СИСТЕМНЫЙ АНАЛИЗ СОЦИАЛЬНО-ЭКОНОМИЧЕСКИХ ПРОЦЕССОВ

УДК 519.862.6 ISSN 1995-5499
DOI: https://doi.org/10.17308/sait.2020.1/2596
Поступила в редакцию 09.11.2019
Подписана в печать 15.03.2020

МНОГОКРИТЕРИАЛЬНЫЙ ПОДХОД К ПОСТРОЕНИЮ ДВУХФАКТОРНЫХ ПОЛНОСВЯЗНЫХ РЕГРЕССИЙ НА ПРИМЕРЕ МОДЕЛИРОВАНИЯ ВВП РОССИИ

© 2020 М. П. Базилевский✉

Иркутский государственный университет путей сообщения
ул. Чернышевского, 15, 664074 Иркутск, Российская Федерация

Аннотация. В настоящее время регрессионные модели чаще всего строятся в предположении, что объясняющие переменные не содержат ошибок. Для регрессионных моделей с ошибками в объясняющих переменных, более известных как «errors-in-variables models», разработан весьма мощный математический аппарат, однако широкого практического применения они почти не находят. Для этого ранее были разработаны модели полносвязной линейной регрессии. Целью данной работы является исследование возможности применения двухфакторной полносвязной регрессии в качестве инструмента для улучшения верифицируемой по нескольким критериям адекватности двухфакторной множественной модели. В статье кратко описаны двухфакторные полносвязные регрессии. Для оценивания суммарного качества регрессионных моделей предложен агрегированный критерий, представляющий собой линейную комбинацию четырех хорошо известных критериев адекватности. На основе этого критерия сформулирована задача выбора оптимальных оценок вторичного уравнения полносвязной регрессии. Такая задача формализована в виде задачи математического программирования. Рассмотрен приближенный способ её решения. С использованием предложенного способа построены регрессионные модели ВВП России для различных условий. При этом полученная в результате модель ВВП оказалась по агрегированному критерию лучше классической множественной регрессии более чем в 2 раза. Предложенную в работе методику можно использовать как инструмент для борьбы с автокорреляцией ошибок и для повышения согласованности поведения фактической и расчетной траекторий изменения значений объясняемой переменной.

Ключевые слова: множественная регрессия; полносвязная регрессия; модель с ошибками во всех переменных; регрессия Деминга; критерии адекватности; автокорреляция ошибок; ВВП России.

ВВЕДЕНИЕ

Задача эффективной обработки и анализа статистических данных с целью извлечения из них неизвестных ранее полезных знаний всегда была и остается актуальной. Одним из признанных во всем мире инструментов анализа данных является регрессионный анализ [1–3]. Чаще всего регрессионные модели оцениваются с помощью метода наименьших квадратов (МНК) в предположении, что объясняющие переменные не содержат ошибок. Если же в них содержатся ошибки, что представляется более реалистичным, то такие модели в зарубежной литературе называют «Errors-In-Variables models» (EIV модели) или «Measurement Error models». Качественный обзор методов их оценивания можно найти в работе Дж. Гилларда [4]. Анализ этой работы

✉ Базилевский Михаил Павлович
e-mail: mik2178@yandex.ru

позволяет сделать вывод, что для оценивания EIV моделей разработан весьма мощный математический аппарат. Кроме того, этот аппарат постоянно совершенствуется, появляются новые статьи. Так, в работе [5] предлагается «compound regression» (смешанная регрессия), в работе [6] исследуется минимаксный подход к построению EIV моделей и т. д. Однако широкого практического применения, в отличие от традиционных регрессий без ошибок, EIV модели почти не находят. Это связано с тем, что они, вообще говоря, не пригодны для точечного прогнозирования, а также возникают проблемы с интерпретацией их оценок. Исключением является регрессия Деминга [7], которая нашла широкое применение в клинической химии [8, 9] и связанных областях. Но стоит признать, что в последнее время прикладной составляющей EIV моделей в научной литературе пытаются уделять гораздо больше внимания, чем раньше. Например, в работах [10, 11] EIV модели применены для обработки экономических данных, в [12–14] — для обработки геопространственных данных, в [15] — для моделирования запасов рыб.

В работах [16, 17] автором разработан синтез простейшей EIV модели и классической парной регрессии, названный двухфакторной моделью полносвязной линейной регрессии. Такие модели целесообразно строить при сильной корреляции между объясняющими переменными, поскольку они по определению лишены мультиколлинеарности. Полносвязные регрессии можно использовать как для интерпретации, так и для прогнозирования. Целью данной работы является исследование возможности применения двухфакторной полносвязной регрессии в качестве инструмента для улучшения верифицируемой по нескольким критериям адекватности двухфакторной множественной модели на примере моделирования валового внутреннего продукта (ВВП) России.

При проведении анализа статистических данных существует два варианта их интерпретации [18]: вероятностно-статистический и логико-алгебраический. В соответствии с первым из них совокупность рядов наблюдений трактуется как выборка из соответствующей генеральной совокупности. При втором подходе исследователь имеет дело со своего рода одной, уникальной выборкой, а всякие априорные сведения о вероятностной природе исходных данных отсутствуют. В настоящей работе основное внимание уделено построению прогнозных моделей, поэтому она выполнена в рамках логико-алгебраического подхода.

МАТЕРИАЛЫ И МЕТОДЫ

Приведем краткое описание двухфакторной модели полносвязной линейной регрессии, подробно рассмотренной в работах [16, 17]. Пусть x_{i1}, x_{i2}, $i=\overline{1,n}$ — наблюдаемые значения объясняющих переменных x_1 и x_2, а y_i, $i=\overline{1,n}$ — наблюдаемые значения объясняемой переменной y. Предположим, что существуют «истинные» значения объясняющих переменных x_1 и x_2 — x_{i1}^*, x_{i2}^*, $i=\overline{1,n}$, которые связаны с наблюдаемыми значениями равенствами:

$$x_{i1}=x_{i1}^*+\varepsilon_i^{(x_1)},\ i=\overline{1,n}, \tag{1}$$

$$x_{i2}=x_{i2}^*+\varepsilon_i^{(x_2)},\ i=\overline{1,n}, \tag{2}$$

где $\varepsilon_i^{(x_1)}$, $\varepsilon_i^{(x_2)}$, $i=\overline{1,n}$ — ошибки измерения. Никаких априорных сведений об этих ошибках нет.

Допустим, что между «истинными» переменными x_1^* и x_2^* имеет место линейная функциональная зависимость:

$$x_{i1}^*=a+bx_{i2}^*,\ i=\overline{1,n}, \tag{3}$$

где a, b — неизвестные параметры.

Совокупность уравнений (1)–(3) представляет собой простейшую EIV модель — регрессию Деминга. Для её оценивания будем использовать метод наименьших полных квадратов, который подразумевает решение следующей оптимизационной задачи:

$$F\left(a,b,x_{12}^*,\ldots,x_{n2}^*\right)=\sum_{i=1}^{n}\left(x_{i1}-a-bx_{i2}^*\right)^2+$$

$$+\frac{1}{\lambda}\sum_{i=1}^{n}\left(x_{i2}-x_{i2}^*\right)^2\to\min, \tag{4}$$

где λ — некоторое положительное число, задающее тип расстояния от точек $\left(x_{i1},x_{i2}\right)$, $i=\overline{1,n}$ до прямой линии (3). При $\lambda\to 0$ имеем

минимизацию суммы квадратов вертикальных расстояний, при $\lambda \to 1$ — евклидовых расстояний, при $\lambda \to \infty$ — горизонтальных расстояний.

Используя необходимое условие экстремума для функции (4), можно получить квадратное уравнение относительно параметра b:

$$K_{x_1x_2} b^2 - \left(D_{x_1} - \frac{D_{x_2}}{\lambda} \right) b - \frac{K_{x_1x_2}}{\lambda} = 0, \qquad (5)$$

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

Минимуму функции (4) соответствует только один из корней уравнения (5):

$$\tilde{b} = \frac{D_{x_1} - \frac{D_{x_2}}{\lambda} + \sqrt{\left(D_{x_1} - \frac{D_{x_2}}{\lambda} \right)^2 + \frac{4K_{x_1x_2}^2}{\lambda}}}{2K_{x_1x_2}}. \qquad (6)$$

Тогда оценка параметра a находится по формуле

$$\tilde{a} = \overline{x_1} - \tilde{b}\overline{x_2}, \qquad (7)$$

а оценки «истинных» значений переменной x_2 по формулам

$$\tilde{x}_{i2}^* = \frac{-\tilde{a}\tilde{b} + \tilde{b}x_{i1} + \frac{1}{\lambda}x_{i2}}{\frac{1}{\lambda} + \tilde{b}^2}, \; i = \overline{1,n}. \qquad (8)$$

Введем модель парной линейной регрессии:

$$y_i = c_0 + c_1\tilde{x}_{i2}^* + \varepsilon_i, \; i = \overline{1,n}, \qquad (9)$$

где c_0, c_1 — неизвестные параметры, для поиска которых будем использовать МНК; ε_i, $i = \overline{1,n}$ — ошибки модели, означающие, что данная связь описывает процесс не точно, а с некоторой погрешностью.

Образованный синтез модели парной линейной регрессии (9) и простейшей EIV модели (1)–(3) называется двухфакторной моделью полносвязной линейной регрессии. Используя (8), можно представить модель (9) в виде:

$$y_i = c_0 + c_1\left(\frac{-\tilde{a}\tilde{b} + \tilde{b}x_{i1} + \frac{1}{\lambda}x_{i2}}{\frac{1}{\lambda} + \tilde{b}^2} \right) + \varepsilon_i, \; i = \overline{1,n}. \qquad (10)$$

Выражение (10) называется вторичной моделью полносвязной линейной регрессии.

Варьирование параметра λ в регрессии (1)–(3) приводит к изменению оценок, а, следовательно, и к изменению критериев адекватности модели (9). В работе [17] исследована задача выбора такого значения параметра λ регрессии (1)–(3), при котором коэффициент детерминации модели (9) R_y^2 максимален:

$$R_y^2(b,\lambda) \to \max, \text{ при условиях (5) и } \lambda > 0. \qquad (11)$$

Результаты исследования были сформулированы в виде следующей теоремы.

Теорема. *Пусть для полносвязной регрессии (1)–(3), (9) найдены числа* $A = D_{x_2}K_{x_1y} - K_{x_2y}K_{x_1x_2}$, $B = D_{x_1}K_{x_2y} - K_{x_1y}K_{x_1x_2}$, *причем,* $A \neq 0$, $B \neq 0$, $K_{x_1x_2} \neq 0$, $K_{x_1y} \neq 0$, $K_{x_2y} \neq 0$. *Пусть* Small, Large *— малое и большое положительные числа. Тогда задача (11) всегда имеет единственное решение, причем:*

(1) если $ABK_{x_1x_2} > 0$, *то в точке* $\tilde{b} = \frac{D_{x_1}A + K_{x_1x_2}B}{D_{x_2}B + K_{x_1x_2}A}$, $\lambda = \frac{K_{x_1x_2}A^2 + D_{x_2}AB}{K_{x_1x_2}B^2 + D_{x_1}AB}$;

(2) если $ABK_{x_1x_2} < 0$, *то:*

при $-\frac{B}{A}K_{x_1x_2} < \frac{K_{x_1x_2}^2}{D_{x_2}}$ *в точке* $\tilde{b} = \frac{D_{x_1}}{K_{x_1x_2}}$, $\lambda = \text{Large}$;

при $\frac{K_{x_1x_2}^2}{D_{x_2}} < -\frac{B}{A}K_{x_1x_2} < D_{x_1}$ *либо в точке* $\tilde{b} = \frac{D_{x_1}}{K_{x_1x_2}}$, $\lambda = \text{Large}$, *либо в точке* $\tilde{b} = \frac{K_{x_1x_2}}{D_{x_2}}$, $\lambda = \text{Small}$;

при $-\frac{B}{A}K_{x_1x_2} > D_{x_1}$ *в точке* $\tilde{b} = \frac{K_{x_1x_2}}{D_{x_2}}$, $\lambda = \text{Small}$.

Из этой теоремы следует, что значение коэффициента детерминации R_y^2 вторичной модели (10) полносвязной регрессии будет наибольшим либо когда (10) принимает вид двухфакторной множественной линейной регрессии при $\lambda = \frac{K_{x_1x_2}A^2 + D_{x_2}AB}{K_{x_1x_2}B^2 + D_{x_1}AB}$, либо вид наилучшей однофакторной парной линейной регрессии при $\lambda = \text{Small}$ или при $\lambda = \text{Large}$. В последнем случае вторичная модель (10) обязательно утрачивает свои аппроксимационные качества. Для того чтобы избежать такой утраты достаточно отказаться от ограничения $\lambda > 0$ в задаче (11). Тогда решением задачи (11) без ограничения $\lambda > 0$ всегда будет

точка $\tilde{b} = \dfrac{D_{x_1} A + K_{x_1x_2} B}{D_{x_2} B + K_{x_1x_2} A}$, $\lambda = \dfrac{K_{x_1x_2} A^2 + D_{x_2} AB}{K_{x_1x_2} B^2 + D_{x_1} AB}$, в которой оцененная модель (10) будет иметь вид двухфакторной множественной регрессии, оцененной с помощью МНК. Это обстоятельство можно использовать следующим образом: попытаться вблизи точки $\lambda = \dfrac{K_{x_1x_2} A^2 + D_{x_2} AB}{K_{x_1x_2} B^2 + D_{x_1} AB}$ отыскать такие оценки вторичного уравнения полносвязной регрессии, для которых значение её агрегированного критерия адекватности будет лучше, чем его значение для множественной модели.

В качестве агрегированного критерия адекватности модели (9) можно использовать следующий показатель:

$$S = w_1 K_1 + w_2 K_2 + w_3 K_3 + w_4 K_4, \qquad (12)$$

где w_1, w_2, w_3, w_4 — некоторые положительные весовые коэффициенты, которые, в случае отсутствия каких-либо приоритетов, можно задать равными; K_1, K_2, K_3, K_4 — нормированные аналоги коэффициента детерминации R^2, критерия Дарбина — Уотсона DW, согласованности поведения SP и средней относительной ошибки аппроксимации E соответственно, которые находятся по формулам:

$$K_1 = 1 - R^2 = \sum_{i=1}^{n} \varepsilon_i^2 \Big/ \sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2, \qquad (13)$$

$$K_2 = 0{,}5\left|2 - DW\right| = 0{,}5\left|2 - \sum_{i=2}^{n} \left(\varepsilon_i - \varepsilon_{i-1}\right)^2 \Big/ \sum_{i=1}^{n} \varepsilon_i^2\right|, \qquad (14)$$

$$K_3 = 0{,}5\left(1 - SP/\left(n-1\right)\right) = \frac{1}{2} - \frac{\sum_{i=1}^{n-1} \operatorname{sign}(y_{i+1} - y_i) \cdot \operatorname{sign}(y_{i+1} - y_i + \varepsilon_i - \varepsilon_{i+1})}{2\left(n-1\right)}, \qquad (15)$$

$$K_4 = 0{,}01E = \left(1/n\right) \sum_{i=1}^{n} \left|\varepsilon_i / y_i\right|, \qquad (16)$$

где ε_i, $i = \overline{1,n}$ — ошибки модели (9) для заданного значения λ.

Подробное описание хорошо известных критериев R^2, DW, SP и E можно найти, например, в работе [18]. Отметим, что идеальным значением для R^2 является 1, для DW — 2, для SP — $n-1$, для E — 0. Область значений каждого из критериев K_1, K_2 и K_3 лежит в интервале от 0 до 1. При этом, чем ближе значение K_1, K_2 или K_3 к 0, тем выше качество модели. Для критерия K_4 наилучшим значением также является 0, однако область его возможных значений не ограничена сверху. Идеальным значением критерия S является 0.

Тогда сформулируем следующую задачу. Пусть требуется выбрать такое значение параметра λ модели (9), для которого

$$S = w_1 K_1 + w_2 K_2 + w_3 K_3 + w_4 K_4 \to \min. \qquad (17)$$

При этом откажемся от ограничения $\lambda > 0$.

Формализуем поставленную задачу в виде задачи математического программирования. Первый этап построения полносвязной регрессии предполагает сначала решение квадратного уравнения (5), из которого находится оценка $\tilde{b}$, затем по формуле (7) определяется $\tilde{a}$, и, наконец, по формулам (8) находятся значения оценки $\tilde{x}_{i2}^*$, $i = \overline{1,n}$. Второй этап построения предполагает МНК-оценивание регрессии (9), заключающееся в решении следующей системы линейных уравнений:

$$\begin{cases} n\tilde{c}_0 + \tilde{c}_1 \sum_{i=1}^{n} \tilde{x}_{i2}^* = \sum_{i=1}^{n} y_i, \\ \tilde{c}_0 \sum_{i=1}^{n} \tilde{x}_{i2}^* + \tilde{c}_1 \sum_{i=1}^{n} \left(\tilde{x}_{i2}^*\right)^2 = \sum_{i=1}^{n} y_i \tilde{x}_{i2}^*. \end{cases} \qquad (18)$$

Ошибки регрессии (9) находятся по формулам:

$$\varepsilon_i = y_i - \left(\tilde{c}_0 + \tilde{c}_1 \tilde{x}_{i2}^*\right), \; i = \overline{1,n}. \qquad (19)$$

Тогда решение задачи нелинейного программирования (17) с ограничениями (13)–(16), (5), (7), (8), (18), (19) дает параметр λ, для которого суммарные качественные характеристики вторичной регрессии (9) оптимальны по критерию S.

Понятно, что если в целевой функции (17) положить $w_1 = 1$, $w_2 = w_3 = w_4 = 0$, то сформулированная задача эквивалентна задаче (11) без ограничения $\lambda > 0$, поэтому её решением будут МНК-оценки множественной регрессии.

Для точного решения задачи нелинейного программирования (17), (13)–(16), (5), (7), (8), (18), (19) можно воспользоваться любым современным оптимизационным программным

обеспечением, например, имеющей Web-интерфейс программой APMonitor. Вместе с тем существует возможность получения приближенного решения данной задачи. Для этого достаточно задать область изменения параметра λ (желательно вблизи точки $\lambda = \frac{K_{x_1x_2}A^2 + D_{x_2}AB}{K_{x_1x_2}B^2 + D_{x_1}AB}$, чтобы существенно не утратить аппроксимационное качество модели), разбить её некоторым множеством точек и выбрать наилучшую из них с точки зрения критерия (17). Главной проблемой при этом является то, что зависимость (5) между оценкой $\tilde{b}$ и параметром λ не является однозначной.

Действительно, покажем, что для любого $\lambda \neq 0$, квадратное уравнение (5) всегда имеет 2 корня. Его дискриминант имеет вид:

$$D = \left(D_{x_1} - \frac{D_{x_2}}{\lambda}\right)^2 + 4\frac{K_{x_1x_2}^2}{\lambda}. \qquad (20)$$

Очевидно, что при $\lambda > 0$ дискриминант (20) всегда положителен.

Преобразуем выражение (20):

$$\left(D_{x_1} - \frac{D_{x_2}}{\lambda}\right)^2 + 4\frac{K_{x_1x_2}^2}{\lambda} =$$

$$= D_{x_1}^2 - 2\frac{D_{x_1}D_{x_2}}{\lambda} + \frac{D_{x_2}^2}{\lambda^2} +$$

$$+4\frac{K_{x_1x_2}^2}{\lambda} + 4\frac{D_{x_1}D_{x_2}}{\lambda} - 4\frac{D_{x_1}D_{x_2}}{\lambda} =$$

$$= \left(D_{x_1} + \frac{D_{x_2}}{\lambda}\right)^2 - 4\frac{D_{x_1}D_{x_2} - K_{x_1x_2}^2}{\lambda}.$$

Поскольку в полученном выражении $D_{x_1}D_{x_2} - K_{x_1x_2}^2 \geq 0$ (неравенство Коши — Буняковского), то можно сделать вывод, что при $\lambda < 0$ дискриминант (20) также положителен. Поэтому квадратное уравнение (5) при $\lambda \neq 0$ всегда имеет два корня:

$$\tilde{b}_1 = \frac{D_{x_1} - \frac{D_{x_2}}{\lambda} + \sqrt{\left(D_{x_1} - \frac{D_{x_2}}{\lambda}\right)^2 + \frac{4K_{x_1x_2}^2}{\lambda}}}{2K_{x_1x_2}},$$

$$\tilde{b}_2 = \frac{D_{x_1} - \frac{D_{x_2}}{\lambda} - \sqrt{\left(D_{x_1} - \frac{D_{x_2}}{\lambda}\right)^2 + \frac{4K_{x_1x_2}^2}{\lambda}}}{2K_{x_1x_2}}.$$

Это означает, что решая задачу (17), (13)–(16), (5), (7), (8), (18), (19) приближенно, необходимо в каждой выбранной точке области изменения λ определить значения агрегированного критерия (12) в зависимости и от оценки $\tilde{b}_1$, и от оценки $\tilde{b}_2$, из которых взять наилучшую с точки зрения условия (17).

РЕЗУЛЬТАТЫ И ИХ ОБСУЖДЕНИЕ

ВВП — это один из основных макроэкономических показателей России [19–21]. 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 моделирования ВВП России на официальном сайте Федеральной службы государственной статистики были собраны годовые данные 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.).

Тот факт, что переменная y действительно связана с переменными x_1 и x_2, подтверждается известной формулой И. Фишера:

$$P \cdot Q = M \cdot V,$$

где P — уровень цен; Q — объем производства; M — объем денежной массы; V — скорость обращения денежной массы.

Оцененная по этим данным с помощью МНК двухфакторная модель множественной регрессии имеет вид:

$$\tilde{y} = 18647{,}8 - 1{,}3286x_1 + 2{,}4897x_2. \qquad (21)$$

Критерии адекватности модели (21):

$$R^2 = 0{,}9924,\ DW = 1{,}6764,\ SP = 10,$$
$$E = 2{,}966\ \%,\ S = 0{,}2823.$$

Отметим, что из-за сильной корреляции между переменными x_1 и x_2 возник эффект мультиколлинеарности, который исказил знак коэффициента при переменной x_1, который по смыслу должен быть с плюсом. Достоинством полносвязной регрессии (1)–(3), (9)

является именно то, что её вторичное уравнение (10) в этом случае дало бы согласующиеся с коэффициентами корреляции знаки оценок при переменных x_1 и x_2. Далее рассматривается задача, связанная только лишь с повышением суммарных качественных характеристик множественной регрессии (21) по критерию S.

Сначала была задана область изменения параметра $\lambda \in [-50,50]$. Затем этот промежуток был равномерно разбит точками с шагом 0,5, причем, $\lambda \neq 0$. После чего в каждой точке интервала были определены оценки $\tilde{b}_1$ и $\tilde{b}_2$, для каждой из которых были вычислены значения критериев адекватности R_1^2, DW_1, SP_1, E_1, S_1 и R_2^2, DW_2, SP_2, E_2, S_2. Графики зависимостей этих критериев от параметра λ приведены на рис. 1–5.

По графикам на рис. 1 видно, что при $w_1 = 1$, $w_2 = w_3 = w_4 = 0$ точное решение задачи (17), (13)–(16), (5), (7), (8), (18), (19) достигается в точке максимума, в которой оценка параметра $b = \tilde{b}_2$. В этой точке, согласно приведенной выше теореме, $\lambda = -1,8679$, а вторичное уравнение полносвязной регрессии имеет вид множественной модели (21).

По графикам на рис. 2 видно, что при $w_2 = 1$, $w_1 = w_3 = w_4 = 0$ приближенное решение задачи достигается в двух точках $\lambda = -5,5$ и $\lambda = -3,5$, в которых оценка параметра $b = \tilde{b}_2$. Из этих двух точек лучше выбрать последнюю, поскольку в ней выше значение коэффициента детерминации. Тогда вторичное уравнение полносвязной регрессии при $\lambda = -3,5$ имеет вид:

$$\tilde{y} = 24057,81 - 2,9279x_1 + 2,9439x_2, \quad (22)$$

Критерии адекватности модели (22):

$$R^2 = 0,9913,\ DW = 2,02,\ SP = 10,$$
$$E = 3,1298,\ S = 0,1334.$$

Как видно, аппроксимационное качество модели (22) несколько хуже, чем для множественной регрессии (21). Но при этом, судя по значению $DW = 2,02$, практически полностью устранена автокорреляция ошибок.

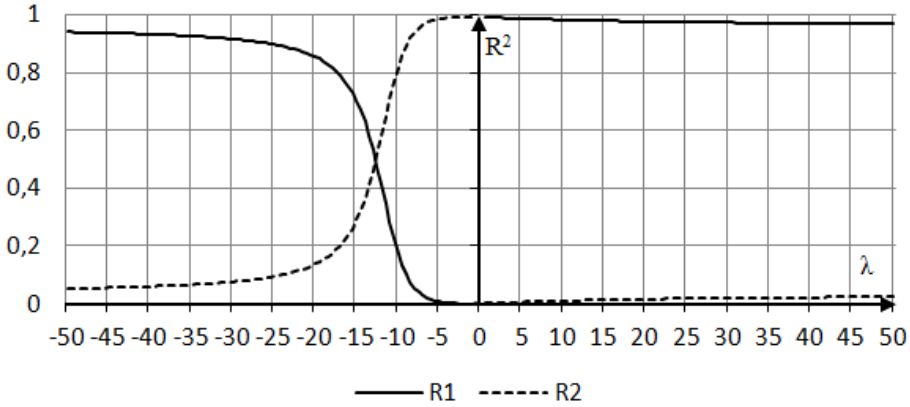

Рис. 1. Зависимости R_1^2 и R_2^2 от λ
[Fig. 1. Dependences of R_1^2 and R_2^2 from λ]

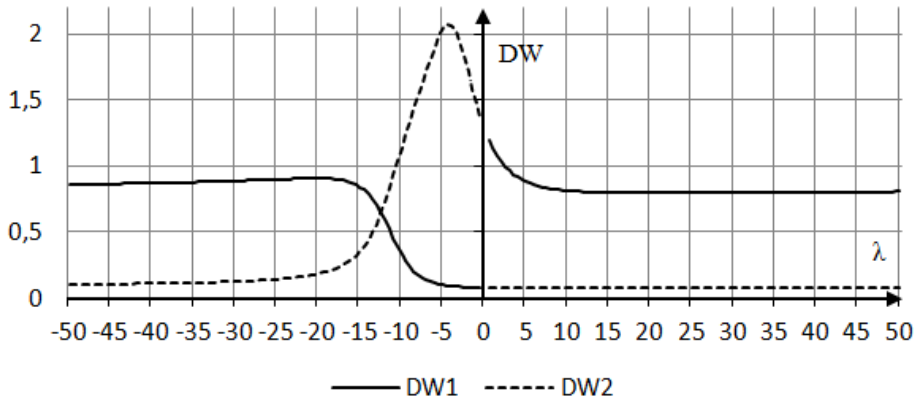

Рис. 2. Зависимости DW_1 и DW_2 от λ
[Fig. 2. Dependences of DW_1 and DW_2 from λ]

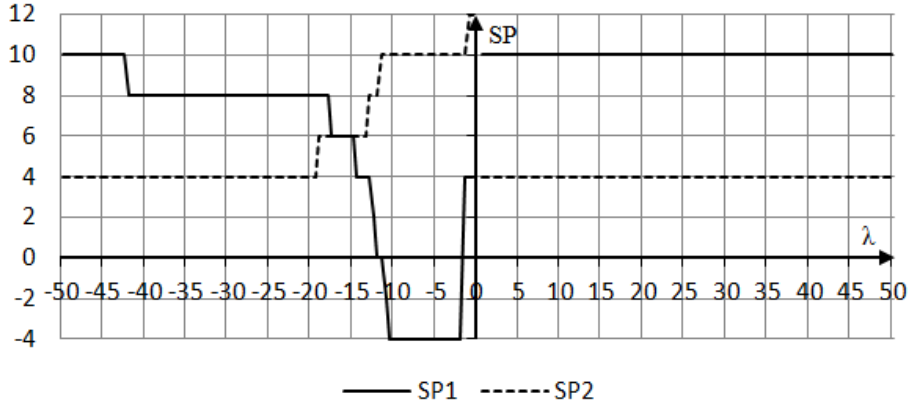

Рис. 3. Зависимости SP_1 и SP_2 от λ
[Fig. 3. Dependences of SP_1 and SP_2 from λ]

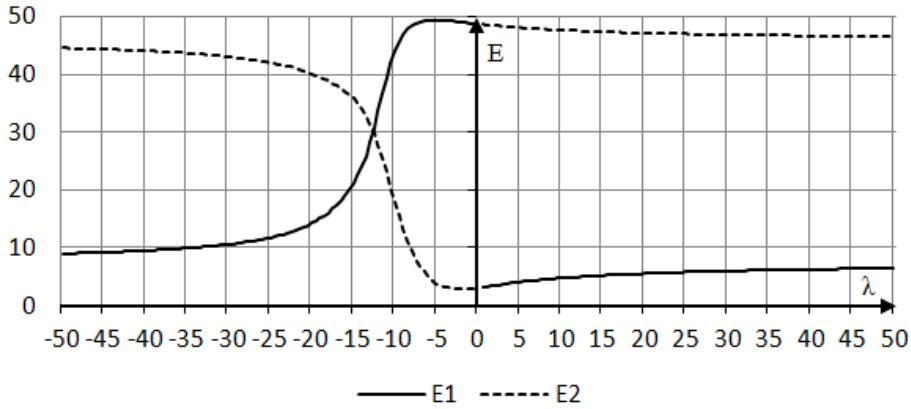

Рис. 4. Зависимости E_1 и E_2 от λ
[Fig. 4. Dependences of E_1 and E_2 from λ]

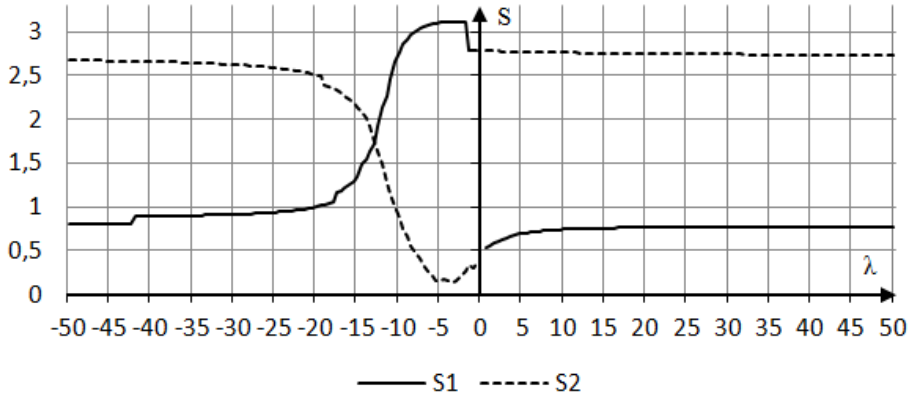

Рис. 5. Зависимости S_1 и S_2 от λ
[Fig. 5. Dependences of S_1 and S_2 from λ]

По графикам на рис. 3 видно, что при $w_3 = 1$, $w_1 = w_2 = w_4 = 0$ приближенное решение задачи достигается в точках из интервала $\lambda \in [-1; -0,5]$, в которых оценка параметра $b = b_2$. Из этих точек лучше выбрать $\lambda = -1$, поскольку в ней выше значение коэффициента детерминации. В этой точке вторичное уравнение полносвязной регрессии имеет вид:

$$\tilde{y} = 16409,38 - 0,6579x_1 + 2,2976x_2, \quad (23)$$

Критерии адекватности модели (23):

$$R^2 = 0,9923,\ DW = 1,47,\ SP = 12,$$

$$E = 2,9428,\ S = 0,302.$$

Аппроксимационное качество модели (23) вновь несколько хуже, чем для множественной регрессии (21). Но при этом, судя по ве-

личине $SP = 12$, обеспечена полная согласованность поведения фактической и расчетной траекторий изменения значений переменной y.

По графикам на рис. 4 видно, что при $w_4 = 1,\ w_1 = w_2 = w_3 = 0$ приближенное решение задачи достигается в точке $\lambda = -1$, в которой оценка параметра $b = \tilde{b}_2$. В этой точке вторичное уравнение полносвязной регрессии имеет вид (23).

По графикам на рис. 5 видно, что приближенное решение четырехкритериальной задачи при $w_1 = w_2 = w_3 = w_4 = 1$ достигается в точке $\lambda = -3,5$, в которой оценка параметра $b = \tilde{b}_2$. В этой точке полносвязная регрессия имеет вид (22), а значение её агрегированного критерия $S = 0,1334$ меньше, чем его величина $S = 0,2823$ для множественной регрессии.

ЗАКЛЮЧЕНИЕ

В данной работе исследована и на примере моделирования ВВП доказана целесообразность применения двухфакторной полносвязной регрессии в качестве инструмента для улучшения верифицируемой по нескольким критериям адекватности двухфакторной множественной модели. При этом полученная в результате модель ВВП оказалась по агрегированному критерию лучше классической множественной регрессии более чем в 2 раза. Предложенную методику можно также использовать как инструмент для борьбы с автокорреляцией ошибок и для повышения согласованности поведения фактической и расчетной траекторий изменения значений объясняемой переменной.

Результаты данной работы в дальнейшем будут использованы при реализации методики многокритериального выбора регрессионных моделей, известной в отечественной литературе, как «конкурс» моделей.

КОНФЛИКТ ИНТЕРЕСОВ

Авторы декларируют отсутствие явных и потенциальных конфликтов интересов, связанных с публикацией настоящей статьи.

СПИСОК ЛИТЕРАТУРЫ

1. *Montgomery, D. C.* Introduction to linear regression analysis / D. C. Montgomery, E. A. Peck, G. G. Vining. – Wiley, 2012. – 672 p.

2. *Kuhn, M.* Applied predictive modeling / M. Kuhn, K. Johnson. – Springer, 2018. – 600 p.

3. *Harrell, Jr.* Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis / Jr. Harrell, E. Frank. – Springer Series in Statistics, 2015. – 582 p.

4. *Gillard, J.* An overview of linear structural models in errors in variables regression / J. Gillard // REVSTAT – Statistical Journal. – 2010. – V. 8, No. 1. – P. 57–80.

5. *Leng, L.* Compound regression and constrained regression: nonparametric regression frameworks for EIV models / L. Leng, W. Zhu // The American Statistician. – 2019. – DOI: 10.1080/00031305.2018.1556734.

6. *Golubev, Yu.* A minimax approach to errors-in-variables linear models / Yu. Golubev // Mathematical methods of statistics. – 2018. – V. 27. – P. 205–225.

7. *Deming, W. E.* Statistical adjustment of data / W. E. Deming. – Wiley, 1943. – 273 p.

8. Skyline performs as well as vendor software in the quantitative analysis of serum 25-hydroxy vitamin D and vitamin D binding globulin / C. M. Henderson [et al] // Clinical Chemistry. – 2018. – V. 64. – P. 408–410.

9. Determination of recent growth hormone abuse using a single dried blood spot / G. Reverter-Branchat et al // Clinical Chemistry. – 2016. – V. 62. – P. 1353–1360.

10. *Chen, H.-Y.* Alternative errors-in-variables models and their applications in finance research / H.-Y. Chen, A.C. Lee, C.-F. Lee // The quarterly review of economics and finance. – 2015. – V. 58. – P. 213–227.

11. Building growth and value hybrid valuation model with errors-in-variables regression / D. Kong [et al] // Applied economics letters. – 2019. – Vol. 26. – P. 370–386.

12. Bayesian inference for the errors-in-variables model / X. Fang [et al] // Studia Geophysica et Geodaetica. – 2017. – V. 61. – P. 35–52.

13. *Wu, Y.* Comparison of total least squares and least squares for four- and seven-parameter model coordinate transformation / Y. Wu, J. Liu, H. Y. Ge // Journal of applied geodesy. – 2016. – V. 10. – P. 259–266.

14. *Awange, J. L.* EIV models and Pareto optimality / J. L. Awange, B. Palancz // Geospatial algebraic computations. – 2016. – P. 155–202.

15. Climate-related variability and stock-recruitment relationship of the North Pacific albacore tuna / A. A. Singh [et al] // Polish journal of natural sciences. – 2018. – V. 33. – P. 131–154.

16. *Базилевский, М. П.* Синтез модели парной линейной регрессии и простейшей EIV-модели / М. П. Базилевский // Моделирование, оптимизация и информационные технологии. – 2019. – Т. 7, № 1 (24). – С. 170–182.

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

18. *Носков, С. И.* Построение регрессионных моделей с использованием аппарата линейно-булевого программирования / С. И. Носков, М. П. Базилевский. – Иркутск : ИрГУПС, 2018. – 176 с.

19. *Aivazian, S. A.* Macroeconomic modeling of the Russian economy / S. A. Aivazian, A. N. Bereznyatsky, B. E. Brodsky // Applied Econometrics. – 2017. – Vol. 47. – P. 5–27.

20. *Кирилюк, И. Л.* Модели производственных функций для российской экономики / И. Л. Кирилюк // Компьютерные исследования и моделирование. – 2013. – Т. 5, № 2. – С. 293–312.

21. *Лычагина, Т. А.* Применение аппарата производственных функций для анализа влияния состояния основных фондов на экономический рост РФ / Т. А. Лычагина, Е. А. Пахомова, Д. А. Писарева // Национальные интересы: приоритеты и безопасность. – 2016. – С. 4–19.

Базилевский Михаил Павлович – канд. техн. наук, доцент, доцент, Иркутский государственный университет путей сообщения.
E-mail: mik2178@yandex.ru
ORCID iD: https://orcid.org/0000-0002-3253-5697

DOI: https://doi.org/10.17308/sait.2020.1/2596 ISSN 1995-5499
Received 09.11.2019
Accepted 15.03.2020

MULTI-CRITERIA APPROACH TO THE CONSTRUCTION OF FULLY CONNECTED TWO-FACTOR REGRESSIONS BASED ON THE MODELLING OF THE GDP OF RUSSIA

© 2020 M. P. Bazilevskiy✉

Irkutsk State Transport University
15, Chernyshevskogo Str., 664074 Irkutsk, Russian Federation

Abstract. CToday, most regression models are based on the assumption that explanatory variables are error free. Although a powerful mathematical apparatus has been developed for regression models with errors in explanatory variables, better known as errors-in-variables models, these models are hardly ever used. The developed mathematical apparatus includes linear regression models. The aim of this paper was to study the possibility of using a two-factor fully connected regression as a tool for improving a two-factor multiple model verified by several adequacy criteria.

✉ Bazilevskiy Mikhail P.
e-mail: mik2178@yandex.ru

The article gives a brief description of fully connected two-factor regressions. To assess the overall quality of the regression models an aggregated criterion is suggested, which is a linear combination of four well-known adequacy criteria. Based on this criterion, the problem of choosing the optimal estimates of the secondary equation of the fully connected regression is formulated. This problem was formalized as a mathematical programming problem. An approximate algorithm for solving this problem was developed. The suggested algorithm was used to create regression models of Russia's GDP under various conditions. The resulting GDP model appeared to be more than 2 times better than classical multiple regression according to the aggregated criterion. The technique proposed in this paper can serve as a tool for combating the autocorrelation of errors. It can also be used to increase the consistency between the actual and calculated trajectories of changes in the values of the dependent variable.
Keywords: multiple regression; fully connected regression; errors-in-variables model; Deming regression; adequacy criteria; autocorrelation of errors; GDP of Russia.

CONFLICT OF INTEREST

The authors declare the absence of obvious and potential conflicts of interest related to the publication of this article.

REFERENCES

1. *Montgomery D. C., Peck E. A., Vining G. G.* Introduction to linear regression analysis. Wiley, 2012. 672 p.

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

3. *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.

4. *Gillard J.* An overview of linear structural models in errors in variables regression. REVSTAT – Statistical Journal. 2010. V. 8, No. 1. P. 57–80 available at https://www.ine.pt/revstat/pdf/rs100104.pdf.

5. *Leng L., Zhu W.* Compound regression and constrained regression: nonparametric regression frameworks for EIV model. The American Statistician. 2019 available at https://doi.org/10.1080/00031305.2018.1556734.

6. *Golubev Yu.* A minimax approach to errors-in-variables linear models. Mathematical methods of statistics. 2018. V. 27. P. 205–225 available at https://doi.org/10.3103/S1066530718030031.

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

8. *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. V. 64. P. 408–410 available at https://doi.org/10.1373/clinchem.2017.282293.

9. *Reverter-Branchat G., Bosch J., Vall J., Farre M., Papaseit E., Pichini S., Segura J.* Determination of recent growth hormone abuse using a single dried blood spot. Clinical Chemistry. 2016. V. 62. P. 1353–1360 available at https://doi.org/10.1373/clinchem.2016.257592.

10. *Chen H.-Y., Lee A. C., Lee C.-F.* Alternative errors-in-variables models and their applications in finance research. The quarterly review of economics and finance. 2015. V. 58. P. 213–227 available at https://ah.lib.nccu.edu.tw/bitstream/140.119/100619/1/407185.pdf.

11. *Kong D., Lin C.-P., Yeh I-C., Chang W.* Building growth and value hybrid valuation model with errors-in-variables regression. Applied economics letters. 2019. V. 26. P. 370–386 available at https://doi.org/10.1080/13504851.2018.1486005.

12. *Fang X., Li B., Alkhatib H., Zeng W., Yao Y.* Bayesian inference for the errors-in-variables model. Studia Geophysica et Geodaetica. 2017. V. 61. P. 35–52 available at https://doi.org/10.1007/s11200-015-6107-9.

13. *Wu Y., Liu J., Ge H. Y.* Comparison of total least squares and least squares for four- and seven-parameter model coordinate transformation. Journal of applied geodesy. 2016. V. 10. P. 259–266 available at https://doi.org/10.1515/jag-2016-0015.

14. *Awange J. L., Palancz B.* EIV models and Pareto optimality. Geospatial algebraic compu-

tations. 2016. P. 155–202 available at https://doi.org/10.1007/978-3-319-25465-4_10.

15. *Singh A. A., Sakuramoto K., Suzuki N., Alok K.* Climate-related variability and stock-recruitment relationship of the North Pacific albacore tuna. Polish journal of natural sciences. 2018. V. 33. P. 131–154 available at http://www.uwm.edu.pl/polish-journal/sites/default/files/issues/articles/singh_et_al_2018.pdf.

16. *Bazilevskiy M. P.* Synthesis of the paired linear regression model and the simplest EIV model. Modeling, optimization and information technology. 2019. V. 24, No. 1. P. 170–182 available at https://moit.vivt.ru/wp-content/uploads/2019/01/Bazilevskiy_1_19_1.pdf.

17. *Bazilevskiy M. P.* Research of a two-factor fully connected linear regression model. Modeling, optimization and information technology. 2019. V. 25, No. 2. P. 80–96 available at https://moit.vivt.ru/wp-content/uploads/2019/05/Bazilevskiy_2_19_1.pdf.

18. *Noskov S. I., Bazilevskiy M. P.* Construction of regression models using linear Boolean programming. IrGUPS. 2018. 176 p.

19. *Aivazian S. A., Bereznyatsky A. N., Brodsky B. E.* Macroeconomic modeling of the Russian economy. Applied Econometrics. 2017. V. 47. P. 5–27.

20. *Kirilyuk I. L.* Models of production functions for the Russian economy. Computer Research and Modeling. 2013. V. 5, No. 2. P. 293–312.

21. *Lychagina T. A., Pakhomova E. A., Pisareva D. A.* The application of production functions to analyze the effect of fixed assets on economic growth in the Russian Federation. National Interests: Priorities and Security, 2016. P. 4–19.

Bazilevskiy Mikhail P. — PhD in Technical Sciences, Associate Professor, Irkutsk State Transport University.
E-mail: mik2178@yandex.ru
ORCID iD: https://orcid.org/0000-0002-3253-5697