – X_1 и X_5 (рис. 3).

```
> library("PerformanceAnalytics")
> tabgr <- tab11[, c(1,2,3,4,5)]
> chart.Correlation(tabgr, histogram=TRUE)
```

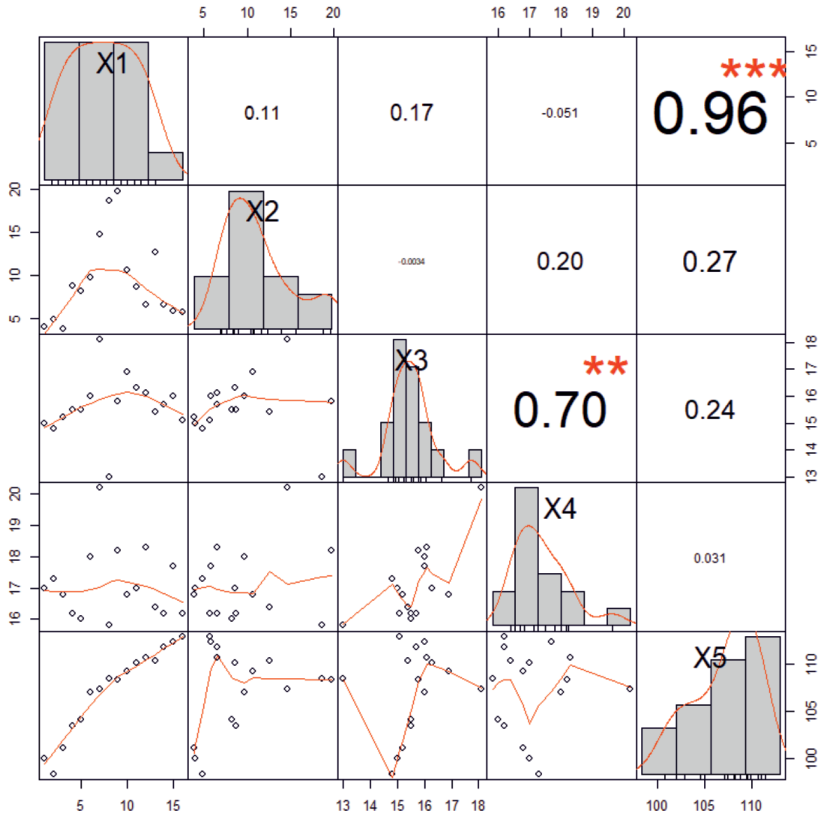

Рис. 1. Матрица коэффициентов парной корреляции

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

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -3017.396   1094.485  -2.757   0.0202 *
X1            -13.419     10.378  -1.293   0.2251
X2              6.672      3.009   2.218   0.0509 .
X3             -6.477     15.779  -0.410   0.6901
X4             12.238     14.410   0.849   0.4156
X5             30.476     11.525   2.644   0.0245 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 41.65 on 10 degrees of freedom
Multiple R-squared:  0.8907,   Adjusted R-squared:  0.8361
F-statistic:  16.3 on 5 and 10 DF,  p-value: 0.0001585
```

Рис. 2. Результат оценки параметров регрессионной модели по всем факторам

Результат выполнения функции eigprop [7–8] (тестирование мультиколлинеарности по методу Белсли) [9–10] приведен на рис. 4. Наибольшее значение индекса обусловленности (CI) в строке 6, равное 376,49, свидетельствует о наличии мультиколлинеарности. В этой строке наибольшие значения среди факторов имеют X_1 и X_5, значит, между этими факторами существует тесная взаимосвязь (заметим, что, по сравнению с предыдущими тестами, появилась новая информация – в пятой строке зафиксирована тесная взаимосвязь между переменными X_3 и X_4). В таких случаях рекомендуется одну из переменных, X_1 или X_5, удалить из модели. В данном случае целесообразно удалить X_1, тем более что и P-значение t-статистики коэффициента регрессии при X_1 равно 0,225, и знак коэффициента регрессии отрицательный, в то время как коэффициент корреляции между Y и X1 положительный.

Однако мы не станем удалять X1, а заменим X_5 на сумму слагаемых $X_5 = \hat{X}_5 + U_5$, где $\hat{X}_5 = \alpha_0^5 + \alpha_1^5 X_1$. Оценивать коэффициенты α_0^5 и α_1^5 будем с помощью МНК. Уравнение регрессии примет вид

$$y_i = \beta_0 + \beta_1 x_1^{(i)} + \beta_2 x_2^{(i)} + \beta_3 x_3^{(i)} + \beta_4 x_4^{(i)} + \beta_5 x_5^{(i)} + \varepsilon_i =$$

$$= \beta_0 + \beta_1 x_1^{(i)} + \beta_2 x_2^{(i)} + \beta_3 x_3^{(i)} + \beta_4 x_4^{(i)} + \beta_5 (\alpha_0^5 + \alpha_1^5 x_1^{(i)} + u_5^i) + \varepsilon_i =$$

$$= \left(\beta_0 + \beta_5 \alpha_0^5\right) + x_1^{(i)} + \left(\beta_1 + \beta_5 \alpha_1^5\right) + \beta_2 x_2^{(i)} + \beta_3 x_3^{(i)} + \beta_4 x_4^{(i)} + \beta_5 u_5^{(i)} + \varepsilon_i =$$

$$= \gamma_0 + \gamma_1 x_1^{(i)} + \gamma_2 x_2^{(i)} + \gamma_3 x_3^{(i)} + \gamma_4 x_4^{(i)} + \gamma_5 u_5^{(i)} + \varepsilon_i, \quad (9)$$

где $\gamma_0 = \beta_0 + \beta_5 \alpha_0^5$, $\gamma_1 = \beta_1 + \beta_5 \alpha_1^5$, $\gamma_j = \beta_j, j = 2, 3, 4, 5$; ε_i – остаточный член регрессии.

Оценим параметры модели $\hat{X}_5 = \alpha_0^5 + \alpha_1^5 X_1$ (рис. 5) и получим остатки $U_5 = X_5 - \hat{X}_5$.

```
fm<-lm(data=tab1,Y~X1+X2+X3+X4+X5 )
> library(car)
> vif(fm)
      X1       X2       X3       X4       X5
21.112    1.889    2.474    2.331   23.389
```

Рис. 3. Факторы инфляции дисперсии параметров регрессии Y по всем переменным

```
library ("mctest")

eigprop(TT,  Inter = TRUE, prop = 0.7)

Call:
eigprop(x = TT, 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. 4. Результат диагностики коллинеарности по методу Белсли [3, с. 133]

```
lm(formula = X5 ~ X1, data = tab1)

Coefficients:
(Intercept)           X1
    99.4950       0.9101
```

Рис. 5. Результат оценки параметров регрессионной модели «выбранного» фактора X_5

```
lm(formula = Y ~ X1 + X2 + X3 + X4 + U5, data = tab5)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   14.798    185.777   0.080 0.938083
X1            14.318      2.442   5.862 0.000159 ***
X2             6.672      3.009   2.218 0.050906 .
X3            -6.477     15.779  -0.410 0.690125
X4            12.238     14.410   0.849 0.415567
U5            30.476     11.525   2.644 0.024548 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 41.65 on 10 degrees of freedom
Multiple R-squared:  0.8907,    Adjusted R-squared:  0.8361
F-statistic:  16.3 on 5 and 10 DF,  p-value: 0.0001585
```

Рис. 6. Результат оценки параметров регрессионной модели по новым переменным

Далее оценим параметры модели $y_i = \gamma_0 + \gamma_1 x_1^{(i)} + \gamma_2 x_2^{(i)} + \gamma_3 x_3^{(i)} + \gamma_4 x_4^{(i)} + \gamma_5 u_5^{(i)}$ (рис. 6).

Коэффициент регрессии β_5 при U_5 оказался тем же самым, что и коэффициент регрессии при X_5. Коэффициент регрессии β_1 при X_1 отличается от нового коэффициента регрессии при X_1 на величину $\beta_5 \alpha_1^5$. Связь между коэффициентами регрессии β_j и γ_j можно получить и без выкладок (9) – по формуле (8): $\boldsymbol{\beta} = A \cdot \boldsymbol{\gamma}$, где матрица А преобразования (3) в данном случае имеет вид

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & -99,495 \\ 0 & 1 & 0 & 0 & 0 & -0,910 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

У нас

$$\boldsymbol{\beta}^{\mathbf{T}} = (-3017,4;\ -13,42;\ 6,67;\ -6,48;\ 12,24;\ 30,48)$$

$$\boldsymbol{\gamma}^{\mathbf{T}} = (14,80;\ 14,32;\ 6,67;\ -6,48;\ 12,24;\ 30,48)$$

Подставляя в (8) А и $\boldsymbol{\gamma}$, получаем приведённый выше вектор $\boldsymbol{\beta}$. Итак, получили, что уравнение регрессии на новые переменные имеет вид

$$\widehat{y_i} = 14,80 + 14,32 x_1^{(i)} + 6,67 x_2^{(i)} - 6,48 x_3^{(i)} + 12,24 x_4^{(i)} + 30,48 u_5^i.$$

В это уравнении регрессии значимо вошли X_1, X_2, U_5 (рис. 6). Коэффициент детерминации равен 0,89.

Далее можно было бы попытаться решить проблему мультиколлинеарности, связанную с X_3, X_4. Однако эти переменные слабо коррелируют с Y и их включение в модель мало изменит коэффициент детерминации. К тому же на основании больших Р-значений у нас есть все основания исключить X_3, X_4 из модели регрессии. Окончательно получаем модель с тремя переменными X_1, X_2, U_5 (рис. 7). Все Р-значения коэффициентов не превосходят 0,02. Коэффициент детерминации уменьшился менее чем на 0,01.

Все VIF(j) не превосходят 1,64 (рис. 8). Как видим, после выполненных преобразований мультиколлинеарность в данных практически отсутствует.

Сравним коэффициенты регрессии и их характеристики уравнений регрессии Y по X_1, X_2, U_5 и Y по X_1, X_2, X_5 (табл. 1) и (табл. 2).

Таблица 1

Параметры результирующей модели регрессии

	γ_0	γ_1	γ_2	γ_5
	const	X_1	X_2	U_5
Коэффициент	121,9	13,85	7,23	30,95
t-статистика	4,01	6,40	2,72	3,01
Р-значение	0,00	<0,0001	0,02	0,01

```
lm(formula = Y ~ X1 + X2 + U5, data = tab5)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  121.867     30.365   4.013  0.00172 **
X1            13.854      2.164   6.402 3.39e-05 ***
X2             7.229      2.654   2.724  0.01847 *
U5            30.951     10.285   3.009  0.01087 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 39.54 on 12 degrees of freedom
Multiple R-squared:  0.8818,    Adjusted R-squared:  0.8523
F-statistic: 29.85 on 3 and 12 DF,  p-value: 7.576e-06
```

Рис. 7. Результат оценки параметров регрессионной модели по переменным X_1, X_2, U_5

```
> vif(fm2)
      X1       X2       U5
1.018483 1.630937 1.612454
```

Рис. 8. Факторы инфляции дисперсии параметров регрессии Y по переменным X_1, X_2, U_5

Таблица 2

Параметры модели регрессии по исходным переменным X_1, X_2, X_5

	β_0	β_1	β_2	β_5
	const	X_1	X_2	X_5
Коэффициент	2958	–14,32	7,23	30,95
t-статистика	2,93	–1,52	2,72	3,01
P-значение	0,01	0,15	0,02	0,01

Коэффициенты β_2 и β_5 и их характеристики совпадают с γ_2 и γ_5, а β_1 не только не близок к γ_1, но имеет отрицательный знак, несмотря на положительный коэффициент корреляции Y с X_1. Очевидно, что по коэффициентам β уравнения регрессии Y по исходным переменным нецелесообразно анализировать степень влияния отдельных регрессоров на Y. В то же время, после замены переменных, если бы регрессоры были бы ортогональны, коэффициенты регрессии нормированного уравнения регрессии были бы равны коэффициентам корреляции регрессоров с Y. У нас же регрессоры не строго ортогональны и коэффициенты корреляции Y с X_1 не строго, а приблизительно равны соответствующим коэффициентам регрессии. Коэффициент корреляции Y с X_1 равен 0,678, а коэффициент нормированного уравнения регрессии при X_1 равен 0,641. Это подтверждает высказанное выше утверждение относительно интерпретации коэффициентов регрессии по новым переменным, полученным после замены переменных.

Заключение

Предложенный метод неполной ортогонализации исходных переменных путём замены переменных позволяет уменьшить степень мультиколлинеарности регрессоров, получить интерпретируемые коэффициенты уравнения регрессии и оценить вклад каждого фактора в изменение эндогенной переменной.

Список литературы

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

2. Шитиков В.К., Мастицкий С.Э. Классификация, регрессия и другие алгоритмы Data Mining с использованием R. 2017351 с. [Электронный ресурс]. URL: https://github.com/ranalytics/data-mining (дата обращения: 12.03.2019).

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

4. Проект R для статистических вычислений. [Электронный ресурс]. URL: http://www.r-project.org/ (дата обращения: 12.03.2019).

5. 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.

6. Орлова И.В. Анализ диагностических индикаторов общей и индивидуальной коллинеарности регрессоров // Фундаментальные исследования. 2019. № 2. С. 16–20.

7. Muhammad Imdad Ullah, Muhammad Aslam Multicollinearity Diagnostic Measures. Package ‘mctest’. [Electronic resource]. URL: https: //cran.r-project.org/web/packages/mctest/mctest.pdf (date of access: 12.03.2019).

8. Muhammad Imdad Ullah, Muhammad Aslam, Saima Altaf mctest: An R Package for Detection of Collinearity among Regressors. The R. Journal. 2016. Vol. 8:2. P. 495–505.

9. Belsley D.A., Kuh E., Welsch R.E. Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. John Wiley & Sons; N.Y., 1980. P. 297.

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