Проблема мультколлинеарности в модели нечеткой линейной регрессии

В. Б. Гисин[1], Б. А. Путко[2], И. З. Ярыгина[3]
Финансовый университет при Правительстве Российской Федерации
[1]vgisin@fa.ru, [2]BAPutko@fa.ru, [3]IYarygina@fa.ru

***Аннотация*. Рассматривается проблема выявления мультиколлинеарности в модели нечеткой линейной регрессии. Предложен подход к выявлению нечеткой линейной зависимости векторов, позволяющий устанавливать наличие линейной зависимости регрессоров.**

Ключевые слова: нечеткая линейная регрессия; мультиколлинеарность; нечеткая линейная зависимость; линейная оптимизация; эластичность

I. Введение

Модели линейной регрессии традиционный, широко распространенный и эффективный инструмент экономических исследований. Регрессионные модели имеют высокую прогностическую способность и позволяют оценивать с контролируемой точностью связь входных и выходных переменных. Точность оценок, получаемых в традиционных регрессионных моделях, обусловлена предположениями, на базе которых строятся такие модели. К числу таких предположений относятся, например, предположение о том, что расхождение между оцениваемыми и наблюдаемыми значениями вызвано случайной ошибкой, или трактовка наблюдаемых значений как случайных величин.

Развитие мягких измерений и вычислений в последние десятилетия позволило в той или иной форме распространить применение ряда моделей, в том числе линейной регрессии, на данные с гораздо более высокой степенью неопределенности. Например, нечеткая (или интервальная) регрессия может использоваться в тех случаях, когда зависимость между объясняющими переменными и объясняемой переменной носит приближенный характер, причем неточность обусловлена не случайными ошибками наблюдений, а природой явления. Хотя результат оказывается менее точным, чем в классической регрессии, однако он оказывается достаточно адекватным используемым данным.

По сравнению с моделями классического регрессионного анализа, в нечетких моделях заметно возрастает роль содержательной интерпретации полученных результатов. Недостаточная точность количественных оценок при этом отчасти компенсируется качественным анализом. Как отмечается в [1] «статистическая» регрессия и регрессия, основанная на «мягких» вычислениях, выступают не оппонентами или конкурентами, а дополняют друг друга.

Ключевая идея, связанная с заменой в модели регрессии случайной ошибки, нечеткими коэффициентами, была представлена в [2]. Нечеткие коэффициенты ищутся таким образом, чтобы, с одной стороны, обеспечить выполнение всех соотношений на заданном уровне уверенности, а с другой, чтобы суммарная нечеткость коэффициентов была минимальной.

Модель нечеткой регрессии может быть построена для трех типов данных: четкие значения регрессоров и четкие значения объясняемой переменной (CICO); четкие значения регрессоров и нечеткие значения объясняемой переменной (CIFO); нечеткие значения регрессоров и нечеткие значения объясняемой переменной (FIFO).

В [2] (см также [3]) оценка коэффициентов нечеткой линейной регрессии получается без использования метода наименьших квадратов в результате решения задачи линейного программирования. Это в определенной степени является математическим выражением смены вероятностно-статистической парадигмы на логико-алгебраическую. Следуя терминологии, предложенной в [1], будем для краткости называть этот подход CFS (conventional fuzzy regression).

Задача построения CFS CICO формулируется следующим образом. Входной информацией служит массив значений объясняющих переменных и объясняемой переменной $((x_{ij}), y_i)$, где $i = 1,\ldots, n$, — номер наблюдения, $j = 0, 1,\ldots, p$, — номер объясняющей переменной, при этом $x_{i0} = 1$ для всех $i = 1,\ldots, n$. На выходе должен быть получен вектор коэффициентов $\boldsymbol{A} = (A_0,\ldots,A_p)^t$, где коэффициенты A_j, $j = 0, 1,\ldots, p$, — симметричные треугольные нечеткие числа с центральным значением a_j и спрэдом $\alpha_j \geq 0$.

Уровневые множества симметричного треугольного нечеткого числа A с центральным значением a и спрэдом $\alpha \geq 0$ представляют собой промежутки вида

$$A^{(h)} = [a - h\alpha, a + h\alpha],\ 0 \leq h \leq 1.$$

Если обозначить через $Poss(x = V)$ меру возможности равенства четкого числа x нечеткому числу V, то

$Poss(x = A) \geq h$ тогда и только тогда, когда $x \in A^{(h)}$.

Коэффициенты A_j, $j = 0, 1,\ldots, p$, подбираются таким образом, чтобы для каждого $i = 1,\ldots, n$ выполнялось условие

$$Poss(y_i = Y_i) \geq h, \qquad (1)$$

где

$$Y_i = A_0 + A_1x_{i1} + \ldots + A_px_{ip},$$

а h — заданное пороговое значение. При этом суммарная неопределенность

$$(|\boldsymbol{X}|^t\, \boldsymbol{I}_n)^t\boldsymbol{\alpha}, \qquad (2)$$

где $|\boldsymbol{X}| = (|x_{ij}|)$, и $\boldsymbol{I}_n = (1,1,\ldots,1)^t$ — вектор размерности $n > 0$, все координаты которого равны единице, а $\boldsymbol{\alpha} = (\alpha_j)$ — вектор-столбец спрэдов

Обозначим через $\boldsymbol{a} = (a_j)$ вектор-столбец центральных значений, через, а через $\boldsymbol{y} = (y_i)$ вектор-столбец значений объясняемой переменной. Далее, пусть $F(\boldsymbol{X})$ — блочная матрица вида

$$F(\boldsymbol{X}) = \begin{pmatrix} \boldsymbol{X} & -(1-h)\,|\,\boldsymbol{X}\,| \\ -\boldsymbol{X} & -(1-h)\,|\,\boldsymbol{X}\,| \end{pmatrix}$$

и $f(\boldsymbol{y}) = (\boldsymbol{y}^t, \ -\boldsymbol{y}^t)^t$. Тогда задача определения коэффициентов нечеткой линейной регрессии может быть сформулирована как задача линейного программирования:

$$(|\boldsymbol{X}|^t\,\boldsymbol{I}_n)^t\boldsymbol{\alpha} \to \min;\ F(\boldsymbol{X})(\boldsymbol{a}^t,\ \boldsymbol{\alpha}^t)^t \le f(\boldsymbol{y});\ \boldsymbol{\alpha} \ge 0. \qquad (3)$$

Простота формулировки задачи CFS CICO делает соответствующие модели достаточно гибкими. К условиям (3) можно легко добавлять дополнительные линейные ограничения, позволяющие учитывать специфику регрессоров. Кроме того, полученные результаты, как правило, допускают достаточно простую и ясную интерпретацию. Это обусловило широкое распространение CFS CICO и применение ее в самых разных сферах (см., например [4]).

В то же время анализ методов CFS CICO выявило ряд серьезных проблем. Некоторые из них обусловлены спецификой линейной оптимизации (например, случаи, когда оптимальное решение не единственно), некоторые присущи подходу, принятому в CFS CICO (например, чрезмерная зависимость от выбросов, см. [5, 6]). Еще одна проблема — проблема мультиколлинеарности, свойственная и классической регрессии, приобретает специфические черты при построении моделей CFS CICO. В [7] проведено обстоятельное сравнение классической и нечеткой линейной регрессии. Однако проблема мультколлинеарности в этой работе по существу не затронута. В многочисленных прикладных исследованиях (там, где применяется CFS CICO) проблема мультколлинеарности, как правило, не находится в фокусе внимания. Мультиколлинеарность может существенно повлиять на оценку параметров нечеткой модели и исказить результат так же, как это может происходить в классических моделях.

В настоящей статье предложен подход к определению нечеткой мультиколлинеарности и ее выявлению.

II. Мультиколлинеарность и нечеткая линейная зависимость

Обозначим через $\boldsymbol{x}_{*j}$ вектор-столбец наблюдений объясняющей переменной j, $j = 0,1,\ldots,p$. В классической модели линейной регрессии под мультиколлинеарностью понимается взаимная зависимость объясняющих переменных. С учетом того, что CFS CICO используется, как правило, тогда, когда статистические методы неприменимы, мы будем понимать под мультиколлинеарностью то, что в классическом случае называют явной формой. В этом случае мультиколлинеарность проявляется как линейная зависимость векторов $\boldsymbol{x}_{*j}$.

Пусть $\boldsymbol{B} = (B_0,\ldots,B_p)^t$ — вектор, составленный из симметричных треугольных нечетких чисел B_j, с центральными значениями b_j и спрэдами $\beta_j \ge 0$, $j = 0,1,\ldots,p$. Соответствующие векторы центральных значений и спрэдов обозначим $\boldsymbol{b}$ и $\boldsymbol{\beta}$. Далее, положим $\boldsymbol{Z} = \boldsymbol{XB}$. Тогда

$$\boldsymbol{X}(\boldsymbol{A} + \boldsymbol{B}) = \boldsymbol{Y} + \boldsymbol{Z}.$$

Если вместе с (1) для всех $i = 1,\ldots, n$ выполняются соотношения

$$Poss(0 = Z_i) \ge h,$$

то

$$Poss(y_i = Y_i + Z_i) \ge h.$$

Таким образом, нечеткий вектор коэффициентов $\boldsymbol{A} + \boldsymbol{B}$ удовлетворяет ограничениям из (3) вместе с вектором $\boldsymbol{A}$. Если спрэды β_j малы, разница значений $(|\boldsymbol{X}|^t\,\boldsymbol{I}_n)^t(\boldsymbol{\alpha} + \boldsymbol{\beta})$ и $(|\boldsymbol{X}|^t\,\boldsymbol{I}_n)^t\boldsymbol{\alpha}$ может быть незначительной (в пределах допустимой точности). В то же время, различие коэффициентов a_j и $a_j + b_j$ для некоторых j может оказаться достаточно существенным.

Чтобы проверить устойчивость оценок коэффициентов нечеткой линейной регрессии, можно применить следующий метод. Пусть $\boldsymbol{a}^*$, $\boldsymbol{\alpha}^*$ — решение задачи (3) и $r^* = (|\boldsymbol{X}|^t\,\boldsymbol{I}_n)^t\boldsymbol{\alpha}^*$. Выберем индекс j, для которого $a_j^* > 0$. Рассмотрим задачу (3) с дополнительным ограничением $a_j \ge (1 + \delta)a_j^*$. Пусть $r = (1 + \Delta)r^*$ — значение целевой функции, соответствующее решению этой задачи. Величину $e_{rj+} = \Delta/\delta$ при небольших значениях δ можно рассматривать как эластичность меры неопределенности r по коэффициенту a_j справа. Аналогичным образом, добавляя к числу ограничений неравенство $a_j \le (1 - \delta)a_j^*$ получаем (с точностью до знака) e_{rj-} — эластичность r относительно a_j слева. Небольшие значения эластичности (порядка коэффициента a_j^*) свидетельствуют об отсутствии мультиколлинеарности и достаточной устойчивости оценок.

Пример 1. Рассматриваются 25 наблюдений с тремя объясняющими переменными.

А) В первом случае векторы $\boldsymbol{x}_{*j}$, $j = 1,2,3$, попарно слабо коррелированы: $\rho_{12} = 0.08$, $\rho_{13} = 0.01$, $\rho_{23} = -0.20$. Оценка целевой функции и коэффициентов регрессии:

$$r = 99.5,\ \boldsymbol{a}^* = (35,\ 1,\ 2,\ 1)^t.$$

В табл. I приведены усредненные значения эластичности.

ТАБЛИЦА I Эластичность. Некоррелированные переменные

j	0	1	2	3
e_{rj+}	1.5	1.3	12	5.2
e_{rj-}	0.8	5.9	8.1	4.2

Б) Во втором случае использовались те же векторы $\boldsymbol{x}_{*1}$ и $\boldsymbol{x}_{*2}$, что и в А. Вектор $\boldsymbol{x}_{*3}$ был сформирован так, что

$$\boldsymbol{x}_{*3} = \boldsymbol{x}_{*1} + \boldsymbol{x}_{*2} + \varepsilon,$$

где ε — случайное возмущение (равномерно распределенное на промежутке [0, 10]). Коэффициенты

корреляции в этом случае оказались равными $\rho_{13} = 0.68$, $\rho_{23} = 0.77$.

Оценка коэффициентов регрессии дает следующие результаты: $\boldsymbol{a}^* = (33.7, 1.03, 1.98, 1.01)^t$, $r^* = 102$.

В табл. II приведены усредненные значения эластичности.

ТАБЛИЦА II ЭЛАСТИЧНОСТЬ. КОРРЕЛИРОВАННЫЕ ПЕРЕМЕННЫЕ

j	0	1	2	3
e_{rj+}	2.8	0.29	1.09	1.3
e_{rj-}	1.08	1.12	1.89	0.5

По таблице видно, что изменение коэффициентов a_1, a_2 и a_3 приводит к незначительному увеличению нечеткости, что свидетельствует о неустойчивости этих коэффициентов и возможной линейной зависимости векторов $\boldsymbol{x}_{*1}$, $\boldsymbol{x}_{*2}$, $\boldsymbol{x}_{*3}$.

Для нахождения нечеткой линейной зависимости между четкими (и нечеткими) векторами могут быть использованы различные методы анализа линейных систем [9, 10]. В то же время, поиск линейной зависимости может быть сведен к задаче линейного программирования, аналогичной задаче построения модели нечеткой линейной регрессии. Для поиска линейной зависимости столбцов матрицы $\boldsymbol{X}$ можно воспользоваться идеей нечеткого нуля из [11]. Для нахождения решения системы $\boldsymbol{XB} = \boldsymbol{0}$ относительно $\boldsymbol{B}$ мы заменяем четкий нулевой вектор $\boldsymbol{0}$ в правой части нечетким нулевым вектором $\boldsymbol{Z}$. Обозначим через $\boldsymbol{b}$ вектор центральных значений, а через $\boldsymbol{\beta}$ — вектор спрэдов вектора нечетких величин $\boldsymbol{B}$. Требуется найти такие компоненты вектора $\boldsymbol{B}$, чтобы при минимальной общей нечеткости хотя бы один из коэффициентов вектора $\boldsymbol{b}$ был достаточно большим (например, не меньше единицы). Если между столбцами матрицы $\boldsymbol{X}$ имеется линейная зависимость, можно добиться при этом того, чтобы суммарная нечеткость была сравнительно невелика.

Таким образом, мы приходим к следующему набору задач линейного программирования:

$$(|\boldsymbol{X}|^t \boldsymbol{I}_p)^t\boldsymbol{\beta} \to \min;\ F(\boldsymbol{X})(\boldsymbol{b}^t, \boldsymbol{\beta}^t)^t \le f(\boldsymbol{0});\ \boldsymbol{\beta} \ge 0,\ \beta_j \ge 1 \qquad (4)$$

с $j = 1,\ldots,p$.

Пример 2. Воспользуемся матрицей $\boldsymbol{X}$ из примера 1Б. Решение задачи (4) при $j = 1$ дает следующие результаты:

$b_0 = 5.7$, $b_1 = 1$, $b_2 = 0.96$, $b_3 = -0.94$, $(|\boldsymbol{X}|^t \boldsymbol{I}_n)^t\boldsymbol{\beta} = 196$;

при $j = 2$:

$b_0 = 5.7$, $b_1 = 0.96$, $b_2 = 1$, $b_3 = -0.94$, $(|\boldsymbol{X}|^t \boldsymbol{I}_n)^t\boldsymbol{\beta} = 196$;

при $j = 3$:

$b_0 = -5.8$, $b_1 = -1.04$, $b_2 = -1.04$, $b_3 = 1$, $(|\boldsymbol{X}|^t \boldsymbol{I}_n)^t\boldsymbol{\beta} = 207$.

Сравнительно низкие оценки нечеткости служат индикатором наличия нечеткой линейной зависимости между столбцами матрицы $\boldsymbol{X}$.

По контрасту для матрицы $\boldsymbol{X}$ из примера 1А имеем:

$j = 1$, $\boldsymbol{b} = (-61, 1, 0.12, 0.07)^t$, $(|\boldsymbol{X}|^t \boldsymbol{I}_n)^t\boldsymbol{\beta} = 699$;

$j = 2$, $\boldsymbol{b} = (-41, -0.02, 1, -0.04)^t$, $(|\boldsymbol{X}|^t \boldsymbol{I}_n)^t\boldsymbol{\beta} = 824$;

$j = 1$, $\boldsymbol{b} = (-115, 0.1, 0.19, 1)^t$, $(|\boldsymbol{X}|^t \boldsymbol{I}_n)^t\boldsymbol{\beta} = 984$.

III. ЗАКЛЮЧЕНИЕ

В работе описан метод выявления мультиколлинеарности в модели нечеткой линейной регрессии, основанный на понятии нечеткой линейной зависимости. Для выявления мультколлинеарности применяется оценка эластичности суммарной нечеткости относительно оценок коэффициентов регрессии. Индикатором мультиколлинеарности служат малые значения эластичности.

СПИСОК ЛИТЕРАТУРЫ

[1] Boukezzoula R., Coquin D. Interval-valued fuzzy regression: Philosophical and methodological issues // Applied Soft Computing. 2021. Т. 103. С. 107145.

[2] Tanaka H., Uejima S., Asai K. Linear regression analysis with fuzzy model // IEEE Transactions on Systems, Man, and Cybernetics. 1982. Т. 12, №. 6. С. 903-907.

[3] Heshmaty B., Kandel A. Fuzzy linear regression and its applications to forecasting in uncertain environment // Fuzzy sets and systems. 1985. Т. 15, №. 2. С. 159-191.

[4] Chukhrova N., Johannssen A. Fuzzy regression analysis: systematic review and bibliography // Applied Soft Computing. 2019. Т. 84. С. 105708.

[5] Redden D.T., Woodall W.H. Properties of certain fuzzy linear regression methods // Fuzzy Sets and Systems. 1994. Т. 64, №. 3. С. 361-375.

[6] Redden D. T., Woodall W. H. Further examination of fuzzy linear regression // Fuzzy Sets and Systems. 1996. Т. 79, №. 2. С. 203-211.

[7] Kim K. J., Moskowitz H., Koksalan M. Fuzzy versus statistical linear regression // European Journal of Operational Research. 1996. Т. 92, №. 2. С. 417-434.

[8] Kim K. J., Chen H. R. A comparison of fuzzy and nonparametric linear regression // Computers & operations research. 1997. Т. 24, №. 6. С. 505-519.

[9] Деменков Н.П., Микрин Е.А., Мочалов И.А. Методы решения нечетких систем линейных уравнений. Ч. 1. Полные системы // Проблемы управления. 2019. № 4. С. 3-14.

[10] Деменков Н.П., Микрин Е.А., Мочалов И.А. Методы решения нечетких систем линейных уравнений. Ч. 2. Неполные системы // Проблемы управления. 2019. №. 5. С. 19-28.

[11] Sevastjanov P., Dymova L. A new method for solving interval and fuzzy equations: linear case //Information sciences. 2009. Т. 179, №. 7. С. 925-937.