ЧИСЛЕННОЕ ОЦЕНИВАНИЕ ТРЕХФАКТОРНЫХ МОДЕЛЕЙ ПОЛНОСВЯЗНОЙ ЛИНЕЙНОЙ РЕГРЕССИИ

М. П. Базилевский

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

В работе рассмотрена трехфакторная модель полносвязной линейной регрессии, проблема оценивания которой заключается в решении системы нелинейных уравнений. Показано, что эту систему можно заменить равносильной и более простой системой, для решения которой могут быть использованы такие численные методы, как метод подстановки и простых итераций. На примере моделирования ВВП России с помощью метода простых итераций найдены оценки трехфакторной полносвязной регрессии. Используя полученные оценки "истинных" значений одной из трех входящих переменных, была построена зависимость ВВП России от грузооборота железнодорожного транспорта, оборота розничной торговли и экспорта. Полученная модель оказалась несколько хуже по величине коэффициента детерминации, чем множественная регрессия. Но при этом все знаки её коэффициентов удовлетворяют содержательному смыслу задачи. Рассмотренный пример демонстрирует, что полносвязные регрессии можно успешно использовать для выпрямления искаженных из-за мультиколлинеарности коэффициентов.

NUMERICAL ESTIMATION OF THREE-FACTOR FULLY CONNECTED LINEAR REGRESSION MODELS

M. P. Bazilevskiy

In this paper a three-factor fully connected linear regression model is considered, the problem of estimating which is to solve a system of nonlinear equations. It is shown that this system can be replaced by an equivalent and simpler system, for the solution of which such numerical methods as the method of substitution and simple iterations can be used. On the example of modeling Russia's GDP using the method of simple iterations, estimates of three-factor fully connected regression are found. Using the obtained estimates of the "true" values of one of the three input variables, the dependence of Russia's GDP on railway freight turnover, retail trade turnover, and export was construct. The resulting model turned out to be slightly worse in terms of the coefficient of determination than multiple regression. But in this case, all the signs of its coefficients satisfy the meaningful meaning of the problem. The above example demonstrates that fully connected regressions can be successfully used to straighten coefficients distorted due to multicollinearity.

Регрессионному анализу в настоящее время посвящено множество работ (см., например, [1–5]). В работах [6,7] впервые рассмотрены и исследованы двухфакторные модели полносвязной линейной регрессии, представляющие собой синтез модели парной линейной регрессии и регрессии Деминга [8]. В работе [9] сформулировано обобщение этих моделей на случай трех входных переменных. Рассмотрим это обобщение подробнее.

Пусть x_{i1}, x_{i2}, x_{i3}, $i=\overline{1,n}$ – наблюдаемые значения трех входных перемен-

ных x_1, x_2, x_3. Предположим, что существуют «истинные» значения этих переменных x_{i1}^*, x_{i2}^*, x_{i3}^*, $i=\overline{1,n}$, которые связаны с наблюдаемыми соотношениями:

$$x_{ij} = x_{ij}^* + \varepsilon_i^{(x_j)},\ i=\overline{1,n},\ j=\overline{1,3}, \tag{1}$$

где $\varepsilon_i^{(x_j)}$ – ошибки аппроксимации для j-й переменной.

Предположим, «истинные» переменные x_1^*, x_2^* и x_3^* связаны функциональными зависимостями:

$$x_{i1}^* = a_1 + b_1 x_{i3}^*, \tag{2}$$

$$x_{i2}^* = a_2 + b_2 x_{i3}^*, \tag{3}$$

где a_1, a_2, b_1, b_2 – неизвестные параметры.

Совокупность уравнений (1) – (3) представляет собой трехфакторную модель полносвязной линейной регрессии без выходных переменных.

Для нахождения неизвестных оценок модели (1) – (3) нужно решить оптимизационную задачу:

$$\lambda_1 \sum_{i=1}^{n}\left(x_{i1} - a_1 - b_1 x_{i3}^*\right)^2 + \lambda_2 \sum_{i=1}^{n}\left(x_{i2} - a_2 - b_2 x_{i3}^*\right)^2 + \sum_{i=1}^{n}\left(x_{i3} - x_{i3}^*\right)^2 \to \min, \tag{4}$$

где $\lambda_1 = \dfrac{\sigma^2_{\varepsilon^{(x_3)}}}{\sigma^2_{\varepsilon^{(x_1)}}}$, $\lambda_2 = \dfrac{\sigma^2_{\varepsilon^{(x_3)}}}{\sigma^2_{\varepsilon^{(x_2)}}}$ – соотношения дисперсий ошибок переменных.

Если соотношения дисперсий ошибок переменных λ_1 и λ_2 известны, то решение задачи (4) сводится в первую очередь к решению системы нелинейных уравнений:

$$\begin{cases} \lambda_1\left(\lambda_2 b_2 K_{x_1x_2} + K_{x_1x_3}\right)b_1^2 + \left(D_{x_3} + \lambda_2^2 b_2^2 D_{x_2} + 2\lambda_2 b_2 K_{x_2x_3} - \lambda_1 D_{x_1} - \lambda_1\lambda_2 b_2^2 D_{x_1}\right)b_1 - \\ \qquad -\left(1+\lambda_2 b_2^2\right)\left(K_{x_1x_3} + \lambda_2 b_2 K_{x_1x_2}\right) = 0, \\ \lambda_2\left(\lambda_1 b_1 K_{x_1x_2} + K_{x_2x_3}\right)b_2^2 + \left(D_{x_3} + \lambda_1^2 b_1^2 D_{x_1} + 2\lambda_1 b_1 K_{x_1x_3} - \lambda_2 D_{x_2} - \lambda_1\lambda_2 b_1^2 D_{x_2}\right)b_2 - \\ \qquad -\left(1+\lambda_1 b_1^2\right)\left(\lambda_1 b_1 K_{x_1x_2} + K_{x_2x_3}\right) = 0, \end{cases} \tag{5}$$

где символом D обозначены дисперсии переменных, а K – ковариации.

Первое уравнение системы (5) является квадратным относительно переменной b_1, а второе – относительно b_2. К сожалению, получить аналитическое решение системы (5) не представляется возможным.

В работе [10] показано, что система (5) равносильна следующей системе:

$$\begin{cases} b_1 = \dfrac{-\left(D_{x_3} + \lambda_2^2 b_2^2 D_{x_2} + 2\lambda_2 b_2 K_{x_2x_3} - \lambda_1 D_{x_1} - \lambda_1\lambda_2 b_2^2 D_{x_1}\right) + \sqrt{D_1}}{2\lambda_1\left(\lambda_2 b_2 K_{x_1x_2} + K_{x_1x_3}\right)}, \\ b_2 = \dfrac{-\left(D_{x_3} + \lambda_1^2 b_1^2 D_{x_1} + 2\lambda_1 b_1 K_{x_1x_3} - \lambda_2 D_{x_2} - \lambda_1\lambda_2 b_1^2 D_{x_2}\right) + \sqrt{D_2}}{2\lambda_2\left(\lambda_1 b_1 K_{x_1x_2} + K_{x_2x_3}\right)}, \end{cases} \tag{6}$$

в которой

$$D_1 = \left(D_{x_3} + \lambda_2^2 b_2^2 D_{x_2} + 2\lambda_2 b_2 K_{x_2 x_3} - \lambda_1 D_{x_1} - \lambda_1 \lambda_2 b_2^2 D_{x_1}\right)^2 +$$
$$+4\lambda_1 \left(\lambda_2 b_2 K_{x_1 x_2} + K_{x_1 x_3}\right)^2 \left(1 + \lambda_2 b_2^2\right), \quad (7)$$
$$D_2 = \left(D_{x_3} + \lambda_1^2 b_1^2 D_{x_1} + 2\lambda_1 b_1 K_{x_1 x_3} - \lambda_2 D_{x_2} - \lambda_1 \lambda_2 b_1^2 D_{x_2}\right)^2 +$$
$$+4\lambda_2 \left(\lambda_1 b_1 K_{x_1 x_2} + K_{x_2 x_3}\right)^2 \left(1 + \lambda_1 b_1^2\right). \quad (8)$$

Для решения системы (6) можно применить, по крайней мере, два численных метода:

1) метод подстановки;

2) метод простых итераций.

В соответствии с первым из них можно, например, подставить в первое уравнение системы (6) вместо переменной b_2 второе уравнение. В результате будет получено линейное уравнение относительно переменной b_1, которое можно решить, например, методом половинного деления.

Второй метод реализуется по следующему итерационному алгоритму. Задается начальное приближение $b^{(0)} = \left(b_1^{(0)}, b_2^{(0)}\right)$ (для этого можно использовать оценки парных линейных регрессий) и малое положительное число ε (точность). По формулам (6) вычисляются новые значения переменных до тех пор, пока не будет обеспечена заданная точность ε.

Пусть из решения системы (6) найдены оценки полносвязной регрессии $\tilde{b}_1$ и $\tilde{b}_2$. Для нахождения её оставшихся параметров нужно воспользоваться формулами [9]:

$$\tilde{a}_1 = \overline{x_1} - \tilde{b}_1 \overline{x_3},\ \tilde{a}_2 = \overline{x_2} - \tilde{b}_2 \overline{x_3}, \quad (9)$$
$$\tilde{x}_{i3}^* = \frac{-\lambda_1 a_1 b_1 - \lambda_2 a_2 b_2 + \lambda_1 b_1 x_{i1} + \lambda_2 b_2 x_{i2} + x_{i3}}{1 + \lambda_1 b_1^2 + \lambda_2 b_2^2},\ i = \overline{1,n}. \quad (10)$$

Для оценивания полносвязной регрессии на официальном сайте Федеральной службы государственной статистики были собраны годовые данные, представленные в табл. 1, за период 2000 – 2018 гг. по следующим показателям:

y – ВВП России (в текущих ценах, млрд руб.);

x_1 – грузооборот железнодорожного транспорта России (млрд т-км);

x_2 – оборот розничной торговли по России (млн руб.);

x_3 – экспорт (млн руб.).

Матрица парных коэффициентов корреляции переменных y, x_1, x_2 и x_3 представлена в табл. 2.

По корреляционной матрице видно, что все переменные сильно коррелируют друг с другом. Поэтому МНК-оценки параметров модели множественной линейной регрессии будут искажены из-за мультиколлинеарности. Действительно, оцененная с помощью МНК модель множественной линейной регрессии имеет вид:

$$\tilde{y} = -1561{,}07 - 0{,}667 x_1 + 0{,}00223 x_2 + 0{,}00131 x_3. \quad (11)$$

Таблица 1

Статистические данные

Год	y	x_1	x_2	x_3
2000	7305,646	1373	2352274	2792051
2001	8943,582	1434	3070014	2816451
2002	10830,5	1510	3765364	3199832
2003	13208,23	1669	4529633	3960851
2004	17027,19	1802	5642498	5124175
2005	21609,77	1858	7041509	6792679
2006	26917,2	1951	8711920	8082559
2007	33247,51	2090	10868976	8864237
2008	41276,85	2116	13944183	11592168
2009	38807,22	1865	14599153	9458444
2010	46308,54	2011	16512047	11921583
2011	60282,54	2128	19104337	15147871
2012	68163,88	2222	21394526	16392649
2013	73133,9	2196	23685914	16620445
2014	79058,48	2301	26356237	19181680
2015	83094,3	2306	27526793	20850458
2016	86014,2	2344	28240885	18615331
2017	92101,35	2493	29745536	20608255
2018	103875,8	2598	31579372	27776058

Таблица 2

Корреляционная матрица

	y	x_1	x_2	x_3
y	1	0,9418	0,9969	0,9892
x_1	0,9418	1	0,9371	0,9476
x_2	0,9969	0,9371	1	0,9812
x_3	0,9892	0,9476	0,9812	1

Критерий детерминации для модели (12) $R^2 = 0,997$, что говорит о её высоком качестве. Как и предполагалось, в уравнении (11) коэффициент при переменной x_1 не удовлетворяет содержательному смыслу задачу, увеличение грузооборота железнодорожного транспорта не может приводить к уменьшению ВВП России.

Для построения полносвязной регрессии (1) – (3) были заданы соотношения дисперсий ошибок $\lambda_1 = D_{x_3} / D_{x_1} = 444774515,92$, $\lambda_2 = D_{x_3} / D_{x_2} = 0,52815$. Методом простых итераций было получено следующее решение системы (6):

$$\tilde{b}_1 = 4,66874 \cdot 10^{-5}, \qquad \tilde{b}_2 = 1,3711.$$

Затем с использованием формул (9) и (10) была построена следующая модель полносвязной линейной регрессии:

$$\tilde{x}_1^* = 1449,3862 + 4,66874 \cdot 10^{-5} \tilde{x}_3^*, \tag{12}$$

$$\tilde{x}_2^* = -863420,217 + 1,3711 \tilde{x}_3^*, \tag{13}$$

$$\tilde{x}_3^* = -9948732{,}882 + 7009{,}723x_1 + 0{,}244x_2 + 0{,}337x_3 . \quad (14)$$

После чего с помощью МНК была оценена модель парной линейной регрессии переменной y от $\tilde{x}_3^*$:

$$\tilde{y} = -4453{,}117 + 0{,}00433\tilde{x}_3^* . \quad (15)$$

Коэффициент детерминации регрессии (15) $R^2 = 0{,}982$.

Подставляя (14) в (15), получим уравнение

$$\tilde{y} = -47565{,}4 + 30{,}376x_1 + 0{,}00106x_2 + 0{,}00146x_3 . \quad (16)$$

Таким образом, качество регрессии (16) по коэффициенту детерминации несколько ниже, чем для (11). Но при этом все знаки коэффициентов в уравнении (16) удовлетворяют содержательному смыслу задачи, поэтому модель (16) можно использовать для интерпретации. Полученный результат доказывает, что полносвязные регрессии могут успешно применяться для выпрямления искаженных из-за мультиколлинеарности коэффициентов.

СПИСОК ЛИТЕРАТУРЫ

1. *Kuhn M., Johnson K.* Applied predictive modeling. Springer, 2018. 600 p.

2. Носков С. И. Метод антиробастного оценивания параметров линейной регрессии: число максимальных по модулю ошибок аппроксимации // Южно-Сибирский научный вестник. 2020. № 1 (29). С. 51-54.

3. Носков С. И. О методе смешанного оценивания параметров линейной регрессии // Информационные технологии и математическое моделирование в управлении сложными системами. 2019. № 1 (2). С. 41-45.

4. Носков С. И., Баенхаева А. В. Множественное оценивание параметров линейного регрессионного уравнения // Современные технологии. Системный анализ. Моделирование. 2016. № 3 (51). С.133-138.

5. Базилевский М. П., Врублевский И. П., Носков С. И., Яковчук И. С. Среднесрочное прогнозирование эксплуатационных показателей функционирования Красноярской железной дороги // Фундаментальные исследования. 2016. № 10-3. С. 471-476.

6. *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.

7. *Базилевский М. П.* Исследование двухфакторной модели полносвязной линейной регрессии // Моделирование, оптимизация и информационные технологии. 2019. Т. 7. № 2 (25). С. 80–96.

8. *Deming W. E.* Statistical adjustment of data. Wiley, 1943. 273 p.

9. *Базилевский М. П.* Проблема оценивания трехфакторных моделей полносвязной линейной регрессии // Актуальные проблемы прикладной математики, информатики и механики: сборник трудов Международной научной конференции. 2020. С. 587-590.

10. *Базилевский М. П.* Методы построения регрессионных моделей с ошибками во всех переменных. Иркутск: ИрГУПС, 2019. 208 с.