УДК 519.862.6 DOI: 10.26102/2310-6018/2019.25.2.008

М.П. Базилевский

ИССЛЕДОВАНИЕ ДВУХФАКТОРНОЙ МОДЕЛИ ПОЛНОСВЯЗНОЙ ЛИНЕЙНОЙ РЕГРЕССИИ

Иркутский государственный университет путей сообщения,
Иркутск, Россия

Данная работа посвящена исследованию модели полносвязной линейной регрессии, представляющей собой синтез модели парной линейной регрессии и регрессии Деминга. Если множественная регрессия строится по принципу «независимые переменные влияют на зависимую», то принципом полносвязной регрессии является «все переменные влияют друг на друга». Полносвязная регрессия достаточно просто оценивается, лишена эффекта мультиколлинеарности, имеет гораздо более разнообразную интерпретацию, чем множественная регрессия, и пригодна для прогнозирования. Однако при построении полносвязной регрессии неизвестным остается соотношение дисперсий ошибок независимых переменных. В данной работе найдено такое соотношение дисперсий ошибок независимых переменных, которое обеспечивает наилучшие аппроксимационные качества вторичной модели полносвязной регрессии. Результаты исследования оформлены в виде теоремы. Из теоремы следует, что значение коэффициента детерминации вторичной модели полносвязной регрессии будет наибольшим либо когда она принимает вид двухфакторной линейной регрессии, либо вид наилучшей по коэффициенту детерминации однофакторной линейной регрессии. Таким образом, осуществляется отбор информативных регрессоров в регрессионной модели. Установлено, что в основе такого отбора лежит полная согласованность знаков коэффициентов при независимых переменных знакам соответствующих коэффициентов корреляции.

Ключевые слова: полносвязная регрессия, множественная регрессия, регрессия Деминга, EIV-модель, коэффициент детерминации, мультиколлинеарность, отбор информативных регрессоров.

Введение. В современном регрессионном анализе выделяется два направления, в соответствии с которыми все регрессионные модели можно разделить на 2 вида.

1. Регрессии без ошибок в объясняющих переменных.
2. Регрессии с ошибками в объясняющих переменных или Errors-In-Variables Models (EIV-модели).

Регрессионные модели без ошибок в объясняющих переменных наиболее просты в применении. С помощью них можно прогнозировать значения объясняемой переменной, можно интерпретировать их оценки, выявлять факторы, оказывающие наиболее или наименее сильное влияние на выходную переменную и т.д. Регрессии без ошибок в объясняющих

переменных легко оценивать, например, с помощью хорошо известного метода наименьших квадратов (МНК). Все эти достоинства делают регрессии первого вида безусловным лидером в исследовательской деятельности с позиции практической применимости, по сравнению с моделями второго вида.

EIV-модели, являясь обобщением моделей без ошибок в объясняющих переменных, практического применения почти не находят. И это несмотря на то, что для их построения на сегодняшний день разработан весьма солидный математический аппарат [1,2]. Простейшей EIV-моделью, содержащей только одну входную и выходную переменную, является регрессия Деминга [3–6], которая находит применение в основном в клинической химии. Описание весьма непростых методов оценивания множественных EIV-моделей можно найти в работах [1,2]. Но даже если удается найти оценки EIV-модели, то бывает и вовсе невозможно использовать полученную модель для прогнозирования или интерпретации. Поэтому на практике EIV-модели применяются реже, чем регрессии без ошибок в переменных.

В работах [7–9] предложен синтез модели парной линейной регрессии и простейшей EIV-модели – регрессии Деминга, названный моделью полносвязной линейной регрессии. Если множественная регрессия строится по принципу «независимые переменные влияют на зависимую», то принципом полносвязной регрессии является «все переменные влияют друг на друга». Полносвязная регрессия достаточно просто оценивается, лишена эффекта мультиколлинеарности, имеет гораздо более разнообразную интерпретацию, чем множественная регрессия, и пригодна для прогнозирования. Однако при построении полносвязной регрессии неизвестным остается соотношение дисперсий ошибок независимых переменных. Целью данной работы является исследование полносвязной регрессии и определение такого соотношения дисперсий ошибок переменных, которое бы обеспечивало наилучшие аппроксимационные качества только, так называемого, вторичного уравнения.

Двухфакторная модель полносвязной линейной регрессии. Пусть исследуется влияние двух объясняющих переменных x_1 и x_2, не содержащих случайных ошибок, на объясняемую переменную y. Для получения статистической зависимости между переменными зачастую находятся оценки параметров двухфакторной модели множественной линейной регрессии:

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

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.

Пусть существуют «истинные» (расчетные мо модели) значения объясняющих переменных x_1 и x_2, которые обозначим x_{i1}^*, x_{i2}^*, $i=\overline{1,n}$. Эти «истинные» значения связаны с наблюдаемыми значениями соотношениями:

$$x_{i1}=x_{i1}^*+\varepsilon_{i1}, \qquad i=\overline{1,n}, \tag{2}$$

$$x_{i2}=x_{i2}^*+\varepsilon_{i2}, \qquad i=\overline{1,n}, \tag{3}$$

где ε_{i1}, ε_{i2}, $i=\overline{1,n}$ – случайные отклонения.

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

$$x_{i1}^*=a+bx_{i2}^*, \qquad i=\overline{1,n}, \tag{4}$$

где a, b – неизвестные параметры.

Тогда совокупность уравнений (2) – (4) представляет собой простейшую EIV-модель – регрессию Деминга. Для её оценивания применяется метод наименьших полных квадратов (МНПК), суть которого состоит в минимизации следующей функции:

$$\sum_{i=1}^{n}\left(x_{i1}-a-bx_{i2}^*\right)^2+\frac{1}{\lambda}\sum_{i=1}^{n}\left(x_{i2}-x_{i2}^*\right)^2\to\min, \tag{5}$$

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

Если значение параметра λ известно, то задача (5) имеет следующее решение:

$$\tilde{b}=\frac{D_{x_1}-\dfrac{1}{\lambda}D_{x_2}+\sqrt{\left(D_{x_1}-\dfrac{D_{x_2}}{\lambda}\right)^2+4\dfrac{K_{x_1x_2}^2}{\lambda}}}{2K_{x_1x_2}}, \tag{6}$$

$$\tilde{a} = \overline{x_1} - \tilde{b}\overline{x_2}, \tag{7}$$

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

Используя значения переменной x_2^*, полученные по формулам (8), составим модель парной линейной регрессии:

$$y_i = c_0 + c_1 x_{i2}^* + \varepsilon_i, \qquad i = \overline{1,n}, \tag{9}$$

где c_0, c_1 - неизвестные параметры, которые находятся с помощью обычного МНК. Переменную x_2^* будем называть связующей.

Полученный синтез модели парной линейной регрессии (9) и простейшей EIV-модели (2) – (4) называется двухфакторной моделью полносвязной линейной регрессии [7–9]. Если значение параметра λ известно, то эта модель оценивается в 2 этапа.

Этап 1. С помощью МНПК находятся «истинные» значения связующей переменной x_2^* по формулам (8).

Этап 2. С помощью МНК находятся оценки регрессии (9).

Оцененная модель полносвязной линейной регрессии представима в виде системы уравнений:

$$y^* = \tilde{c}_0 + \tilde{c}_1 x_2^*, \tag{10}$$

$$x_1^* = \tilde{a} + \tilde{b} x_2^*, \tag{11}$$

$$x_2^* = A_0 + A_1 x_1 + A_2 x_2, \tag{12}$$

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

В работе [7] отмечается, что при построении полносвязной регрессии параллельно формируется зависимость между переменной y и переменными x_1 и x_2. Действительно, перепишем уравнение (9), используя выражение (12):

$$y_i = c_0 + A_0 c_1 + A_1 c_1 x_{i1} + A_2 c_1 x_{i2} + \varepsilon_i, \qquad i = \overline{1,n}. \tag{13}$$

Будем называть зависимость (13) моделью вторичной линейной регрессии. Из уравнения (13) следует очевидный факт: для любого λ сумма квадратов остатков вторичной регрессии не может быть меньше суммы квадратов остатков множественной регрессии (1).

В регрессии (13) 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 (13) становится парной линейной регрессией y от x_2. А если $\lambda \to \infty$, то $A_0 = -\frac{a}{b}$, $A_1 = \frac{1}{b}$, $A_2 = 0$, а значит, регрессия (13) становится парной линейной регрессией y от x_1. Таким образом, с ростом значений параметра λ от 0 до ∞ происходит трансформация классической однофакторной регрессии с независимой переменной x_2 в регрессию с независимой переменной x_1. При других значениях параметра λ вторичная модель сохранят двухфакторную спецификацию (13).

Возникает вопрос: существует ли значение параметра λ, при котором вторичная регрессия (13) становится классической двухфакторной линейной регрессией (1)? Ответ на него будет дан ниже.

Построение полносвязной регрессии при неизвестном значении лямбда-параметра. Как уже было отмечено, варьирование значений параметра λ приводит к изменению оценок полносвязной регрессии, что, в свою очередь, отражается на её качественных характеристиках. Будем считать, что цель, которую ставит перед собой исследователь, состоит в том, чтобы обеспечить высокие аппроксимационные качества только для вторичной модели (13), не принимая во внимание те же качества для моделей (4) и (5). Сформулируем следующую задачу: найти такое значение параметра λ, при котором качество вторичной модели (13) в полносвязной регрессии будет наилучшим. Решение этой задачи способно дать ответ на поставленный выше вопрос о том, может ли вторичная регрессия (13) совпадать с множественной моделью (1).

Поскольку аппроксимационные качества моделей (13) и (9) одинаковы, то для удобства будем работать с парной регрессией (9). Так как она оценивается с помощью МНК, то об её аппроксимационных качествах можно судить, например, по величине коэффициента детерминации R^2. Известно, что область его возможных значений принадлежит отрезку $[0,1]$, при этом, чем ближе R^2 к 1, тем выше качество модели.

Будем искать коэффициент детерминации регрессии (9) как квадрат коэффициента линейной корреляции между зависимой переменной y и «истинной» переменной x_2^*:

$$R^2 = \mathrm{corr}^2\left(y, x_2^*\right), \tag{14}$$

где $\mathrm{corr}\left(y, x_2^*\right)$ – коэффициент корреляции между переменными y и x_2^*.

Коэффициент корреляции находится по формуле:

$$\mathrm{corr}\left(y, x_2^*\right) = \frac{\mathrm{cov}\left(y, x_2^*\right)}{\sigma_y \sigma_{x_2^*}}, \tag{15}$$

где $\mathrm{cov}\left(y, x_2^*\right)$ – ковариация между переменными y и x_2^*; σ_y, $\sigma_{x_2^*}$ – среднеквадратические отклонения переменных y и x_2^*.

Подставляя выражение (15) в равенство (14), имеем:

$$R^2 = \frac{\mathrm{cov}^2\left(y, x_2^*\right)}{D_y D_{x_2^*}}. \tag{16}$$

Перепишем уравнение (16), используя равенство (12) и свойства ковариации. В результате получим аналитическое выражение для коэффициента детерминации регрессии (9):

$$R^2\left(b, \lambda\right) = \frac{\left(bK_{x_1 y} + \frac{1}{\lambda} K_{x_2 y}\right)^2}{D_y\left(b^2 D_{x_1} + 2\frac{b}{\lambda} K_{x_1 x_2} + \frac{1}{\lambda^2} D_{x_2}\right)}. \tag{17}$$

Для организации синтеза с EIV-моделью (2) – (4), функцию (17), во-первых, обязательно следует дополнить уравнением связи для переменных b и λ [4–6]:

$$\lambda = \frac{K_{x_1 x_2} - D_{x_2} b}{K_{x_1 x_2} b^2 - D_{x_1} b}. \tag{18}$$

Во-вторых, необходимо учесть, что соотношение дисперсий ошибок переменных должно удовлетворять ограничению $\lambda > 0$.

Таким образом, если параметр $\lambda > 0$ и переменные b и λ связаны соотношением (18), то коэффициент детерминации R^2 модели (9)

полносвязной регрессии находится по формуле (17). Понятно, что качество модели (9) будет наилучшим тогда, когда функция (17) принимает свое максимальное значение, т.е. необходимо исследовать эту функцию на экстремум.

Исследуем функцию двух переменных (17) на экстремум только при одном условии, что справедливо равенство (18), т.е. ограничение на параметр λ ставить пока не будем.

Используя уравнение связи (18), перейдем от функции двух переменных (17) к функции одной переменной:

$$R^2(b)=\frac{b^2\left(b\left(D_{x_2}K_{x_1y}-K_{x_2y}K_{x_1x_2}\right)+D_{x_1}K_{x_2y}-K_{x_1y}K_{x_1x_2}\right)^2}{D_y\left(D_{x_1}D_{x_2}-K_{x_1x_2}^2\right)b^2\left(D_{x_2}b^2-2K_{x_1x_2}b+D_{x_1}\right)}. \qquad (19)$$

Производная функции (19) имеет вид:

$$\left(R^2(b)\right)'=-\frac{\left(D_{x_2}AB+K_{x_1x_2}A^2\right)b^2+\left(D_{x_2}B^2-D_{x_1}A^2\right)b-\left(D_{x_1}AB+K_{x_1x_2}B^2\right)}{D_y\left(D_{x_1}D_{x_2}-K_{x_1x_2}^2\right)\left(D_{x_2}b^2-2K_{x_1x_2}b+D_{x_1}\right)^2}, \qquad (20)$$

где $A=D_{x_2}K_{x_1y}-K_{x_2y}K_{x_1x_2}$, $B=D_{x_1}K_{x_2y}-K_{x_1y}K_{x_1x_2}$. Будем считать далее, что $A\neq 0$ и $B\neq 0$.

Как видно, знак производной (20) зависит только от знака числителя, поскольку знаменатель всегда положителен. А числитель представляет собой уравнение параболы, направление ветвей которой зависит от величины $-(D_{x_2}AB+K_{x_1x_2}A^2)$: если $D_{x_2}AB+K_{x_1x_2}A^2>0$, то ветви направлены вниз, а если $D_{x_2}AB+K_{x_1x_2}A^2<0$, то наоборот.

Критические точки функции (19) находятся из условия:

$$\left(D_{x_2}AB+K_{x_1x_2}A^2\right)b^2+\left(D_{x_2}B^2-D_{x_1}A^2\right)b-\left(D_{x_1}AB+K_{x_1x_2}B^2\right)=0. \qquad (21)$$

Квадратное уравнение (21) всегда имеет 2 корня:

$$b_1=-\frac{B}{A}, \qquad b_2=\frac{D_{x_1}A+K_{x_1x_2}B}{D_{x_2}B+K_{x_1x_2}A}. \qquad (22)$$

Рассмотрим возможные случаи взаимного расположения корней уравнения (21) на числовой оси.

Случай № 1 ($b_1=b_2$). Приравнивая корни (22), имеем равенство:

$$\frac{D_{x_1}A^2 + 2K_{x_1x_2}AB + D_{x_2}B^2}{D_{x_2}AB + K_{x_1x_2}A^2} = 0. \qquad (23)$$

Применяя неравенство Коши-Буняковского $-\sqrt{D_{x_1}D_{x_2}} \le K_{x_1x_2} \le \sqrt{D_{x_1}D_{x_2}}$, легко показать, что числитель дроби (23) неотрицателен. Более того, поскольку считаем, что $A \neq 0,\ B \neq 0$, то числитель всегда положителен, т.е.

$$D_{x_1}A^2 + 2K_{x_1x_2}AB + D_{x_2}B^2 > 0.$$

Следовательно, выражение (23) никогда не обращается в тождество, а значит, корни уравнения (21) всегда различны.

Случай № 2 ($b_1 < b_2$). С использованием корней (22) составим неравенство:

$$\frac{D_{x_1}A^2 + 2K_{x_1x_2}AB + D_{x_2}B^2}{D_{x_2}AB + K_{x_1x_2}A^2} > 0. \qquad (24)$$

Как уже отмечалось, в неравенстве (24) числитель по определению положителен, а значит, и знаменатель должен иметь знак «плюс». Если это так, то ветви параболы в числителе (20) направлены вниз, следовательно, легко идентифицировать знаки первой производной, используя метод интервалов (Рисунок 1).

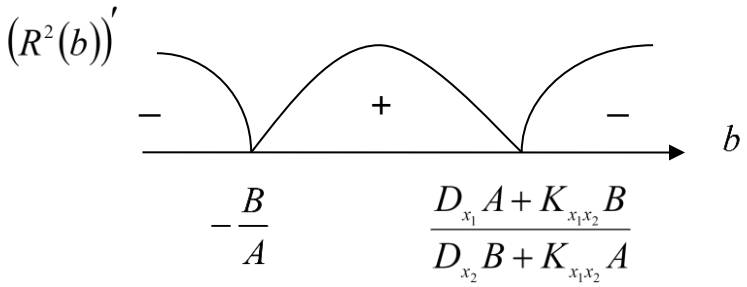

Рисунок 1 – Экстремумы функции (случай № 2)

Как следует из Рисунка 1, на промежутке $\left(-\infty, -\frac{B}{A}\right) \cup \left(\frac{D_{x_1}A + K_{x_1x_2}B}{D_{x_2}B + K_{x_1x_2}A}, \infty\right)$ функция монотонно убывает, а на

промежутке $\left(-\frac{B}{A}, \frac{D_{x_1}A + K_{x_1x_2}B}{D_{x_2}B + K_{x_1x_2}A}\right)$ – монотонно возрастает. Следовательно, в точке $b_1 = -\frac{B}{A}$ минимум, а в точке $b_2 = \frac{D_{x_1}A + K_{x_1x_2}B}{D_{x_2}B + K_{x_1x_2}A}$ – максимум.

Случай № 3 ($b_1 > b_2$). Действуя аналогично, можно установить, что ветви параболы в числителе (20) для этого случая направлены вверх. Графическая иллюстрация метода интервалов представлена на Рисунке 2.

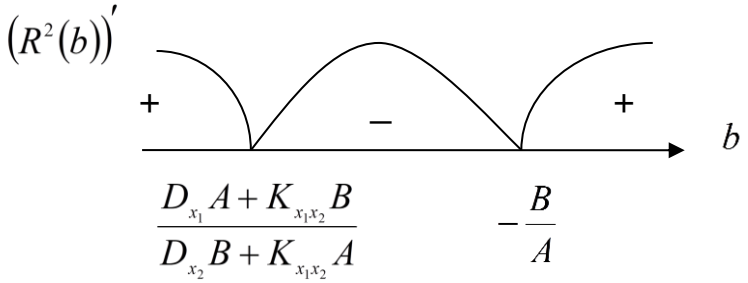

Рисунок 2 – Экстремумы функции (случай № 3)

Из Рисунка 2 следует, что на промежутке $\left(-\infty, \frac{D_{x_1}A + K_{x_1x_2}B}{D_{x_2}B + K_{x_1x_2}A}\right) \cup \left(-\frac{B}{A}, \infty\right)$ функция монотонно возрастает, а на промежутке $\left(\frac{D_{x_1}A + K_{x_1x_2}B}{D_{x_2}B + K_{x_1x_2}A}, -\frac{B}{A}\right)$ – монотонно убывает. Это означает, что в точке $b_1 = -\frac{B}{A}$ минимум, а в точке $b_2 = \frac{D_{x_1}A + K_{x_1x_2}B}{D_{x_2}B + K_{x_1x_2}A}$ – максимум.

Обобщая все случаи, можно сделать вывод, что функция (17) при условии (18), всегда имеет минимум в точке $b_1 = -\frac{B}{A}$ и максимум в точке $b_2 = \frac{D_{x_1}A + K_{x_1x_2}B}{D_{x_2}B + K_{x_1x_2}A}$. По уравнению связи (18) значения переменной λ в указанный точках экстремума совпадают и составляют $\lambda = \frac{K_{x_1x_2}A^2 + D_{x_2}AB}{K_{x_1x_2}B^2 + D_{x_1}AB}$.

Сравним значение функции (17) при условии (18) в точке максимума со значением коэффициента детерминации множественной регрессии (1).

Сначала подставим координаты точки максимума $b = \dfrac{D_{x_1} A + K_{x_1x_2} B}{D_{x_2} B + K_{x_1x_2} A}$, $\lambda = \dfrac{K_{x_1x_2} A^2 + D_{x_2} AB}{K_{x_1x_2} B^2 + D_{x_1} AB}$ в функцию (17). Получим:

$$R^2_{\max} = \frac{\left(AK_{x_1y} + BK_{x_2y} \right)^2}{D_y \left(A^2 D_{x_1} + 2ABK_{x_1x_2} + B^2 D_{x_2} \right)}. \qquad (25)$$

Известно, что МНК-оценки параметров для двухфакторной регрессии (1) находятся по формулам [10]:

$$\tilde{\alpha}_1 = \frac{A}{D_{x_1} D_{x_2} - K^2_{x_1x_2}}, \qquad \tilde{\alpha}_2 = \frac{B}{D_{x_1} D_{x_2} - K^2_{x_1x_2}}. \qquad (26)$$

Вычисляя коэффициент детерминации регрессии (1) как квадрат коэффициента линейной корреляции между фактическими и расчетными значениями переменной y с использованием формул (26), можно получить для него аналитическое выражение, полностью совпадающее с выражением (25).

Таким образом, в единственной точке максимума значение функции (17) при условии (18) всегда совпадает с коэффициентом детерминации множественной регрессии (1). Но при этом может получиться так, что в точке максимума значение лямбда-параметра будет неположительным $\lambda \le 0$, что противоречит предпосылкам полносвязной регрессии, поскольку λ – соотношение дисперсий ошибок переменных. Из этого следует, что вторичная регрессия (13) совпадает с множественной моделью (1) только тогда, когда выполняется условие $\dfrac{K_{x_1x_2} A^2 + D_{x_2} AB}{K_{x_1x_2} B^2 + D_{x_1} AB} > 0$. Возникает вопрос, а что происходит в противном случае?

Дополним исследование ограничением на параметр λ, т.е. рассмотрим теперь функцию (17) при условии (18) и при $\lambda > 0$. Используя равенство (18), можно определить область значений переменной b при $\lambda > 0$. Для этого составим неравенство

$$\frac{K_{x_1x_2} - D_{x_2} b}{K_{x_1x_2} b^2 - D_{x_1} b} > 0. \qquad (27)$$

Решение неравенства (27) имеет вид:

если $K_{x_1x_2} > 0$, то $b \in \left(\frac{K_{x_1x_2}}{D_{x_2}}, \frac{D_{x_1}}{K_{x_1x_2}} \right)$;

если $K_{x_1x_2} < 0$, то $b \in \left(\frac{D_{x_1}}{K_{x_1x_2}}, \frac{K_{x_1x_2}}{D_{x_2}} \right)$.

Из этого следует, что для простейшей EIV-модели (2) – (4), знак оценки параметра b всегда совпадает со знаком ковариации между переменными x_1 и x_2, т.е. $bK_{x_1x_2} > 0$. Тогда это свойство всегда будет выполняться и для полносвязной регрессии.

Поставим задачу максимизации функции (17):

$$R^2(b, \lambda) \to \max, \tag{28}$$

при условии (18) и ограничении $\lambda > 0$.

Теорема. Пусть $A \neq 0$, $B \neq 0$, $K_{x_1x_2} \neq 0$, $K_{x_1y} \neq 0$, $K_{x_2y} \neq 0$; m, M – малое и большое положительные числа. Тогда задача (28) при условиях (18) и $\lambda > 0$ всегда имеет единственное решение, причем:

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

(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 $b = \frac{D_{x_1}}{K_{x_1x_2}}$, $\lambda = M$;

при $\frac{K_{x_1x_2}^2}{D_{x_2}} < -\frac{B}{A}K_{x_1x_2} < D_{x_1}$ либо в точке $b = \frac{D_{x_1}}{K_{x_1x_2}}$, $\lambda = M$, либо в точке $b = \frac{K_{x_1x_2}}{D_{x_2}}$, $\lambda = m$;

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

Доказательство. Рассмотрим все возможные варианты знаков коэффициентов A, B, и ковариации $K_{x_1x_2}$.

Случай № 1. Пусть $AB>0$ и $K_{x_1x_2}>0$. Тогда координата точки минимума функции (17) при условии (18) $b_1=-\frac{B}{A}<0$, а точки максимума $b_2=\frac{D_{x_1}A+K_{x_1x_2}B}{D_{x_2}B+K_{x_1x_2}A}>0$. Поскольку при $K_{x_1x_2}>0$ переменная $b\in\left(\frac{K_{x_1x_2}}{D_{x_2}},\frac{D_{x_1}}{K_{x_1x_2}}\right)$, то точка минимума b_1 в этот промежуток никогда не попадает. Используя неравенство Коши-Буняковского, можно показать справедливость двойного неравенства:

$$\frac{K_{x_1x_2}}{D_{x_2}}<\frac{D_{x_1}A+K_{x_1x_2}B}{D_{x_2}B+K_{x_1x_2}A}<\frac{D_{x_1}}{K_{x_1x_2}},$$

из чего следует, что при $AB>0$ и $K_{x_1x_2}>0$ задача (28) при условиях (18) и $\lambda>0$ имеет решение в точке b_2.

Случай № 2. Пусть $AB>0$ и $K_{x_1x_2}<0$. Тогда координата точки минимума $b_1=-\frac{B}{A}<0$, а координата точки максимума b_2 может иметь любой знак. Известно, что при $K_{x_1x_2}<0$ переменная $b\in\left(\frac{D_{x_1}}{K_{x_1x_2}},\frac{K_{x_1x_2}}{D_{x_2}}\right)$. Легко показать, что если точка минимума b_1 расположена правее указанного промежутка, т.е. $-\frac{B}{A}>\frac{K_{x_1x_2}}{D_{x_2}}$, то точка максимума b_2 всегда расположена левее него. Следовательно, т.к. $b_1>b_2$ (см. рис. 3), то функция на промежутке $b\in\left(\frac{D_{x_1}}{K_{x_1x_2}},\frac{K_{x_1x_2}}{D_{x_2}}\right)$ монотонно убывает. Значит, наибольшее значение она принимает при $b\to\frac{D_{x_1}}{K_{x_1x_2}}$, т.е. при $\lambda\to\infty$.

Если $b_1=-\frac{B}{A}<\frac{K_{x_1x_2}}{D_{x_2}}$, то $b_2>\frac{K_{x_1x_2}}{D_{x_2}}$. При этом, если $b_1\in\left(\frac{D_{x_1}}{K_{x_1x_2}},\frac{K_{x_1x_2}}{D_{x_2}}\right)$, т.е. минимум внутри промежутка, то наибольшее значение будет либо при $b\to\frac{D_{x_1}}{K_{x_1x_2}}$, либо при $b\to\frac{K_{x_1x_2}}{D_{x_2}}$. Если $b_1<\frac{D_{x_1}}{K_{x_1x_2}}$, то $b_1<b_2$, следовательно,

функция монотонно возрастает на промежутке. Значит, наибольшее значение она принимает при $b \to \frac{K_{x_1x_2}}{D_{x_2}}$, т.е. при $\lambda \to 0$.

Равенство $-\frac{B}{A} = \frac{K_{x_1x_2}}{D_{x_2}}$ равносильно уравнению $\left(D_{x_1}D_{x_2} - K_{x_1x_2}^2\right)K_{x_2y} = 0$, которое обращается в нуль либо когда переменные x_1 и x_2 линейно зависимы, либо при $K_{x_2y} = 0$, что противоречит условию теоремы.

Случай № 3. Пусть $AB < 0$ и $K_{x_1x_2} > 0$. Тогда координата точки минимума $b_1 = -\frac{B}{A} > 0$, а координата точки максимума b_2 может иметь любой знак. Известно, что при $K_{x_1x_2} > 0$ переменная $b \in \left(\frac{K_{x_1x_2}}{D_{x_2}}, \frac{D_{x_1}}{K_{x_1x_2}}\right)$.

Если $b_1 = -\frac{B}{A} < \frac{K_{x_1x_2}}{D_{x_2}}$, то $b_2 > \frac{D_{x_1}}{K_{x_1x_2}}$, откуда $b_1 < b_2$. Следовательно, функция монотонно возрастает на $b \in \left(\frac{K_{x_1x_2}}{D_{x_2}}, \frac{D_{x_1}}{K_{x_1x_2}}\right)$. Значит, наибольшее значение она принимает при $b \to \frac{D_{x_1}}{K_{x_1x_2}}$, т.е. при $\lambda \to \infty$.

Если $b_1 = -\frac{B}{A} > \frac{K_{x_1x_2}}{D_{x_2}}$, то $b_2 < \frac{K_{x_1x_2}}{D_{x_2}}$. При этом, если минимум внутри промежутка, т.е. $b_1 \in \left(\frac{K_{x_1x_2}}{D_{x_2}}, \frac{D_{x_1}}{K_{x_1x_2}}\right)$, то наибольшее значение функция принимает либо при $b \to \frac{D_{x_1}}{K_{x_1x_2}}$, либо при $b \to \frac{K_{x_1x_2}}{D_{x_2}}$. Если $b_1 > \frac{D_{x_1}}{K_{x_1x_2}}$, то $b_1 > b_2$, следовательно, функция монотонно убывает на промежутке. Значит, её наибольшее значение будет при $b \to \frac{K_{x_1x_2}}{D_{x_2}}$, т.е. при $\lambda \to 0$.

Случай № 4. Пусть $AB<0$ и $K_{x_1x_2}<0$. Тогда $b_1>0$, $b_2<0$. Т.к. $K_{x_1x_2}<0$, то $b\in\left(\frac{D_{x_1}}{K_{x_1x_2}},\frac{K_{x_1x_2}}{D_{x_2}}\right)$, т.е. точка минимума не попадает в этот промежуток. Можно показать, что $b_2\in\left(\frac{D_{x_1}}{K_{x_1x_2}},\frac{K_{x_1x_2}}{D_{x_2}}\right)$. Отсюда решение задачи (28) при условиях (18) и $\lambda>0$ в точке b_2.

Заменив предельные равенства $\lambda\to 0$, $\lambda\to\infty$ на $\lambda=m$, $\lambda=M$, и обобщая все рассмотренные случаи, умножив неравенства на $K_{x_1x_2}$, нетрудно получить условия теоремы (1) и (2). Теорема доказана.

Из теоремы следует, что значение коэффициента детерминации вторичной модели (13) полносвязной регрессии будет наибольшим либо когда уравнение (13) принимает вид двухфакторной линейной регрессии (1), либо вид наилучшей по коэффициенту детерминации однофакторной линейной регрессии. Таким образом, решение задачи (28) при условии (18) и $\lambda>0$ имитирует процедуру отбора информативных регрессоров в регрессионной модели (1).

В чем состоит принцип такого отбора? Если все переменные y, x_1 и x_2 тесно коррелируют друг с другом, то во множественной регрессии (1), из-за эффекта мультиколлинеарности, могут искажаться знаки коэффициентов $\tilde{\alpha}_1$ и $\tilde{\alpha}_2$. Однако вторичная модель, получаемая при решении задачи (28) с условиями (18) и $\lambda>0$, не допускает такого искажения. Действительно, согласно формулам (26), знаки коэффициентов A и B совпадают со знаками оценок двухфакторной линейной регрессии (1) $\tilde{\alpha}_1$ и $\tilde{\alpha}_2$. Тогда, например, первое условие теоремы $ABK_{x_1x_2}>0$ равносильно условию $\tilde{\alpha}_1\tilde{\alpha}_2K_{x_1x_2}>0$, которое означает, что для того чтобы вторичное уравнение (13) приняло вид двухфакторной регрессии (1) необходимо, чтобы: при $K_{x_1x_2}>0$ оценки регрессии (1) $\tilde{\alpha}_1$ и $\tilde{\alpha}_2$ были одного знака; при $K_{x_1x_2}<0$ оценки $\tilde{\alpha}_1$ и $\tilde{\alpha}_2$ были разных знаков. Если это условие не выполняется, то, как следует из теоремы, происходит исключение одного из факторов и уравнение (13) принимает вид наилучшей однофакторной регрессии. Таким образом, в основе отбора информативных регрессоров в модели (1) при решении задачи (28) с условиями (18) и $\lambda>0$ лежит полная согласованность знаков коэффициентов при переменных x_1 и x_2 знакам соответствующих коэффициентов корреляции r_{yx_1} и r_{yx_2}.

Завершение. В данной работе исследована модель полносвязной линейной регрессии и найдено такое соотношение дисперсий ошибок независимых переменных, которое обеспечивает наилучшие аппроксимационные качества только вторичной модели. Результаты исследования оформлены в виде теоремы. Из теоремы следует, что значение коэффициента детерминации вторичной модели полносвязной регрессии будет наибольшим либо когда она принимает вид двухфакторной линейной регрессии, либо вид наилучшей по коэффициенту детерминации однофакторной линейной регрессии. Таким образом, осуществляется отбор информативных регрессоров в регрессионной модели. Установлено, что в основе такого отбора лежит полная согласованность знаков коэффициентов при независимых переменных знакам соответствующих коэффициентов корреляции.

ЛИТЕРАТУРА

1. Кендалл М. Статистические выводы и связи / М. Кендалл, А. Стьюарт. – Главная редакция физико-математической литературы изд-ва «Наука», 1973. – 899 с.
2. Демиденко Е.З. Линейная и нелинейная регрессия / Е.З. Демиденко. – М.: Финансы и статистика, 1981. – 304 с.
3. Deming W.E. Statistical adjustment of data / W.E. Deming. – New York, Dover Publications, 2011. – 288 p.
4. 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.
5. Базилевский М.П. Аналитические зависимости для некоторых критериев адекватности модели регрессии Деминга // Вестник ИрГТУ. – Иркутск, 2016. – Т.20 – №10. – С. 81-89.
6. Базилевский М.П. Методика многокритериального выбора лямбда-параметра в модели парной линейной регрессии со стохастическими переменными // Вестник ИрГТУ. – Иркутск, 2017. – Т.21 – №3. – С. 59-72.
7. Базилевский М.П. Синтез модели парной линейной регрессии и простейшей EIV-модели // Моделирование, оптимизация и информационные технологии. – Воронеж, 2019. – Т. 7. – № 1. – Режим

доступа: https://moit.vivt.ru/wp-content/uploads/2019/01/Bazilevskiy_1_19_1.pdf.

8. Базилевский М.П. Двухфакторная модель полносвязной регрессии с квадратом связующей переменной // Молодежь и современные информационные технологии: сборник трудов XVI Международной научно-практической конференции студентов, аспирантов и молодых ученых. – Томск, 2018. – С. 26–27.
9. Базилевский М.П. Оценивание параметров простейшей модели полносвязной линейной регрессии // Достижения и приложения современной информатики, математики и физики: материалы VII Всероссийской научно-практической конференции. – Нефтекамск, 2018. – С. 179-184.
10. Гефан Г.Д. Эконометрика. – Иркутск: ИрГУПС, 2005. – 84 с.

M.P. Bazilevskiy

INVESTIGATION OF A TWO-FACTOR FULLY CONNECTED LINEAR REGRESSION MODEL

Irkutsk State Transport University,

Irkutsk, Russia

This paper is devoted to the study of a fully connected linear regression model, which is a synthesis of the pairing linear regression model and the Deming regression model. If multiple regression is based on the principle "independent variables influence dependent", then the principle of fully connected regression is "all variables influence each other". A fully connected regression is fairly simply estimated, devoid of multicollinearity effect, has a much more diverse interpretation than multiple regression, and is suitable for prediction. However, when building a fully connected regression, the ratio of error variances of independent variables remains unknown. In this paper, we find the ratio of error variances of independent variables that provides the best approximation qualities of the secondary fully connected regression model. The research results are presented in the form of a theorem. It follows from the theorem that the value of the coefficient of determination of the secondary model of a fully connected regression will be greatest either when it takes the form of a two-factor linear regression or the best one in the coefficient of determination of a single-factor linear regression. Thus, the selection of informative regressors in the regression model is carried out. It is established that the basis of such a selection is the complete consistency of the signs of the coefficients with independent variable signs of the corresponding correlation coefficients.

Keywords: fully connected regression, multiple regression, Deming regression, EIV-model, coefficient of determination, multicollinearity, subset selection in regression.

REFERENCES

1. Kendall M., St'yuart A. Statisticheskie vyvody i svyazi. Glavnaya redakciya fiziko-matematicheskoj literatury izd-va «Nauka», 1973, 899 p. (in Russian)
2. Demidenko E.Z. Linejnaya i nelinejnaya regressiya. Moscow: Finansy i statistika, 1981. 304 p. (in Russian)
3. Deming W.E. Statistical adjustment of data. New York, Dover Publications, 2011. 288 p.
4. Bazilevskij M.P. Analiticheskie zavisimosti mezhdu koefficientami determinacii i sootnosheniem dispersij oshibok issleduemyh priznakov v modeli regressii Deminga. Matematicheskoe modelirovanie i chislennye metody. 2016, no. 2, vol. 10, pp. 104–116. (in Russian)
5. Bazilevskij M.P. Analiticheskie zavisimosti dlya nekotoryh kriteriev adekvatnosti modeli regressii Deminga. Vestnik IrGTU. Irkutsk, 2016, vol. 20, no. 10, pp. 81–89. (in Russian)
6. Bazilevskij M.P. Metodika mnogokriterial'nogo vybora lyambda-parametra v modeli parnoj linejnoj regressii so stohasticheskimi peremennymi. Vestnik IrGTU. Irkutsk, 2017, vol. 21, no. 3, pp. 59–72. (in Russian)
7. Bazilevskij M.P. Sintez modeli parnoj linejnoj regressii i prostejshej EIV-modeli. Modelirovanie, optimizaciya i informacionnye tekhnologii. Voronezh, 2019, vol. 7, no. 1. URL: https://moit.vivt.ru/wp-content/uploads/2019/01/Bazilevskiy_1_19_1.pdf. (in Russian)
8. Bazilevskij M.P. Dvuhfaktornaya model' polnosvyaznoj regressii s kvadratom svyazuyushchej peremennoj. Molodezh' i sovremennye informacionnye tekhnologii: sbornik trudov XVI Mezhdunarodnoj nauchno-prakticheskoj konferencii studentov, aspirantov i molodyh uchenyh. Tomsk, 2018, pp. 26–27. (in Russian)
9. Bazilevskij M.P. Ocenivanie parametrov prostejshej modeli polnosvyaznoj linejnoj regressii. Dostizheniya i prilozheniya sovremennoj informatiki, matematiki i fiziki: materialy VII Vserossijskoj nauchno-prakticheskoj konferencii. Neftekamsk, 2018, pp. 179–184. (in Russian)
10. Gefan G.D. Ekonometrika. Irkutsk, 2005, 84 p. (in Russian)