УДК 519.862.6 doi: 10.26102/2310-6018/2019.24.1.033

М. П. Базилевский

СИНТЕЗ МОДЕЛИ ПАРНОЙ ЛИНЕЙНОЙ РЕГРЕССИИ И ПРОСТЕЙШЕЙ EIV-МОДЕЛИ

Иркутский государственный университет путей сообщения, Иркутск, Россия

Данная работа посвящена синтезу модели парной линейной регрессии и простейшей EiV-модели (Errors-In-Variables model), более известной как регрессия Деминга. EIV-модель – это регрессия, в которой все переменные содержат случайные ошибки. Такие модели имеют ряд существенных недостатков, что затрудняет работу с ними. Предлагаемый в работе синтез, названный двухфакторной моделью полносвязной линейной регрессии, не только лишен этих недостатков, но и имеет определенные достоинства. Рассмотрены основные этапы построения и анализа двухфакторных моделей полносвязной линейной регрессии. Предложенная модель полносвязной линейной регрессии имеет много общего с классической моделью множественной регрессии, однако в основе этих двух видов лежат совершенно разные подходы. Если множественная регрессия строится по принципу «независимые переменные влияют на зависимую», то принципом полносвязной регрессии является «все переменные влияют друг на друга». Установлено, что аппроксимационные способности полносвязных моделей не превосходят способностей множественных регрессий, но зато первые имеют гораздо более разнообразную интерпретацию. Разработанный синтез можно использовать при построении множественных моделей как инструмент для решения задач снижения размерности данных, устранения мультиколлинеарности и отбора информативных регрессоров.

Ключевые слова: регрессионная модель, метод наименьших квадратов, метод наименьших полных квадратов, регрессия Деминга, EIV-модель, модель полносвязной линейной регрессии.

Введение. Часто при проведении анализа данных возникает задача выявления неизвестной статистической зависимости между объясняемой (зависимой) переменной и одной или несколькими объясняющими (независимыми) переменными. Это классическая задача регрессионного анализа, которая решается с помощью оценивания неизвестных параметров специальным образом выбранной регрессионной модели. Для нахождения оценок параметров модели регрессии разработано немало методов (см., например, [1–3]), однако большинство исследователей отдают предпочтение исключительно методу наименьших квадратов (МНК). Причиной такой популярности МНК, безусловно, является его вычислительная простота по сравнению с другими известными методами оценивания.

Одной из предпосылок МНК является то, что значения независимых переменных должны быть неслучайными. При этом считается, что объясняемая переменная содержит некоторого рода случайные ошибки. Наличие случайных ошибок в регрессионной модели означает, например, что при измерении значений зависимой переменной были допущены неточности. Однако такие погрешности могут быть допущены и при регистрации независимых переменных, а значит, они также могут быть случайными на практике. Найденные с помощью МНК оценки регрессионных моделей с ошибками в объясняющих переменных могут обладать такими негативными свойствами, как смещенность, несостоятельность и неэффективность.

Первые упоминания о регрессионных моделях с ошибками в объясняющих переменных, известных в настоящее время в англоязычных источниках как «Errors-In-Variables models» (EIV-модели), можно найти в работах Р. Эдкока и К. Куммеля, опубликованных еще в 1877 – 1879 гг. Однако эти идеи оставались незамеченными более 50 лет, пока их в 1937 г. не возродил Т. Купманс и позднее в 1943 г. не развил В. Деминг.

Регрессия Деминга – простейший случай EIV-модели, которая содержит только одну независимую переменную. Такие модели достаточно хорошо изучены, а их подробное описание можно найти в работах [4, 5]. Множественная EIV-модель рассмотрена в [6]. К сожалению, приходиться констатировать, что в настоящее время регрессионные модели с ошибками в независимых переменных практического применения почти не находят. Так, например, при поиске в научной электронной библиотеке «Elibrary.ru», по ключевым словам, «регрессия Деминга» было найдено всего лишь 3 работы [7–9] прикладного характера. В зарубежной литературе это направление чуть более популяризировано и особенно актуально при обработке медицинских данных. Так, регрессия Деминга в клинических лабораториях служит прекрасным инструментом для численного сопоставления новых измерительных методов с существующими. Множество статей по данной тематике можно найти на сайте журнала Clinical Chemistry [10].

Отсутствие интереса у исследователей к EIV-моделям можно объяснить следующим образом. Для чего вообще предназначены регрессионные модели? Во-первых, такие модели можно интерпретировать, например, анализируя полученные оценки параметров регрессии. Во-вторых, они служат прекрасным инструментом для прогнозирования зависимой переменной. В EIV-моделях, как показывает практика, приходиться работать с так называемыми «истинными»

значениями независимых переменных. Но поскольку эти «истинные» значения никогда априори не известны исследователю, то несмотря даже на возможность интерпретации EIV-моделей, прогноз по ним и вовсе осуществить нельзя, что является их значительным недостатком. При этом интерпретация тоже может быть ошибочной, поскольку в EIV-моделях независимые переменные содержат ошибки, а значит, не выполняются условия теоремы Гаусса-Маркова, следовательно, оценки параметров регрессии могут стать смещенными. Для решения этой проблемы необходимо точно знать дисперсии ошибок объясняющих переменных, что также не всегда возможно. Кроме того, для оценивания множественных EIV-моделей необходимо привлекать специальные численные методы. Все эти многочисленные обстоятельства справедливо вызывают негативное отношение исследователей к EIV-моделям.

Вместе с тем нельзя не согласиться, что идея построения регрессионных моделей с ошибками в независимых переменных является перспективной и требует дальнейшего развития. Данная работа посвящена слиянию воедино классических моделей парной линейной регрессии и простейших EIV-моделей.

Синтез моделей. Пусть исследователь постулирует наличие зависимости между объясняемой переменной y и двумя объясняющими переменными x_1 и x_2. Предположим, что объясняющие переменные не содержат случайных ошибок. В этом случае для получения статистической зависимости между переменными можно воспользоваться классической двухфакторной регрессионной моделью:

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

где y_i, $i = \overline{1,n}$ – значения зависимой переменной; x_{i1}, x_{i2}, $i = \overline{1,n}$ – значения независимых переменных; ε_i, $i = \overline{1,n}$ – случайные ошибки; α_0, α_1, α_2 – неизвестные параметры модели; n – число наблюдений.

Если объясняющие переменные x_1 и x_2 содержат ошибки, что на практике распространено довольно часто, то применять модель (1) уже не представляется возможным. Для такой ситуации разработана схема, подробно рассмотренная в работе [6]. Предполагается, что существуют «истинные» (расчетные мо модели) значения переменных y, x_1 и x_2, которые обозначим y_i^*, x_{i1}^*, x_{i2}^*, $i = \overline{1,n}$. Между этими переменными существует линейная функциональная зависимость:

$$y_i^* = \beta_0 + \beta_1 x_{i1}^* + \beta_2 x_{i2}^*, \qquad i = \overline{1,n}, \tag{2}$$

где β_0, β_1, β_2 – неизвестные параметры.

«Истинные» значения переменных y, x_1 и x_2 пока также неизвестны, но зато известны их наблюдаемые значения, которые отличаются от первых на случайные отклонения:

$$y_i = y_i^* + \varepsilon_{i0}, \qquad i = \overline{1,n}, \tag{3}$$

$$x_{i1} = x_{i1}^* + \varepsilon_{i1}, \qquad i = \overline{1,n}, \tag{4}$$

$$x_{i2} = x_{i2}^* + \varepsilon_{i2}, \qquad i = \overline{1,n}. \tag{5}$$

Совокупность уравнений (2) – (5) будем называть двухфакторной линейной EIV-моделью. Для её оценивания можно воспользоваться методом наименьших полных квадратов (МНПК), который предполагает решение оптимизационной задачи:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \beta_1 x_{i1}^* - \beta_2 x_{i2}^*\right)^2 + \frac{1}{\lambda_1}\sum_{i=1}^{n}\left(x_{i1} - x_{i1}^*\right)^2 + \frac{1}{\lambda_2}\sum_{i=1}^{n}\left(x_{i2} - x_{i2}^*\right)^2 \to \min, \tag{6}$$

где $\lambda_1 = \dfrac{\sigma_{\varepsilon_1}^2}{\sigma_{\varepsilon_0}^2}$, $\lambda_2 = \dfrac{\sigma_{\varepsilon_2}^2}{\sigma_{\varepsilon_0}^2}$ – соотношения дисперсий ошибок (лямбда-параметры); $\sigma_{\varepsilon_0}^2$, $\sigma_{\varepsilon_1}^2$, $\sigma_{\varepsilon_2}^2$ – дисперсии ошибок переменных y, x_1 и x_2.

Как можно заметить, при построении EIV-модели (2) – (5) возникает множество проблем.

1. Практически никогда априори неизвестны дисперсии ошибок переменных $\sigma_{\varepsilon_0}^2$, $\sigma_{\varepsilon_1}^2$, $\sigma_{\varepsilon_2}^2$, а значит, непонятно, какие нужно взять значения лямбда-параметров для решения задачи (6). А выбор неправильных значений λ_1 и λ_2 приведет к смещению оценок параметров β_0, β_1, β_2 в уравнении (2), следствием чего будет ложная интерпретация EIV-модели.

2. Решение задачи (6) оценивания EIV-модели возможно только с использованием специальных численных методов, поскольку целевая функция является нелинейной.

3. Прогнозировать по EIV-модели, используя уравнение (2), можно только в том случае, если известны «истинные» прогнозные значения независимых переменных. Иначе прогноз будет плохим. Проблема в том, что эти «истинные» значения также априори неизвестны.

4. Поскольку EIV-модели, в силу предположения (2), всё же представляют собой множественные модели, то возникает естественный вопрос о тесноте линейной зависимости между «истинными» независимыми переменными x_1^* и x_2^*, т.е. о мультиколлинеарности. Стоит отметить, что хотя автору и не удалось отыскать литературу, касающуюся строгого исследования этого вопроса, но не вызывает сомнения то, что по аналогии с классической регрессией (1) проблема мультиколлинеарности и следствия из неё имеют место и для EIV-моделей.

Для решения приведенных проблем предлагается следующая схема построения EIV-моделей. В нашем распоряжении, по-прежнему, одна зависимая переменная y и две независимых переменных x_1 и x_2. «Истинные» значения этих переменных – y_i^*, x_{i1}^*, x_{i2}^*, $i=\overline{1,n}$. Не обращая пока внимания на переменную y^*, предположим, что между переменными x_1^* и x_2^* имеет место линейная функциональная зависимость:

$$x_{i1}^* = a + bx_{i2}^*, \qquad i=\overline{1,n}, \tag{7}$$

где a, b – неизвестные параметры.

Тогда совокупность уравнений (4), (5) и (7) представляет собой простейшую EIV-модель – регрессию Деминга. Целевая функция для её оценивания с помощью МНПК имеет вид:

$$\sum_{i=1}^{n}\left(x_{i1}-a-bx_{i2}^*\right)^2+\frac{1}{\lambda}\sum_{i=1}^{n}\left(x_{i2}-x_{i2}^*\right)^2\to\min, \tag{8}$$

где $\lambda=\dfrac{\sigma_{\varepsilon_2}^2}{\sigma_{\varepsilon_1}^2}$.

Если значение параметра λ известно, то оценки регрессии Деминга, удовлетворяющие функционалу (8), вычисляются последовательно [5]. Сначала находится оценка параметра b из квадратного уравнения:

$$K_{x_1x_2}b^2-\left(D_{x_1}-\frac{D_{x_2}}{\lambda}\right)b-\frac{K_{x_1x_2}}{\lambda}=0, \tag{9}$$

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

Причем, смыслу задачи удовлетворяет только один из корней уравнения (9):

$$\tilde{b} = \frac{D_{x_1} - \frac{1}{\lambda} D_{x_2} + \sqrt{\left(D_{x_1} - \frac{D_{x_2}}{\lambda} \right)^2 + 4 \frac{K_{x_1 x_2}^2}{\lambda}}}{2K_{x_1 x_2}}. \qquad (10)$$

Затем вычисляется оценка параметра a по формуле:

$$\tilde{a} = \overline{x_1} - \tilde{b}\overline{x_2}, \qquad (11)$$

и «истинные» значения переменной x_2 по формулам:

$$x_{i2}^* = -\frac{\tilde{a}\tilde{b}}{\frac{1}{\lambda} + \tilde{b}^2} + \frac{\tilde{b}}{\frac{1}{\lambda} + \tilde{b}^2} x_{i1} + \frac{\frac{1}{\lambda}}{\frac{1}{\lambda} + \tilde{b}^2} x_{i2}, \qquad i = \overline{1,n}. \qquad (12)$$

«Истинные» значения переменной x_1 можно найти по формулам (7).

Из выражения (12) следует, что если $\lambda \to \infty$, то $x_2^* = -\frac{\tilde{a}}{\tilde{b}} + \frac{1}{\tilde{b}} x_1$, а из уравнения (7) переменная $x_1^* = x_1$, т.е. имеем МНК-оценки обратной регрессии $x_2 = -\frac{a}{b} + \frac{1}{b} x_1 + \varepsilon$; если $\lambda \to 0$, то $x_2^* = x_2$, $x_1^* = \tilde{a} + \tilde{b} x_2$, т.е. имеем МНК-оценки прямой регрессии $x_1 = a + b x_2 + \varepsilon$. Иными словами, варьирование значений лямбда-параметра приводит к изменению направления связи между переменными x_1 и x_2: при $\lambda \to 0$ переменная x_2 влияет на x_1, а при $\lambda \to \infty$ наоборот.

Поскольку «истинная» переменная x_2^* является линейной комбинацией (12) наблюдаемых переменных x_1 и x_2, то возникла идея интегрировать x_2^* в модель парной линейной регрессии:

$$y_i = c_0 + c_1 x_{i2}^* + \varepsilon_i, \qquad i = \overline{1,n}, \qquad (13)$$

где c_0, c_1 - неизвестные параметры, которые находятся с помощью обычного МНК.

Рассмотрим некоторые основные этапы построения и анализа предлагаемого синтеза модели парной линейной регрессии (13) и простейшей EIV-модели (4), (5), (7).

1. Идентификация модели.

Синтезированная модель достаточно легко оценивается. Если значение параметра λ, т.е. соотношение дисперсий ошибок переменных x_1 и x_2, известно, то, последовательно вычислив по формулам (10) – (12) оценки параметров $\tilde{b}$, $\tilde{a}$ и x_{i2}^*, $i=\overline{1,n}$, можно найти МНК-оценки параметров c_1 и c_0 по известным формулам:

$$\tilde{c}_1 = \frac{\overline{yx_2^*} - \overline{y}\cdot\overline{x_2^*}}{\overline{\left(x_2^*\right)^2} - \left(\overline{x_2^*}\right)^2}, \qquad \tilde{c}_0 = \overline{y} - \tilde{c}_1\overline{x_2^*}. \tag{14}$$

Тогда оцененная синтезированная модель представима в виде системы уравнений:

$$y^* = \tilde{c}_0 + \tilde{c}_1 x_2^*, \tag{15}$$

$$x_1^* = \tilde{a} + \tilde{b}x_2^*, \tag{16}$$

$$x_2^* = A_0 + A_1 x_1 + A_2 x_2, \tag{17}$$

где $A_0 = -\dfrac{\tilde{a}\tilde{b}}{\dfrac{1}{\lambda}+\tilde{b}^2}$, $A_1 = \dfrac{\tilde{b}}{\dfrac{1}{\lambda}+\tilde{b}^2}$, $A_2 = \dfrac{\dfrac{1}{\lambda}}{\dfrac{1}{\lambda}+\tilde{b}^2}$.

В соответствии с системой (15) – (17), структурная спецификация предложенной синтезированной модели представлена на рис. 1.

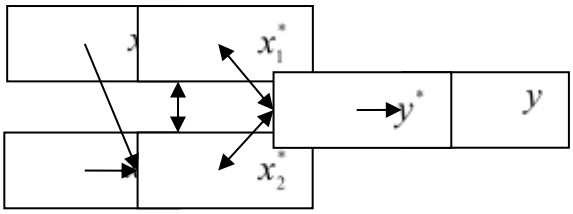

Рис. 1. Структурная спецификация синтезированной модели

Изображенная на рис. 1 спецификация очень сильно напоминает полносвязную топологию нейронных сетей, в которой любой нейрон может передавать свой сигнал любому другому нейрону напрямую. В нашем случае, все «истинные» переменные x_1^*, x_2^* и y^* взаимосвязаны между собой. Зная значение любой из этих переменных, можно единственным образом вычислить значения остальных. В соответствии с этим, будем называть предложенный синтез модели парной линейной регрессии (13) и

простейшей EIV-модели (4), (5), (7) – двухфакторной моделью полносвязной линейной регрессией. Такая модель оценивается в 2 этапа.

Этап 1. С помощью МНПК находятся «истинные» значения независимых переменных x_1 и x_2.

Этап 2. С помощью МНК находятся «истинные» значения зависимой переменной y.

Важно отметить, что интеграция в модель (13) сразу двух «истинных» переменных x_1^* и x_2^* приведёт к её неидентифицируемости из-за совершенной мультиколлинеарности, поскольку между этими переменными изначально постулируется функциональная зависимость (7). А интеграция в модель (13) переменной x_1^* вместо x_2^*, конечно же, повлияет на оценки параметров c_0 и c_1, однако, в силу равенства (7), это будет всё та же модель лишь в другом математическом представлении, т.е. её качественные характеристики останутся неизменными.

2. Верификация модели (проверка адекватности).

Поскольку в модель полносвязной регрессии входит 3 уравнения (4), (5) и (13), содержащие ошибки и оцениваемые с помощью МНПК и МНК, то проверка адекватности может осуществляться для каждого из них. Для этого можно использовать такие классические критерии адекватности, как: коэффициент детерминации, критерий Фишера, Дарбина-Уотсона, «согласованность поведения» и др.

3. Прогнозирование по модели.

Синтезированную модель, в отличие от EIV-модели (4), (5), (7), вполне можно использовать для прогнозирования. Для этого, используя выражение (12), необходимо найти расчетные значения переменной x_2 и подставить их в оцененную модель (13).

4. Интерпретация модели.

Интерпретация полносвязных регрессий гораздо более разнообразна, чем интерпретация множественных моделей.

Для уравнения (15) интерпретация коэффициента $\tilde{c}_1$: с изменением «истинного» значения x_2^* на 1 единицу, значение переменной y изменяется в среднем на $\tilde{c}_1$ единиц. Интерпретация $\tilde{c}_0$: если «истинное» значение переменной $x_2^* = 0$, то $y = \tilde{c}_0$. Если из равенства (16) выразить переменную x_2^* и подставить её в уравнение (15), то получим зависимость y^* от x_1^*, которая интерпретируется аналогично.

Для уравнения (16) интерпретация $\tilde{b}$: с изменением «истинного» значения x_2^* на 1 единицу, «истинное» значение переменной x_1^* изменится

ровно на $\tilde{b}$. Интерпретация $\tilde{a}$: если «истинное» значение переменной $x_2^* = 0$, то $x_1^* = \tilde{a}$.

Для уравнения (17), связывающего «истинное» значение переменной x_2^* с зашумленными переменными x_1 и x_2, интерпретация A_0: если значения переменных $x_1 = x_2 = 0$, то $x_2^* = A_0$. Интерпретация A_1: с изменением x_1 на 1 единицу, значение переменной x_2^* изменится ровно на 1 единицу. Аналогично для коэффициента A_2. Если в равенство (16) вместо переменной x_2^* подставить выражение (17), то получим линейную зависимость переменной x_1^* от x_1 и x_2, которая интерпретируется аналогично.

Отметим, что помимо уравнений (15) – (17) можно дополнительно получить и интерпретировать статистические зависимости между зашумленными и «истинными» переменными.

Подчеркнем два свойства полносвязных регрессий.

Свойство 1. Для построения полносвязной регрессии вообще не требуется заниматься проблемой мультиколлинеарности, потому что в основе синтеза изначально лежит предположение о том, что «истинные» переменные x_1^* и x_2^* связаны функциональной зависимостью (7). Это подтверждается тем, что в процессе двухэтапной идентификации синтезированной модели всегда приходиться оценивать только однофакторные регрессии. Таким образом, множественная регрессия (1) представляет собой некую «смесь» простых зависимостей, по которой можно прогнозировать, но которую не всегда можно корректно интерпретировать из-за мультиколлинеарности. Полносвязная же регрессия не только пригодна для получения прогнозов зависимой переменной, но и позволяет разделять образованную «смесь» на простые и интерпретируемые взаимосвязанные уравнения.

Свойство 2. Полносвязная регрессия может использоваться как инструмент для решения задачи отбора регрессоров в регрессионной модели. Действительно, перепишем регрессию (13), используя выражение (17):

$$y_i = c_0 + A_0 c_1 + A_1 c_1 x_{i1} + A_2 c_1 x_{i2} + \varepsilon_i, \qquad i = \overline{1,n}. \tag{18}$$

Из уравнения (18) следует очевидный факт: для любого λ сумма квадратов остатков регрессии (13) в полносвязной модели не может быть меньше суммы квадратов остатков множественной регрессии (1).

В регрессии (18) коэффициенты A_0, A_1, A_2 зависят от параметра λ. Так, если $\lambda \to 0$, то $A_0 = 0$, $A_1 = 0$, $A_2 = 1$, следовательно, модель (18) становится однофакторной линейной регрессией:

$$y_i = c_0 + c_1 x_{i2} + \varepsilon_i, \qquad i = \overline{1,n}, \tag{19}$$

Если же $\lambda \to \infty$, то $A_0 = -\frac{a}{b}$, $A_1 = \frac{1}{b}$, $A_2 = 0$, а значит, регрессия (18) принимает вид:

$$y_i = c_0 + c_1\left(-\frac{a}{b} + \frac{1}{b}x_{i1}\right) + \varepsilon_i, \qquad i = \overline{1,n}. \tag{20}$$

Таким образом, с ростом значений параметра λ от 0 до ∞ происходит трансформация классической однофакторной регрессии (19) с независимой переменной x_2 в регрессию (20) с независимой переменной x_1. При других значениях параметра λ модель сохранят вид двухфакторной модели (18). Возникает вопрос: существует ли значение лямбда-параметра, при котором регрессия (18) становится классической двухфакторной линейной регрессией (1)? Ответ на него будет дан в последующих работах автора.

Завершение. В данной работе впервые предложен синтез модели парной линейной регрессии и простейшей EIV-модели, названный моделью полносвязной линейной регрессии. Рассмотрены основные этапы построения и анализа таких моделей.

Модель полносвязной линейной регрессии имеет много общего с классической моделью множественной регрессии, однако в основе этих двух видов совершенно разные подходы. Если множественная регрессия строится по принципу «независимые переменные влияют на зависимую», то принципом полносвязной регрессии является «все переменные влияют друг на друга». Отсюда следует, что для построения полносвязных регрессий, в отличие от множественных, вообще не требуется заниматься проблемой мультиколлинеарности. Как установлено, прогностические способности линейных полносвязных моделей не превосходят способностей множественных регрессий, но зато первые имеют гораздо более разнообразную интерпретацию. Помимо этого, полносвязные регрессии можно использовать как инструмент при построении множественных моделей для решения задач снижения размерности данных, устранения мультиколлинеарности и отбора информативных регрессоров.

ЛИТЕРАТУРА

1. Айвазян С.А. Прикладная статистика: исследование зависимостей / С.А. Айвазян, И.С. Енюков, Л.Д. Мешалкин. – М. : Финансы и статистика, 1985. – 487 с.
2. Носков С.И. Технология моделирования объектов с нестабильным функционированием и неопределенностью в данных / С.И. Носков. – Иркутск : Облинформпечать, 1996. – 321 с.
3. Пирогов Г.Г. Проблемы структурного оценивания в эконометрии / Г.Г. Пирогов, Ю.П. Федоровский. – М.: Статистика, 1979. – 327 с.
4. Deming W.E. Statistical adjustment of data / W.E. Deming. – New York, Dover Publications, 2011. – 288 p.
5. Базилевский М.П. Аналитические зависимости между коэффициентами детерминации и соотношением дисперсий ошибок исследуемых признаков в модели регрессии Деминга / М.П. Базилевский // Математическое моделирование и численные методы, 2016. – №2 (10). – С. 104-116.
6. Демиденко Е.З. Линейная и нелинейная регрессия / Е.З. Демиденко. – М.: Финансы и статистика, 1981. – 304 с.
7. Кудрина М.А. Алгоритм волновой скелетизации растровых изображений / М.А. Кудрина, В.С. Мишенев // IV Международная конференция и молодежная школа «Информационные технологии и нанотехнологии» : сборник трудов ИТНТ-2018. Самара, 2018. – С. 784-792.
8. Каллнер А. Сравнение результатов измерений глюкозы крови с помощью интерактивной клинико-лабораторной оценки и метода решеток ошибок Кларка / А. Каллнер, О.В. Черничук, Л.А. Хоровская // Клинико-лабораторный консилиум. – 2009. – № 4. – С. 14-15.
9. Смирнов М.Б. Зависимости между основными структурно-групповыми параметрами состава нефтей Волго-Уральского нефтегазоносного бассейна по данным ЯМР 1Н и 13С / М.Б. Смирнов, Н.А. Ванюкова // Нефтехимия. 2017. Т. 57, № 3. С. 269-277.
10. Clinical Chemistry. Available at: http://clinchem.aaccjnls.org/.

M. P. Bazilevskiy

SYNTHESIS OF LINEAR REGRESSION MODEL AND EIV-MODEL

Irkutsk State Transport University,
Irkutsk, Russia

This paper is devoted to a synthesis of pair-wise linear regression model and simplest EiV-model (Errors-In-Variables model), better known as the Deming regression. The EIV model is a regression in which all variables contain random errors. Such models have a number of significant drawbacks, which makes it difficult to work with them. The synthesis proposed in the paper, called the two-factor model of a fully connected linear regression, is not only devoid of these shortcomings, but also has certain advantages. The main stages of the construction and analysis of two-factor models of fully connected linear regression are considered. The proposed fully connected linear regression model has much in common with the classical multiple regression model; however, these two types are based on completely different approaches. If multiple regression is based on the principle "independent variables affect the dependent one", then the principle of fully connected regression is "all variables influence each other". It is established that the approximation abilities of fully connected models do not exceed the capabilities of multiple regressions, but the former have a much more diverse interpretation. The developed synthesis can be used in the construction of multiple models as a tool for solving problems of reducing the dimensionality of data, eliminating multicollinearity and selecting informative regressors.

Keywords: regression model, ordinary least squares, total least squares, Deming regression, EIV-model, fully connected linear regression model.

REFERENCES

1. Ajvazyan S.A., Enyukov I.S., Meshalkin L.D. Prikladnaya statistika: issledovanie zavisimostej. Moscow, Finansy i statistika, 1985. 487 p. (in Russian)
2. Noskov S.I. Tehnologija modelirovanija ob'ektov s nestabil'nym funkcionirovaniem i neopredelennost'ju v dannyh. Irkutsk: RIC GP «Oblinformpechat'» Publ., 1996, 321 p. (in Russian)
3. Pirogov G.G., Fedorovskij Y.P. Problemy strukturnogo ocenivaniya v ehkonometrii. Moscow: Statistics Publ., 1979, 327 p. (in Russian)
4. Deming W.E. Statistical adjustment of data / W.E. Deming. – New York, Dover Publications, 2011. – 288 p.
5. Bazilevskiy M.P. Analytical dependences between the determination coefficients and the ratio of error variances of the test items in Deming regression model. Matematicheskoe modelirovanie i chislennye metody [Mathematical modeling and numerical methods]. 2016, no. 2, vol. 10, pp. 104-116. (in Russian)

6. Demidenko E.Z. Linejnaja i nelinejnaja regressii [Linear and nonlinear regressions]. Moscow: Finance and Statistics Publ., 1981, 304 p. (in Russian)
7. Kudrina M.A., Mishenev V.S. Algorithm of wave skeletonization of raster images. IV Mezhdunarodnaya konferenciya i molodezhnaya shkola «Informacionnye tekhnologii i nanotekhnologii» : sbornik trudov ITNT-2018 [IV International Conference and Youth School "Information Technologies and Nanotechnologies": a collection of works of ITNT-2018]. 2018, pp. 784-792. (in Russian)
8. Kallner A., Chernichuk O.V., Horovskaya L.A. Comparison of the results of blood glucose measurements with the help of an interactive clinical laboratory evaluation and the method of Clark's error gratings. Kliniko-laboratornyj konsilium [Clinical laboratory consultation]. 2009, no. 4, pp. 14-15. (in Russian)
9. Smirnov M.B., Vanyukova N.A. Dependencies between the main structural and group parameters of the composition of the oils of the Volga-Ural oil and gas bearing basin according to NMR 1H and 13C. Neftekhimiya [Petrochemistry]. 2017, vol. 57, no. 3, pp. 269-277. (in Russian)
10. Clinical Chemistry. Available at: http://clinchem.aaccjnls.org/.