ДВУХФАКТОРНАЯ МОДЕЛЬ ПОЛНОСВЯЗНОЙ ЛИНЕЙНОЙ РЕГРЕССИИ ДИНАМИКИ ВВП РОССИИ

М. П. Базилевский

Иркутский государственный университет путей сообщения, Россия
E-mail: mik2178@yandex.ru

Эффективным инструментом интеллектуального анализа данных в области техники, экономики, бизнеса и др., является регрессионный анализ. Поскольку функционирование экономических объектов или процессов, как правило, определяются большим количеством одновременно и совокупно действующих факторов, то возникает задача исследования зависимости одной переменной от нескольких объясняющих переменных. Эта задача решается с помощью построения моделей множественной линейной регрессии. При этом предполагается, что объясняющие переменные не содержат ошибок. Но на практике даже при использовании современных технических средств объясняющие переменные часто оказываются не вполне правильно измеренными. В работе рассмотрена двухфакторная модель полносвязной линейной регрессии, учитывающая не только наличие ошибок в объясняющих переменных, но и не требующая для своего построения решения проблемы мультиколлинеарности. По данным за 2005 – 2017 гг. построена полносвязная регрессия зависимости ВВП от стоимости фиксированного набора потребительских товаров и услуг и денежной массы М2.

TWO-FACTOR FULLY CONNECTED LINEAR REGRESSION MODEL OF GDP DYNAMICS IN RUSSIA

M. P. Bazilevskiy

An effective tool for data mining in the field of technology, economics, business, etc., is regression analysis. Since the functioning of economic objects or processes, as a rule, is determined by a large number of simultaneously acting factors together, the task arises of studying the dependence of one variable on several explanatory variables. This problem is solved by constructing a multiple linear regression model. It is assumed that the explanatory variables are error free. But in practice, even when using modern technical means, explanatory variables are often not quite correctly measured. In this paper a two-factor fully connected linear regression model, which takes into account not only the presence of errors in the explanatory variables, but also does not require a solution to the multicollinearity problem for its construction, is considered. According to the data for 2005 – 2017 years fully-connected regression of the dependence of GDP on the value of a fixed set of consumer goods and services and money supply M2 was built.

Классическая двухфакторная модель множественной линейной регрессии [1–3] имеет вид

$$y_i = \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \varepsilon_i,\ i = \overline{1,n}, \qquad (1)$$

где y_i, $i = \overline{1,n}$ – значения зависимой (объясняемой, выходной) переменной; x_{i1}, x_{i2}, $i = \overline{1,n}$ – значения независимых (объясняющих, входных) переменных; ε_i,

$i=\overline{1,n}$ – ошибки аппроксимации; α_0, α_1, α_2 – неизвестные параметры модели; n – количество наблюдений.

Если объясняющие переменные x_1 и x_2 содержат ошибки, то оценивать модель (1) с помощью метода наименьших квадратов (МНК) некорректно. Обозначим «истинные» значения переменных y, x_1 и x_2 – y_i^*, x_{i1}^*, x_{i2}^*, $i=\overline{1,n}$. «Истинные» значения переменных связаны с наблюдаемыми значениями уравнениями:

$$y_i = y_i^* + \varepsilon_{i0},\ i=\overline{1,n}, \tag{2}$$

$$x_{i1} = x_{i1}^* + \varepsilon_{i1},\ i=\overline{1,n}, \tag{3}$$

$$x_{i2} = x_{i2}^* + \varepsilon_{i2},\ i=\overline{1,n}. \tag{4}$$

Предположим, что между переменными x_1^*, x_2^* линейная функциональная зависимость:

$$x_{i1}^* = a + bx_{i2}^*,\ i=\overline{1,n}, \tag{5}$$

где a, b – неизвестные параметры.

Тогда совокупность уравнений (3) – (5) представляет собой известную регрессию Деминга [4,5], успешно применяемую в клинической химии [6]. Она оценивается с помощью метода наименьших полных квадратов (МНПК) [7]:

$$\sum_{i=1}^{n}\left(x_{i1}-a-bx_{i2}^*\right)^2+\frac{1}{\lambda}\sum_{i=1}^{n}\left(x_{i2}-x_{i2}^*\right)^2\to\min, \tag{6}$$

где $\lambda=\sigma_{\varepsilon_2}^2/\sigma_{\varepsilon_1}^2$ – соотношение дисперсий ошибок (лямбда-параметр); $\sigma_{\varepsilon_1}^2$, $\sigma_{\varepsilon_2}^2$ – дисперсии ошибок переменных x_1 и x_2.

Если значение λ известно, то оценки регрессии Деминга находятся по формулам:

$$\tilde{b}=\frac{D_{x_1}-\frac{1}{\lambda}D_{x_2}+\sqrt{\left(D_{x_1}-\frac{D_{x_2}}{\lambda}\right)^2+4\frac{K_{x_1x_2}^2}{\lambda}}}{2K_{x_1x_2}}, \tag{7}$$

$$\tilde{a}=\overline{x_1}-\tilde{b}\overline{x_2}, \tag{8}$$

$$x_{i2}^*=-\frac{\tilde{a}\tilde{b}}{\frac{1}{\lambda}+\tilde{b}^2}+\frac{\tilde{b}}{\frac{1}{\lambda}+\tilde{b}^2}x_{i1}+\frac{\frac{1}{\lambda}}{\frac{1}{\lambda}+\tilde{b}^2}x_{i2},\ i=\overline{1,n}, \tag{9}$$

где D_{x_1}, D_{x_2} – дисперсии переменных; $K_{x_1x_2}$ – ковариация.

Стоит отметить, что в регрессии Деминга для любого значения λ справедливо неравенство $bK_{x_1x_2}>0$, т.е. оценка параметра b всегда совпадает со знаком ковариации $K_{x_1x_2}$ между переменными x_1 и x_2.

Так как переменная x_2^* является линейной комбинацией (9) переменных

x_1 и x_2, то используем её в качестве независимой переменной в модели парной линейной регрессии:

$$y_i = c_0 + c_1 x_{i2}^* + \varepsilon_i, \; i = \overline{1,n}, \tag{10}$$

где неизвестные параметры c_0 и c_1 находятся с помощью МНК.

Совокупность уравнений (3) – (5), (10) называется двухфакторной моделью полносвязной линейной регрессии [8–10]. Так как при построении полносвязной регрессии одновременно оцениваются две однофакторные зависимости – (5) и (10), то для неё не требуется решать проблему мультиколлинеарности.

Валовой внутренний продукт (ВВП) [11–13] – это один из основных макроэкономических показателей России. Для стабильной экономической ситуации в стране необходимо постоянно поддерживать темпы его роста на достойном уровне. При этом актуальной задачей является исследование влияния различных финансово-экономических показателей на ВВП. Для построения двухфакторной полносвязной регрессии требовалось определиться с выходной переменной y и входными переменными x_1 и x_2. Согласно известной формуле Ирвинга Фишера:

$$P \cdot Q = M \cdot V, \tag{11}$$

где P – уровень цен; Q – объем производства (например, годовой ВВП); M – объем денежной массы; V – скорость обращения денежной массы.

Таким образом, из равенства (11) следует, что годовой ВВП Q связан как с уровнем цен P, так и с объемом денежной массы M. Используя этот факт, на официальном сайте Федеральной службы государственной статистики были собраны годовые данные, представленные в табл. 1, за период 2005 – 2017 гг. по следующим показателям: y – ВВП (в текущих ценах, млрд руб.); x_1 – стоимость фиксированного набора потребительских товаров и услуг (руб.); x_2 – денежная масса М2 (млрд руб.).

Таблица 1

Статистические данные

Год	ВВП, y	Цены, x_1	Масса, x_2
2005	21609,77	4540,225	4353,9
2006	26917,2	5108,766	6032,1
2007	33247,51	5746,28	8970,7
2008	41276,85	6731,188	12869
2009	38807,22	7577,886	12975,9
2010	46308,54	8320,481	15267,6
2011	60282,54	9183,053	20011,9
2012	68163,88	9518,888	24204,8
2013	73133,9	10419,55	27164,6
2014	79199,66	11455,34	31155,6
2015	83387,19	13028,25	31615,7
2016	86148,57	13889,8	35179,7
2017	92037,18	14668,68	38418

Матрица парных коэффициентов корреляции переменных y, x_1 и x_2 представлена в табл. 2.

Таблица 2

Корреляционная матрица

	y	x_1	x_2
y	1	0,9793	0,9958
x_1	0,9793	1	0,9879
x_2	0,9958	0,9879	1

По корреляционной матрице видно, что все переменные сильно коррелируют друг с другом. С одной стороны это хорошо, поскольку подтверждается справедливость зависимости (11). А с другой стороны плохо, потому что из-за тесной корреляции между переменными x_1 и x_2 возникает явление мультиколлинеарности, из-за чего МНК-оценки параметров модели множественной линейной регрессии будут искажены. Действительно, оцененная с помощью МНК модель множественной линейной регрессии имеет вид:

$$y^* = 18647,8 - 1,3286x_1 + 2,4897x_2. \qquad (12)$$

Критерий детерминации для модели (12) $R^2 = 0,9924$, что говорит о её высоком качестве. В принципе, регрессию (12), даже не смотря на мультиколлинеарность, можно использовать для прогнозирования. Но интерпретировать её коэффициенты категорически нельзя! Достаточно заметить, что в уравнении (12) знак коэффициента при переменной x_1 не соответствует содержательному смыслу задачи.

Построим двухфакторную модель полносвязной регрессии. При этом главной проблемой является то, что неизвестно значение параметра λ, т.е. соотношение дисперсий ошибок переменных x_1 и x_2. Предположим, что это отношение равно отношению дисперсий переменных, т.е. $\lambda = D_{x_2} / D_{x_1}$. Такую регрессию принято называть диагональной [14, 15]. Для такого параметра λ двухфакторная модель полносвязной регрессии имеет вид

$$y^* = 14327,833 + 2,1037x_2^*, \qquad (13)$$

$$x_1^* = 3251,483 + 0,2905x_2^*, \qquad (14)$$

$$x_2^* = -5596,26 + 1,721x_1 + 0,5x_2. \qquad (15)$$

Коэффициенты детерминации модели (13) – (15) по переменным y, x_1 и x_2 составляют $R_y^2 = 0,9812$, $R_{x_1}^2 = 0,9939$, $R_{x_2}^2 = 0,9939$, что также говорит о её весьма высоком качестве.

Если из уравнения (14) выразить переменную x_2^* и подставить в её в равенство (13), то получим

$$y^* = -9218,276 + 7,2415x_1^*. \qquad (16)$$

Таким образом, в уравнениях (13), (14) и (16) знаки всех угловых коэффициентов (2,1037, 0,2905 и 7,2415) совпадают с соответствующими знаками

парных коэффициентов корреляции (табл. 2), поэтому эти коэффициенты можно интерпретировать. Также научный интерес вызывает тот факт, что полученные предложенным методом оценки отличаются от МНК-оценок соответствующих моделей парной линейной регрессии. Кроме того, если подставить в уравнения (13) и (14) выражение (15), то получим зависимости «истинных» переменных y^* и x_1^* от наблюдаемых переменных x_1 и x_2. Это означает, что регулируя наблюдаемые переменные x_1 и x_2, можно прогнозировать «истинные» значения всех входящих в модель факторов.

СПИСОК ЛИТЕРАТУРЫ

1. *Harrell Jr., Frank E.* Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis. Springer Series in Statistics, 2015. 582 p.
2. *Kuhn M., Johnson K.* Applied predictive modeling. Springer, 2018. 600 p.
3. *Носков С. И., Базилевский М. П.* Построение регрессионных моделей с использованием аппарата линейно-булевого программирования. Иркутск : ИрГУПС, 2018. 176 с.
4. *Wu C., Yu J. Z.* Evaluation of linear regression techniques for atmospheric applications: the importance of appropriate weighting // Atmospheric Measurement Techniques. 2018. Vol. 11. P. 1233–1250.
5. *Deming W. E.* Statistical adjustment of data. Wiley, 1943. 273 p.
6. *Henderson C. M., Shulman N. J., MacLean B., MacCoss M. J., Hoofnagle A. N.* Skyline performs as well as vendor software in the quantitative analysis of serum 25-hydroxy vitamin D and vitamin D binding globulin // Clinical Chemistry. 2018. Vol. 64. P. 408–410.
7. *Gillard J.* An overview of linear structural models in errors in variables regression // REVSTAT – Statistical Journal. 2010. Vol. 8. No. 1. P. 57–80.
8. *Базилевский М. П.* Синтез модели парной линейной регрессии и простейшей EIV-модели // Моделирование, оптимизация и информационные технологии. 2019. Т. 7. № 1 (24). С. 170–182.
9. *Базилевский М.П.* Исследование двухфакторной модели полносвязной линейной регрессии // Моделирование, оптимизация и информационные технологии. 2019. Т. 7. № 2 (25). С. 80–96.
10. *Базилевский М. П.* Оценивание параметров простейшей модели полносвязной линейной регрессии // Достижения и приложения современной информатики, математики и физики: материалы VII Всероссийской научно-практической конференции. 2018. С. 179–184.
11. *Aivazian S. A., Bereznyatsky A. N., Brodsky B. E.* Macroeconomic modeling of the Russian economy // Applied Econometrics. 2017. Vol. 47. P. 5–27.
12. *Кирилюк И. Л.* Модели производственных функций для российской экономики // Компьютерные исследования и моделирование. 2013. Т. 5. № 2. С. 293–312.
13. *Лычагина Т. А., Пахомова Е. А., Писарева Д. А.* Применение аппарата производственных функций для анализа влияния состояния основных фондов на экономический рост РФ // Национальные интересы: приоритеты и безопасность. 2016. С. 4–19.
14. *Базилевский М. П.* Аналитические зависимости между коэффициентами детерминации и соотношением дисперсий ошибок исследуемых признаков в модели регрессии Деминга // Математическое моделирование и численные методы. 2016. № 2 (10). С. 104–116.
15. *Базилевский М. П.* Аналитические зависимости для некоторых критериев адекватности модели регрессии Деминга // Вестник Иркутского государственного технического университета. 2016. Т. 20. № 10. С. 81–89.