Моделирование и анализ данных
2023. Том 13. № 3. С. 52–65
DOI: https://doi.org/10.17759/mda.2023130304
ISSN: 2219-3758 (печатный)
ISSN: 2311-9454 (online)

Modelling and Data Analysis
2023. Vol. 13, no. 3, pp. 52–65
DOI: https://doi.org/10.17759/mda.2023130304
ISSN: 2219-3758 (print)
ISSN: 2311-9454 (online)

◇◇◇◇◇◇◇ МЕТОДЫ ОПТИМИЗАЦИИ ◇◇◇◇◇◇◇

УДК 519.862.6

Идентификация областей возможных оценок параметров моделей полносвязной линейной регрессии

Базилевский М.П.*
Иркутский государственный университет путей сообщения
(ФГБОУ ВО ИрГУПС), г. Иркутск, Российская Федерация
ORCID: https://orcid.org/0000-0002-3253-5697
e-mail: mik2178@yandex.ru

Статья посвящена исследованию моделей полносвязной линейной регрессии, в которых наблюдаемые переменные содержат ошибки, а пары истинных переменных связаны между собой линейными функциональными зависимостями. При оценивании полносвязных регрессий главной проблемой считается правильный выбор соотношений дисперсий ошибок переменных. Если выбор сделан неверно, то оценки полносвязной регрессии будут смещенными. Цель статьи состоит в поиске областей возможных оценок основных параметров полносвязных регрессий в зависимости от возможных соотношений дисперсий ошибок переменных. Впервые с помощью элементов матричной алгебры решена обратная задача – получены аналитические зависимости соотношений дисперсий ошибок переменных от основных параметров. Эти зависимости позволяют идентифицировать области возможных оценок параметров, в которых выполняется необходимое условие экстремума целевой функции. Доказано, что при определенных условиях для любых соотношений дисперсий ошибок переменных оценки параметров всегда лежат внутри открытого выпуклого многоугольника, расположенного только в одном из ортантов многомерного пространства. При этом знаки оценок всегда согласуются со знаками соответствующих коэффициентов корреляции. Проведен численный эксперимент, подтверждающий корректность полученных результатов.

Ключевые слова: модель с ошибками в переменных, модель полносвязной линейной регрессии, взвешенный метод наименьших полных квадратов, оценка параметров, выпуклый многоугольник.

Для цитаты:
Базилевский М.П. Идентификация областей возможных оценок параметров моделей полносвязной линейной регрессии // Моделирование и анализ данных. 2023. Том 13. № 3. С. 52–65. DOI: https://doi.org/10.17759/mda.2023130304

Базилевский Михаил Павлович, кандидат технических наук, доцент кафедры математики, Иркутский государственный университет путей сообщения (ФГБОУ ВО ИрГУПС), г. Иркутск, Российская Федерация, ORCID: https://orcid.org/0000-0002-3253-5697, e-mail: mik2178@yandex.ru

1. ВВЕДЕНИЕ

В настоящее время при проведении регрессионного анализа большинство исследователей по-прежнему отдаёт предпочтение моделям множественной линейной регрессии [1], которые представляют собой зависимости выходной (объясняемой) переменной y от одной или нескольких входных (объясняющих) переменных x_1, x_2, ..., x_m, не содержащих ошибок. Оценки параметров таких регрессий найти довольно просто, например, с помощью метода наименьших квадратов. Реже применяются так называемые «errors-in-variables models» (EIV-модели) [2], в которых объясняющие переменные содержат ошибки. Такой подход к моделированию более реалистичен, но его применимость осложняет незнание в большинстве случае дисперсий ошибок переменных, из-за чего полученные результаты могут быть неточными. Тем не менее, EIV-модели достаточно подробно изучены. Для их оценки разработаны, например, метод ортогональной регрессии [3], метод наименьших полных квадратов [4] и др. Подробное описание этих и других методов со ссылками на первоисточники можно найти в монографии [5].

В той же монографии автором были предложены модели полносвязной линейной регрессии (МПЛР), в которых все наблюдаемые переменные содержат ошибки, а каждая истинная переменная связана со всеми другими линейными функциональными зависимостями. МПЛР имеет вид:

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

$$x_{ij}^* = a_j + b_j \cdot x_{im}^*,\ i=\overline{1,n},\ j=\overline{1,m-1}, \tag{2}$$

где n – объем выборки; x_{ij}, $i=\overline{1,n}$, $j=\overline{1,m}$ – наблюдаемые значения m взаимосвязанных переменных x_1, x_2, ..., x_m; x_{ij}^*, $i=\overline{1,n}$, $j=\overline{1,m}$ – их истинные значения, которые неизвестны; a_j, b_j, $j=\overline{1,m-1}$ – неизвестные параметры; $\varepsilon_i^{(x_j)}$, $i=\overline{1,n}$, $j=\overline{1,m}$ – ошибки переменных. Подчеркнём, что в МПЛР (1), (2) все переменные с одной стороны выходные (объясняемые), а с другой – входные (объясняющие), поэтому правильнее называть их взаимосвязанными. Понятно, что МПЛР следует строить при сильной корреляции всех переменных.

Таким образом, МПЛР (1), (2) можно считать новым видом EIV-моделей. Если $m=2$, то МПЛР вырождается в известную регрессию Деминга [6], которая находит широкое применение в клинической химии [7].

МПЛР (1), (2) оценивается с помощью взвешенного метода наименьших полных квадратов:

$$S\left(a_1,...,a_{m-1},b_1,...,b_{m-1},x_{1m}^*,...,x_{nm}^*\right)=\sum_{j=1}^{m-1}\lambda_j\sum_{i=1}^{n}\left(x_{ij}-a_j-b_j x_{im}^*\right)^2+\sum_{i=1}^{n}\left(x_{im}-x_{im}^*\right)^2\to\min, \tag{3}$$

где λ_j, $j=\overline{1,m-1}$ – положительные весовые коэффициенты, представляющие собой с вероятностно-статистической точки зрения отношения дисперсий ошибок переменных, т.е. $\lambda_j=\sigma^2_{\varepsilon(x_m)}\Big/\sigma^2_{\varepsilon(x_j)}$, $j=\overline{1,m-1}$. Геометрически это числа, задающие тип расстояния от точек $\left(x_{i1},x_{i2},...,x_{im}\right)$, $i=\overline{1,n}$ до линии регрессии.

В [8] доказано, что выбор связующей переменной x^*_m, входящей в правые части равенств (2), не влияет на решение задачи (3). А в [9,10] предложены алгоритмы численного решения задачи (3) при известных λ_j, $j=\overline{1,m-1}$.

Данная работа посвящена нахождению областей возможных оценок параметров b_j, $j=\overline{1,m-1}$, играющих главную роль при интерпретации МПЛР, в зависимости от любых значений λ_j, $j=\overline{1,m-1}$.

2. ПОИСК ЗАВИСИМОСТЕЙ КОЭФФИЦИЕНТОВ λ_j ОТ ПАРАМЕТРОВ b_j

При известных коэффициентах λ_j, $j=\overline{1,m-1}$ применение необходимых условий экстремума функции (3) приводит к нелинейной системе [5,9]:

$$b_p\cdot D_{x^*_m}=K_{x_px^*_m},\ p=\overline{1,m-1}, \tag{4}$$

где $K_{x_px^*_m}=\left(1+\sum_{j=1}^{m-1}\lambda_j b_j^2\right)^{-1}\left(\sum_{j=1}^{m-1}\lambda_j b_j K_{x_jx_p}+K_{x_mx_p}\right)$,

$$D_{x^*_m}=\left(D_{x_m}+\sum_{j=1}^{m-1}\lambda_j^2b_j^2D_{x_j}+2\sum_{j_1=1}^{m-2}\sum_{j_2=j_1+1}^{m-1}\lambda_{j_1}\lambda_{j_2}b_{j_1}b_{j_2}K_{x_{j_1}x_{j_2}}+2\sum_{j=1}^{m-1}\lambda_jb_jK_{x_jx_m}\right)\left(1+\sum_{j=1}^{m-1}\lambda_jb_j^2\right)^{-2},$$

символом D обозначены дисперсии переменных, K – ковариации.

Сначала была поставлена задача, используя уравнения (4), выразить коэффициенты λ_j, $j=\overline{1,m-1}$ от параметров b_j, $j=\overline{1,m-1}$.

Как это сделано в [5,9], обозначим $\lambda_pb_p=q_p$, $p=\overline{1,m-1}$. Тогда система (4) примет вид

$$\begin{pmatrix}A_1q_1-B & A_1q_2 & \dots & A_1q_{m-1}\\ A_2q_1 & A_2q_2-B & \dots & A_2q_{m-1}\\ \dots & \dots & \dots & \dots\\ A_{m-1}q_1 & A_{m-1}q_2 & \dots & A_{m-1}q_{m-1}-B\end{pmatrix}\times\begin{pmatrix}b_1\\ b_2\\ \dots\\ b_{m-1}\end{pmatrix}=-\begin{pmatrix}A_1\\ A_2\\ \dots\\ A_{m-1}\end{pmatrix}, \tag{5}$$

где $B=D_{x_m}+\sum_{j=1}^{m-1}q_j^2D_{x_j}+2\sum_{j_1=1}^{m-2}\sum_{j_2=j_1+1}^{m-1}q_{j_1}q_{j_2}K_{x_{j_1}x_{j_2}}+2\sum_{j=1}^{m-1}q_jK_{x_jx_m}$; $A_p=\sum_{j=1}^{m-1}q_jK_{x_jx_p}+K_{x_mx_p}$, $p=\overline{1,m-1}$.

Будем решать систему (5) относительно параметров b_j методом Крамера по формулам $b_p=\dfrac{\Delta^*_p}{\Delta^*}$, $p=\overline{1,m-1}$, где Δ^*_p – определитель, полученный из Δ^* путём замены p-го столбца на столбец свободных членов $\left(-A_1\ \ -A_2\ \ \dots\ \ -A_{m-1}\right)^T$.

В [9] доказано, что определитель основной матрицы системы (5)

$$\Delta^* = (-1)^{m-1}\, B^{m-2}\left(B - \sum_{j=1}^{m-1} A_j q_j\right).$$

Используя свойства определителей, найдём, например, Δ_1^*:

$$\Delta_1^* = \begin{vmatrix} -A_1 & A_1 q_2 & \dots & A_1 q_{m-1} \\ -A_2 & A_2 q_2 - B & \dots & A_2 q_{m-1} \\ \dots & \dots & \dots & \dots \\ -A_{m-1} & A_{m-1} q_2 & \dots & A_{m-1} q_{m-1} - B \end{vmatrix} = \begin{vmatrix} -A_1 & 0 & \dots & 0 \\ -A_2 & -B & \dots & 0 \\ \dots & \dots & \dots & \dots \\ -A_{m-1} & 0 & \dots & -B \end{vmatrix} = A_1 (-1)^{m-1}\, B^{m-2}.$$

Действуя по аналогии, получим

$$\Delta_p^* = A_p (-1)^{m-1}\, B^{m-2},\ p = \overline{1, m-1}.$$

Тогда по формулам Крамера впервые получим зависимости параметров b_j от переменных q_j:

$$b_p = \frac{A_p}{B - \sum_{j=1}^{m-1} A_j q_j},\ p = \overline{1, m-1}. \tag{6}$$

Число B в матричной форме имеет вид

$$B = \begin{pmatrix} q_1 \\ \dots \\ q_{m-1} \\ 1 \end{pmatrix}^T \begin{pmatrix} D_{x_1} & K_{x_1 x_2} & \dots & K_{x_1 x_m} \\ K_{x_1 x_2} & D_{x_2} & \dots & K_{x_2 x_m} \\ \dots & \dots & \dots & \dots \\ K_{x_1 x_m} & K_{x_2 x_m} & \dots & D_{x_m} \end{pmatrix} \begin{pmatrix} q_1 \\ \dots \\ q_{m-1} \\ 1 \end{pmatrix},$$

а сумма

$$\sum_{j=1}^{m-1} A_j q_j = \begin{pmatrix} q_1 \\ \dots \\ q_{m-1} \\ 1 \end{pmatrix}^T \begin{pmatrix} D_{x_1} & K_{x_1 x_2} & \dots & K_{x_1 x_{m-1}} & 0 \\ K_{x_1 x_2} & D_{x_2} & \dots & K_{x_2 x_{m-1}} & 0 \\ \dots & \dots & \dots & \dots & \dots \\ K_{x_1 x_m} & K_{x_2 x_m} & \dots & K_{x_{m-1} x_m} & 0 \end{pmatrix} \begin{pmatrix} q_1 \\ \dots \\ q_{m-1} \\ 1 \end{pmatrix}.$$

Тогда

$$B - \sum_{j=1}^{m-1} A_j q_j = \begin{pmatrix} q_1 \\ \dots \\ q_{m-1} \\ 1 \end{pmatrix}^T \begin{pmatrix} 0 & \dots & 0 & K_{x_1 x_m} \\ 0 & \dots & 0 & K_{x_2 x_m} \\ \dots & \dots & 0 & \dots \\ 0 & \dots & 0 & D_{x_m} \end{pmatrix} \begin{pmatrix} q_1 \\ \dots \\ q_{m-1} \\ 1 \end{pmatrix} = \sum_{j=1}^{m-1} q_j K_{x_j x_m} + D_{x_m}.$$

Следовательно, формулы (6) принимают вид:

$$b_p = \frac{\sum_{j=1}^{m-1} q_j K_{x_j x_p} + K_{x_m x_p}}{\sum_{j=1}^{m-1} q_j K_{x_j x_m} + D_{x_m}}, \quad p = \overline{1, m-1}. \tag{7}$$

А используя формулы (7), запишем линейную систему с переменными q_j:

$$\sum_{j=1}^{m-1} \left(b_p K_{x_j x_m} - K_{x_j x_p} \right) q_j = K_{x_m x_p} - b_p D_m, \quad p = \overline{1, m-1}, \tag{8}$$

которая может быть представлена в матричной форме:

$$\left[\begin{pmatrix} b_1 \\ b_2 \\ \dots \\ b_{m-1} \end{pmatrix} \begin{pmatrix} K_{x_1 x_m} \\ K_{x_2 x_m} \\ \dots \\ K_{x_{m-1} x_m} \end{pmatrix}^T - \begin{pmatrix} D_{x_1} & K_{x_1 x_2} & \dots & K_{x_1 x_{m-1}} \\ K_{x_1 x_2} & D_{x_2} & \dots & K_{x_2 x_{m-1}} \\ \dots & \dots & \dots & \dots \\ K_{x_1 x_{m-1}} & K_{x_2 x_{m-1}} & \dots & D_{x_{m-1}} \end{pmatrix} \right] \times \begin{pmatrix} q_1 \\ q_2 \\ \dots \\ q_{m-1} \end{pmatrix} = \begin{pmatrix} K_{x_1 x_m} \\ K_{x_2 x_m} \\ \dots \\ K_{x_{m-1} x_m} \end{pmatrix} - D_m \begin{pmatrix} b_1 \\ b_2 \\ \dots \\ b_{m-1} \end{pmatrix}.$$

Будем решать систему (8) относительно переменных q_j методом Крамера по формулам $q_p = \frac{\Delta_p}{\Delta}$, $p = \overline{1, m-1}$, где Δ – определитель основной матрицы, а Δ_p – определитель, полученный из Δ путём замены p-го столбца на столбец свободных членов.

Если $m = 2$, то из (8) вытекает необходимое условие $\left(K_{x_1 x_2} b_1 - D_{x_1} \right) \cdot q_1 = \left(K_{x_1 x_2} - D_{x_2} b_1 \right)$ для регрессии Деминга, решение которого по методу Крамера можно записать в виде:

$$q_1 = \frac{\Delta_1}{\Delta} = \left(- \begin{vmatrix} D_{x_1} & b_1 \\ K_{x_1 x_2} & 1 \end{vmatrix} \right) \Big/ \left(- \begin{vmatrix} b_1 & K_{x_1 x_2} \\ 1 & D_{x_2} \end{vmatrix} \right).$$

Если $m = 3$, то система (8) принимает вид:

$$\begin{pmatrix} K_{x_1 x_3} b_1 - D_{x_1} & K_{x_2 x_3} b_1 - K_{x_1 x_2} \\ K_{x_1 x_3} b_2 - K_{x_1 x_2} & K_{x_2 x_3} b_2 - D_{x_2} \end{pmatrix} \times \begin{pmatrix} q_1 \\ q_2 \end{pmatrix} = \begin{pmatrix} K_{x_1 x_3} - D_{x_3} b_1 \\ K_{x_2 x_3} - D_{x_3} b_2 \end{pmatrix}.$$

Используя свойства определителей, найдем Δ:

$$\Delta = \begin{vmatrix} K_{x_1 x_3} b_1 - D_{x_1} & K_{x_2 x_3} b_1 - K_{x_1 x_2} \\ K_{x_1 x_3} b_2 - K_{x_1 x_2} & K_{x_2 x_3} b_2 - D_{x_2} \end{vmatrix} =$$

$$= \begin{vmatrix} K_{x_1 x_3} b_1 & K_{x_2 x_3} b_1 \\ K_{x_1 x_3} b_2 - K_{x_1 x_2} & K_{x_2 x_3} b_2 - D_{x_2} \end{vmatrix} - \begin{vmatrix} D_{x_1} & K_{x_1 x_2} \\ K_{x_1 x_3} b_2 - K_{x_1 x_2} & K_{x_2 x_3} b_2 - D_{x_2} \end{vmatrix} =$$

$$= \begin{vmatrix} K_{x_1 x_3} b_1 & K_{x_2 x_3} b_1 \\ K_{x_1 x_3} b_2 & K_{x_2 x_3} b_2 \end{vmatrix} - \begin{vmatrix} K_{x_1 x_3} b_1 & K_{x_2 x_3} b_1 \\ K_{x_1 x_2} & D_{x_2} \end{vmatrix} - \begin{vmatrix} D_{x_1} & K_{x_1 x_2} \\ K_{x_1 x_3} b_2 & K_{x_2 x_3} b_2 \end{vmatrix} + \begin{vmatrix} D_{x_1} & K_{x_1 x_2} \\ K_{x_1 x_2} & D_{x_2} \end{vmatrix} =$$

$$= b_1 \begin{vmatrix} K_{x_1x_2} & D_{x_2} \\ K_{x_1x_3} & K_{x_2x_3} \end{vmatrix} - b_2 \begin{vmatrix} D_{x_1} & K_{x_1x_2} \\ K_{x_1x_3} & K_{x_2x_3} \end{vmatrix} + \begin{vmatrix} D_{x_1} & K_{x_1x_2} \\ K_{x_1x_2} & D_{x_2} \end{vmatrix} = \begin{vmatrix} D_{x_1} & K_{x_1x_2} & b_1 \\ K_{x_1x_2} & D_{x_2} & b_2 \\ K_{x_1x_3} & K_{x_2x_3} & 1 \end{vmatrix}.$$

Аналогично, можно получить:

$$\Delta_1 = \begin{vmatrix} b_1 & K_{x_1x_2} & K_{x_1x_3} \\ b_2 & D_{x_2} & K_{x_2x_3} \\ 1 & K_{x_2x_3} & D_{x_3} \end{vmatrix}, \ \Delta_2 = \begin{vmatrix} D_{x_1} & b_1 & K_{x_1x_3} \\ K_{x_1x_2} & b_2 & K_{x_2x_3} \\ K_{x_1x_3} & 1 & D_{x_3} \end{vmatrix}.$$

Таким образом, при $m = 3$ переменная q_1 в системе (8) находится как отношение определителя, полученного из матрицы ковариаций переменных $V_{3\times3} = \begin{pmatrix} D_{x_1} & K_{x_1x_2} & K_{x_1x_3} \\ K_{x_1x_2} & D_{x_2} & K_{x_2x_3} \\ K_{x_1x_3} & K_{x_2x_3} & D_{x_3} \end{pmatrix}$ путем замены первого столбца на вектор $\begin{pmatrix} b_1 & b_2 & 1 \end{pmatrix}^T$, к определителю, полученному из матрицы $V_{3\times3}$ путем замены третьего столбца на тот же вектор. Аналогично находится переменная q_2.

То же самое происходит при $m = 4$. Причем, перед определителями Δ_1, Δ_2, Δ_3 и Δ появляется знак «–». Но это никак не влияет на переменные q_1, q_2 и q_3.

Обобщим полученный результат. Переменные q_p зависят от параметров λ_p по формулам:

$$q_p = \frac{\Delta_p}{\Delta_m}, \ p = \overline{1, m-1}, \tag{9}$$

где Δ_i – определитель, полученный из матрицы ковариаций $V_{m\times m}$ путём замены i-го столбца на столбец $\begin{pmatrix} b_1 & b_2 & \dots & b_{m-1} & 1 \end{pmatrix}^T$.

Учитывая, что $\lambda_p b_p = q_p$, $p = \overline{1, m-1}$, из (9) получим зависимости коэффициентов λ_j, $j = \overline{1, m-1}$ от параметров b_j, $j = \overline{1, m-1}$:

$$\lambda_p = \frac{\Delta_p}{b_p \cdot \Delta_m}, \ p = \overline{1, m-1}. \tag{10}$$

3. ПОИСК ОБЛАСТИ ВОЗМОЖНЫХ ОЦЕНОК ПАРАМЕТРОВ b_j

Поскольку по определению λ_j, $j = \overline{1, m-1}$ – положительные коэффициенты, то, используя зависимости (10), получим условия, которым удовлетворяют все стационарные точки функции (3):

$$\frac{\Delta_p}{b_p \cdot \Delta_m} > 0, \ p = \overline{1, m-1}. \tag{11}$$

Если из (10) выразить произведения $\lambda_p b_p$, $p=\overline{1,m-1}$, и подставить их в правую часть системы (4), то получим, что $D^*_{x_m}=\left(1+\sum_{j=1}^{m-1}\lambda_j b_j^2\right)^{-1}\cdot\Delta_m^{-1}\cdot\left|V_{m\times m}\right|$. Из этого следует, что определитель Δ_m в (11) всегда положителен.

В [5] установлено, что для любых $\lambda_j>0$, $j=\overline{1,m-1}$, система (4) всегда имеет 2^{m-1} решений. Причем, все эти точки в $(m-1)$-мерном пространстве расположены в разных ортантах и только в одной из них целевая функция (3) достигает минимум. Следовательно, поскольку определители Δ_i, $i=\overline{1,m}$ представляют собой линейные комбинации параметров b_j, $j=\overline{1,m-1}$, то решение системы неравенств (11) представляет собой совокупность 2^{m-1} открытых выпуклых областей.

Введём матрицу $\mathrm{M}_{(m-1)\times m}$, полученную из матрицы ковариаций $V_{m\times m}$ как результат поэлементного деления в ней первых $(m-1)$ строк на последнюю строку. И введём в рассмотрение m точек, координатами которых являются элементы столбцов матрицы M:

$$P_1\left(\frac{D_{x_1}}{K_{x_1x_m}},\frac{K_{x_1x_2}}{K_{x_1x_m}},\frac{K_{x_1x_3}}{K_{x_1x_m}},...,\frac{K_{x_1x_{m-1}}}{K_{x_1x_m}}\right),\ P_2\left(\frac{K_{x_1x_2}}{K_{x_2x_m}},\frac{D_{x_2}}{K_{x_2x_m}},\frac{K_{x_2x_3}}{K_{x_2x_m}},...,\frac{K_{x_2x_{m-1}}}{K_{x_2x_m}}\right),$$

$$\ldots,$$

$$P_m\left(\frac{K_{x_1x_m}}{D_{x_m}},\frac{K_{x_2x_m}}{D_{x_m}},\frac{K_{x_3x_m}}{D_{x_m}},...,\frac{K_{x_{m-1}x_m}}{D_{x_m}}\right).$$

Рассмотрим определитель Δ_1. Преобразуем его следующим образом:

$$\Delta_1=\begin{vmatrix} b_1 & K_{x_1x_2} & \ldots & K_{x_1x_m}\\ b_2 & D_{x_2} & \ldots & K_{x_2x_m}\\ \ldots & \ldots & \ldots & \ldots\\ 1 & K_{x_2x_m} & \ldots & D_{x_m}\end{vmatrix}=\begin{vmatrix} b_1 & K_{x_1x_2}-K_{x_2x_m}b_1 & \ldots & K_{x_1x_m}-D_{x_m}b_1\\ b_2 & D_{x_2}-K_{x_2x_m}b_2 & \ldots & K_{x_2x_m}-D_{x_m}b_2\\ \ldots & \ldots & \ldots & \ldots\\ 1 & 0 & \ldots & 0\end{vmatrix}.$$

Отсюда следует, что $\Delta_1=0$ в точках P_2, P_3, …, P_{m-1}, P_m.

Аналогично можно установить, что $\Delta_2=0$ в точках P_1, P_3, …, P_{m-1}, P_m, $\Delta_3=0$ в точках P_1, P_2, P_4, P_5, ..., P_{m-1}, P_m и т.д. Тогда справедливо следующее утверждение: точка P_k является точкой пересечения плоскостей $\Delta_j=0$, $j\in\{1,2,...,m-1\}\setminus\{k\}$.

Рассмотрим некоторые возможные решения задачи (3) при условии, когда только у одной из переменных дисперсия ошибок стремится к нулю.

Если $\sigma^2_{\varepsilon(x_m)}\to 0$, то $\lambda_j\to 0$, $j=\overline{1,m-1}$. В этом случае $x^*_{im}\to x_{im}$, $i=\overline{1,n}$, поэтому оценки параметров b_j МПЛР стремятся к оценкам соответствующих моделей парной линейной регрессии x_1 от x_m, x_2 от x_m, …, x_{m-1} от x_m:

$$b_1\to\frac{K_{x_1x_m}}{D_{x_m}},\ b_2\to\frac{K_{x_2x_m}}{D_{x_m}},\ ...,\ b_{m-1}\to\frac{K_{x_{m-1}x_m}}{D_{x_m}}.$$

Иными словами, оценки параметров b_j МПЛР стремятся к координатам точки P_m.

Если $\sigma^2_{\varepsilon^{(x_1)}} \to 0$, то $\lambda_1 \to \infty$. В этом случае $x_{i1} - a_1 - b_1 x^*_{im} \to 0$, $i = \overline{1,n}$, откуда $x^*_{im} \to -\dfrac{a_1}{b_1} + \dfrac{1}{b_1} x_{i1}$, $i = \overline{1,n}$. Поэтому оценки параметров b_j МПЛР связаны с оценками моделей парной линейной регрессии x_2 от x_1, x_3 от x_1, …, x_m от x_1 следующими соотношениями:

$$\frac{b_2}{b_1} \to \frac{K_{x_1x_2}}{D_{x_1}}, \ \frac{b_3}{b_1} \to \frac{K_{x_1x_3}}{D_{x_1}}, \ \ldots, \ \frac{b_{m-1}}{b_1} \to \frac{K_{x_1x_{m-1}}}{D_{x_1}}, \ \frac{1}{b_1} \to \frac{K_{x_1x_m}}{D_{x_1}},$$

откуда

$$b_1 \to \frac{D_{x_1}}{K_{x_1x_m}}, \ b_2 \to \frac{K_{x_1x_2}}{K_{x_1x_m}}, \ \ldots, \ b_{m-1} \to \frac{K_{x_1x_{m-1}}}{K_{x_1x_m}}.$$

Иными словами, оценки параметров b_j МПЛР стремятся к координатам точки P_1.

Аналогично можно получить оставшиеся решения задачи (3) при условиях $\sigma^2_{\varepsilon^{(x_j)}} \to 0$, $j = \overline{2,m-1}$. Таким образом, при $\sigma^2_{\varepsilon^{(x_j)}} \to 0$, $j = \overline{1,m}$, оценки параметров b_j МПЛР стремятся к координатам соответствующих точек P_j, $j = \overline{1,m}$. Из этого следует, что в зависимости от коэффициентов λ_j, $j = \overline{1,m-1}$, оценки параметров b_j могут быть противоречивыми, т.е. для них могут не выполняться условия $K_{x_jx_m} b_j > 0$, $j = \overline{1,m-1}$, что негативно сказывается на интерпретационных характеристиках МПЛР.

Для того чтобы условия $K_{x_jx_m} b_j > 0$, $j = \overline{1,m-1}$, выполнялись всегда, т.е. для любых коэффициентов λ_j, $j = \overline{1,m-1}$, необходимо потребовать, чтобы все точки P_j, $j = \overline{1,m}$, лежали в одном ортанте $(m-1)$-мерного пространства. Это требование равносильно тому, чтобы в матрице M элементы каждой строки были одинакового знака.

Теорема. Если в матрице M элементы каждой строки одного знака, то для любых $\lambda_j > 0$, $j = \overline{1,m-1}$, оценки параметров b_j, $j = \overline{1,m-1}$, МПЛР (1), (2) для метода (3) всегда лежат внутри открытого выпуклого m-угольника с вершинами в точках P_j, $j = \overline{1,m}$, расположенного в том ортанте $(m-1)$-мерного пространства, в котором $K_{x_jx_m} b_j > 0$, $j = \overline{1,m-1}$.

Доказательство. Если в матрице M элементы каждой строки одного знака, то точки P_j, $j = \overline{1,m}$, лежат в одном ортанте $(m-1)$-мерного пространства. При этом будут справедливы равенства

$$\operatorname{sgn}\left(K_{x_jx_m}\right) = \operatorname{sgn}\left(\frac{K_{x_jx_p}}{K_{x_px_m}}\right), \ j = \overline{1,m-1}, \ p = \overline{1,m}, \tag{12}$$

т.е. первые координаты точек P_j, $j = \overline{1,m}$, имеют знак ковариации $K_{x_1x_m}$, вторые – знак ковариации $K_{x_2x_m}$ и т.д., поэтому все эти точки лежат в ортанте θ, в котором $K_{x_jx_m} b_j > 0$, $j = \overline{1,m-1}$.

Для ортанта θ условия (11) упрощаются:

$$\frac{\Delta_p}{\Delta_m} K_{x_p x_m} > 0\,,\ \ p = \overline{1, m-1}\,. \tag{13}$$

Учитывая, что $\Delta_m > 0$, из (13) следуют равенства

$$\operatorname{sgn}\left(\Delta_j\right) = \operatorname{sgn}\left(K_{x_j x_m}\right),\ \ j = \overline{1, m-1}\,. \tag{14}$$

Идентифицируем выпуклую область в ортанте θ. Рассмотрим в нём открытый выпуклый многоугольник (m-угольник), вершинами которого являются точки P_j, $j = \overline{1, m}$, а сторонами – плоскости $\Delta_j = 0$, $j = \overline{1, m}$. На плоскости $\Delta_m = 0$ из точек P_j, $j = \overline{1, m}$, не лежит только точка P_m. Подставим координаты точки P_m в определитель Δ_m:

$$\Delta_m\left(P_m\right) = \begin{vmatrix} D_{x_1} & K_{x_1 x_2} & \dots & \dfrac{K_{x_1 x_m}}{D_{x_m}} \\ K_{x_1 x_2} & D_{x_2} & \dots & \dfrac{K_{x_2 x_m}}{D_{x_m}} \\ \dots & \dots & \dots & \dots \\ K_{x_1 x_m} & K_{x_2 x_m} & \dots & 1 \end{vmatrix} = \frac{1}{D_{x_m}} \begin{vmatrix} D_{x_1} & K_{x_1 x_2} & \dots & K_{x_1 x_m} \\ K_{x_1 x_2} & D_{x_2} & \dots & K_{x_2 x_m} \\ \dots & \dots & \dots & \dots \\ K_{x_1 x_m} & K_{x_2 x_m} & \dots & D_{x_m} \end{vmatrix} > 0.$$

Таким образом, решением неравенства $\Delta_m > 0$ будет полупространство, направленное внутрь m-угольника.

Аналогично можно определить, что $\Delta_j\left(P_j\right) = \dfrac{1}{K_{x_j x_m}} \left|V_{m\times m}\right|$, $j = \overline{1, m-1}$, поэтому решением неравенств $\Delta_p \cdot K_{x_p x_m} > 0$, $p = \overline{1, m-1}$, будут полупространства, также направленные внутрь m-угольника.

Таким образом, единственной областью в ортанте θ, каждая точка которой удовлетворяет системе (13), будет открытый выпуклый m-угольник с вершинами P_j, $j = \overline{1, m}$.

В [5,10] установлено, что необходимое условие минимума функции (3) представляет собой систему неравенств

$$b_p \cdot G_p > 0\,,\ \ p = \overline{1, m-1}\,, \tag{15}$$

где $G_p = K_{x_p x_m} + \sum\limits_{j\in\{1,\dots,m-1\}\setminus p} \lambda_j b_j K_{x_j x_p}$.

Учитывая зависимости (10), получим

$$G_p = K_{x_p x_m} + \Delta_m^{-1} \sum_{j\in\{1,\dots,m-1\}\setminus p} \Delta_j \cdot K_{x_j x_p}\,,\ \ p = \overline{1, m-1}\,.$$

Из соотношений (12), (14) следует, что $\operatorname{sgn}\left(\Delta_j\right) = \operatorname{sgn}\left(\dfrac{K_{x_j x_p}}{K_{x_p x_m}}\right)$, $j = \overline{1, m-1}$, $p = \overline{1, m}$. Откуда $\operatorname{sgn}\left(\Delta_j \cdot K_{x_j x_p}\right) = \operatorname{sgn}\left(K_{x_p x_m}\right)$, $j = \overline{1, m-1}$, $p = \overline{1, m}$. Это означает, что

$\operatorname{sgn}\left(G_p\right)=\operatorname{sgn}\left(K_{x_p x_m}\right)$, $p=\overline{1,m-1}$. Тогда необходимое условие минимума (15) для стационарных точек из ортанта θ примет вид $b_p \cdot K_{x_p x_m}>0$. Таким образом, условие (15) выполняется в любой точке открытого выпуклого m-угольника с вершинами P_j, $j=\overline{1,m}$.

Поскольку необходимое условие минимума выполняется абсолютно во всех точках открытого выпуклого m-угольника с вершинами P_j, $j=\overline{1,m}$, расположенного в ортанте θ, то оно никак не может выполняться в других ортантах $(m-1)$-мерного пространства. Следовательно, для любых $\lambda_j>0$, $j=\overline{1,m-1}$, оценки параметров b_j, $j=\overline{1,m-1}$, МПЛР всегда лежат внутри этого m-угольника.

Теорема доказана.

Вопрос о том, выполняется ли достаточное условие минимума во всех точках выпуклого m-угольника, пока остаётся открытым. Тем не менее, из теоремы вытекает следующее важное следствие.

Следствие. Если в матрице M элементы каждой строки одного знака, то для любых $\lambda_j>0$, $j=\overline{1,m-1}$, знаки оценок параметров b_j, $j=\overline{1,m-1}$, МПЛР (1), (2) для метода (3) всегда согласуются со знаками соответствующих коэффициентов корреляции $r_{x_j x_m}$, $j=\overline{1,m-1}$.

4. ЧИСЛЕННЫЙ ЭКСПЕРИМЕНТ

Решалась задача построения МПЛР по ежегодным статистическим данным о численности и составе населения в Иркутской области (https://rosstat.gov.ru/) за период 2000–2020 гг. по следующим переменным:

x_1 – население в трудоспособном возрасте (в процентах от общей численности населения);

x_2 – численность рабочей силы (тыс. человек);

x_3 – численность пенсионеров (тыс. человек).

Вычисленная по этим данным матрица ковариаций имеет вид:

$$V=\begin{pmatrix} 9,096 & 121,234 & -93,284 \\ 121,234 & 2403,759 & -1451,256 \\ -93,284 & -1451,256 & 1176,521 \end{pmatrix}.$$

По этой матрице были определены коэффициенты парной корреляции переменных $r_{12}=0,8199$, $r_{13}=-0,9017$, $r_{23}=-0,863$, что подтверждает их весьма тесную линейную зависимость.

Матрица M в этом случае имеет вид:

$$\mathrm{M}=\begin{pmatrix} -0,0975 & -0,0835 & -0,0793 \\ -1,2996 & -1,6563 & -1,2335 \end{pmatrix}.$$

В этой матрице элементы каждой строки одного знака, поэтому в силу доказанной теоремы, для любых коэффициентов λ_j, $j=\overline{1,m-1}$, область возможных оценок

параметров b_j, $j=\overline{1,m-1}$, будет представлять собой открытый треугольник (рис. 1) с вершинами $P_1\left(-0.0975,-1.2996\right)$, $P_2\left(-0.0835,-1.6563\right)$, $P_3\left(-0.0793,-1.2335\right)$.

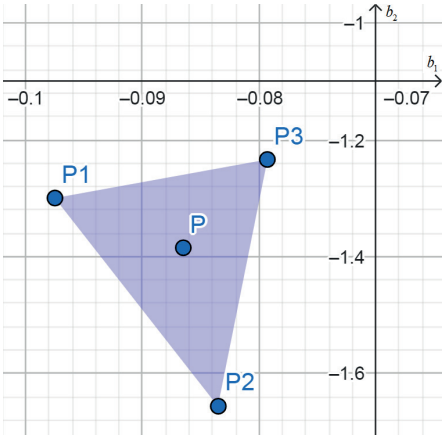

Рис. 1. Область возможных оценок МПЛР

Как видно, размеры полученной области (рис. 1) возможных оценок параметров b_j, $j=\overline{1,m-1}$, оказались довольно малы. Из этого следует, что даже если исследователь выберет значения коэффициентов λ_j, $j=\overline{1,m-1}$, случайно, то разница между полученными и несмещенными оценками МПЛР будет несущественна. В целом, если все абсолютные значения коэффициентов взаимосвязанных переменных в МПЛР стремятся к 1, то область возможных оценок параметров b_j, $j=\overline{1,m-1}$, сужается в точку.

В [10] коэффициенты λ_j, $j=\overline{1,m-1}$, предложено выбирать по формулам $\lambda_j = D_{x_m} \big/ D_{x_j}$, $j=\overline{1,m-1}$. В нашем примере $\lambda_1 = D_{x_3} / D_{x_1} = 129,34$, $\lambda_2 = D_{x_3} \big/ D_{x_2} = 0,489$. Численно полученные для этих значений оценки параметров b_1 и b_2 составили $\tilde{b}_1 = -0,08653$, $\tilde{b}_2 = -1,38486$. На рис. 1 этим оценкам соответствует точка P, расположенная очень близко к середине треугольника. Тем самым, точку P можно считать точечной характеризацией построенной области. В этой точке оцененная МПЛР имеет вид:

$$\tilde{x}_1 = 121,505 - 0,0865\tilde{x}_3,\ \tilde{x}_2 = 2247,444 - 1,3848\tilde{x}_3.$$

Механизм интерпретации МПЛР, в частности, модели (16), будет исследован в будущих работах автора.

5. ЗАКЛЮЧЕНИЕ

В работе доказано, что если в матрице M элементы каждой строки одного знака, что всегда выполняется при сильной корреляции всех пар взаимосвязанных

переменных, то для любых $\lambda_j > 0$, $j = \overline{1, m-1}$, оценки параметров b_j, $j = \overline{1, m-1}$, МПЛР всегда лежат внутри открытого выпуклого m-угольника и удовлетворяют неравенствам $K_{x_j x_m} b_j > 0$, $j = \overline{1, m-1}$. Таким образом, во-первых, условие, накладываемое на элементы матрицы M, можно считать критерием применимости полносвязных регрессий: если оно выполняется, то какими бы ни были дисперсии ошибок переменных, знаки оценок параметров b_j, $j = \overline{1, m-1}$, всегда будут согласованы со знаками соответствующих коэффициентов корреляции, т.е. даже случайный выбор коэффициентов $\lambda_j > 0$, $j = \overline{1, m-1}$, будет давать оценки, согласованные по знакам с несмещенными. Во-вторых, если дисперсии ошибок переменных неизвестны, то какими бы они ни были, всегда можно точно определить, в каких пределах будут меняться оценки параметров b_j, $j = \overline{1, m-1}$. Для этого достаточно найти в каждой строке матрицы M минимальный и максимальный элемент и составить соответствующие промежутки для каждого параметра. Причём, если $m = 2, 3, 4$, то выпуклый многоугольник можно изобразить графически. При сильной корреляции переменных выпуклый многоугольник такой узкий, что выбор в качестве оценок координат точки его центра сводит к минимуму разницу между ними и несмещенными оценками. В дальнейшем автор планирует исследовать вопросы интерпретации оценок МПЛР.

Литература

1. *Montgomery D.C., Peck E.A., Vining G.G.* Introduction to linear regression analysis. John Wiley & Sons, 2021.
2. *Xu P.* Improving the weighted least squares estimation of parameters in errors-in-variables models // Journal of the Franklin Institute. 2019. Vol. 356. № 15. P. 8785–8802. DOI:10.1016/j.jfranklin.2019.06.016
3. *Демиденко Е.З.* Линейная и нелинейная регрессии. М.: Финансы и статистика, 1981. 304 с.
4. *Golub G.H., Van Loan C.F.* An analysis of the total least squares problem // SIAM Journal on Numerical Analysis. 1980. Vol. 17. № 6. P. 883–893.
5. *Базилевский М.П.* Методы построения регрессионных моделей с ошибками во всех переменных. Иркутск: ИрГУПС, 2019. 208 с.
6. *Deming W.E.* Statistical adjustment of data. New York, Wiley, 2011. 288 p.
7. *Koh N.W.X., Markus C., Loh T.P., Lim C.Y.* Comparison of six regression-based lot-to-lot verification approaches // Clinical Chemistry and Laboratory Medicine. 2022. Vol. 60. № 8. P. 1175–1185. DOI:10.1515/cclm-2022-0274
8. *Базилевский М.П.* Исследование поведения относительных вкладов переменных в общую детерминацию в оцененном на основе метода выпрямления искаженных коэффициентов регрессионном уравнении // Вестник СибГУТИ. 2022. № 1(57). С. 89–96.
9. *Базилевский М.П.* Многофакторные модели полносвязной линейной регрессии без ограничений на соотношения дисперсий ошибок переменных // Информатика и её применения. 2020. Т. 14. № 2. С. 92–97. DOI:10.14357/19922264200213
10. *Базилевский М.П.* Метод выпрямления искаженных из-за мультиколлинеарности коэффициентов в регрессионных моделях // Информатика и её применения. 2021. Т. 15. № 2. С. 60–65. DOI:10.14357/19922264210209

Identification of Possible Estimates Areas for Parameters of Fully connected Linear Regression Models

Mikhail P. Bazilevskiy*
Irkutsk State Transport University (ISTU), Irkutsk, Russia
ORCID: https://orcid.org/0000-0002-3253-5697
e-mail: mik2178@yandex.ru

This article is devoted to the study of fully connected linear regression models, in which the observed variables contain errors, and the pairs of true variables are interconnected by linear functional dependencies. When estimating fully connected regressions, the main problem is the correct choice of the error variances ratios of the variables. If the choice is made incorrectly, then the fully connected regression estimates will be biased. The purpose of this article is to find the dependence of main parameters possible estimates areas on the possible error variances ratios of the variables in fully connected regressions. For the first time, with the help of matrix algebra elements, the inverse problem is solved – analytical dependences of the error variances ratios of variables on the main parameters are obtained. These dependences make it possible to identify the parameters possible estimates areas in which the necessary condition for the extremum of the objective function is satisfied. It is proved that, under certain conditions, for any error variances ratios of the variables, the parameters estimates always lie inside an open convex polygon located only in one of the orthants of the multidimensional space. In this case, the signs of the estimates always agree with the signs of the corresponding correlation coefficients. A numerical experiment was carried out, confirming the correctness of the results obtained.

Keywords: errors-in-variables model, fully connected linear regression model, weighted total least squares, parameter estimation, convex polygon.

For citation:
Bazilevskiy M.P. Identification of possible estimates areas for parameters of fully connected linear regression models. *Modelirovanie i analiz dannykh = Modelling and Data Analysis*, 2023. Vol. 13, no. 3, pp. 52–65. DOI: 10.17759/mda.2023130304 (In Russ., abstr. in Engl.).

References

1. Montgomery D.C., Peck E.A., Vining G.G. *Introduction to linear regression analysis*. John Wiley & Sons, 2021.
2. Xu P. Improving the weighted least squares estimation of parameters in errors-in-variables models, *Journal of the Franklin Institute*, 2019, vol. 356, no. 15, pp. 8785–8802. DOI:10.1016/j.jfranklin.2019.06.016

Mikhail P. Bazilevskiy, PhD (Engineering), Associate Professor, Department of Mathematics, Irkutsk State Transport University (ISTU), Irkutsk, Russia, ORCID: https://orcid. org/0000-0002-3253-5697, e-mail: mik2178@yandex.ru

3. Demidenko E.Z. *Lineynaya i nelineynaya regressii* [Linear and nonlinear regressions]. Moscow, Finansy i statistika, 1981. 304 p.
4. Golub G.H., Van Loan C.F. An analysis of the total least squares problem, *SIAM Journal on Numerical Analysis*, 1980, vol. 17, no. 6, pp. 883–893.
5. Bazilevskiy M.P. *Metody postroeniya regressionnykh modeley s oshibkami vo vsekh peremennykh* [Methods for constructing errors-in-variables regression models]. Irkutsk, IrGUPS, 2019. 208 p.
6. Deming W.E. *Statistical adjustment of data*. New York, Wiley, 2011. 288 p.
7. Koh N.W.X., Markus C., Loh T.P., Lim C.Y. Comparison of six regression-based lot-to-lot verification approaches, *Clinical Chemistry and Laboratory Medicine*, 2022, vol. 60, no. 8, pp. 1175–1185. DOI:10.1515/cclm-2022-0274
8. Bazilevskiy M.P. Issledovanie povedeniya otnositel'nykh vkladov peremennykh v obshchuyu determinatsiyu v otsenennom na osnove metoda vypryamleniya iskazhennykh koeffitsientov regressionnom uravnenii [Researching the behavior of variables relative contributions to the total determination in regression equation estimated using the method of distorted coefficients straightening], *The Herald of the Siberian State University of Telecommunications and Information Science*, 2022, no. 1(57), pp. 89–96.
9. Bazilevskiy M.P. Mnogofaktornye modeli polnosvyaznoy lineynoy regressii bez ogranicheniy na sootnosheniya dispersiy oshibok peremennykh [Multifactor fully connected linear regression models without constraints to the ratios of variables errors variances], *Informatics and Applications*, 2020, vol. 14, no. 2, pp. 92–97. DOI:10.14357/19922264200213
10. Bazilevskiy M.P. Metod vypryamleniya iskazhennykh iz-za mul'tikollinearnosti koeffitsientov v regressionnykh modelyakh [Method of straightening distorted due to multicollinearity coefficients in regression models], *Informatics and Applications,* 2021, vol. 15, no. 2, pp. 60–65. DOI:10.14357/19922264210209

Получена 10.07.2023
Принята в печать 09.08.2023

Received 10.07.2023
Accepted 09.08.2023