Интерпретация оценок параметров моделей полносвязной линейной регрессии

М. П. Базилевский

***Аннотация*—Данная работа посвящена исследованию вопросов интерпретации оценок параметров моделей полносвязной линейной регрессии. В таких моделях все наблюдаемые переменные содержат ошибки, а истинные переменные связаны между собой линейными функциональными зависимостями. Частным случаем полносвязной регрессии является хорошо изученная регрессия Деминга. Ранее для оценивания полносвязных регрессий применялся взвешенный метод наименьших полных квадратов. В данной статье установлено, что полученные таким методом оценки полносвязной линейной регрессии совпадают с оценками метода максимального правдоподобия. Выявлено, что интерпретировать полносвязные регрессии по аналогии с множественными регрессиями нельзя, поскольку первые строятся в предположении, что все переменные сильно коррелируют между собой. Доказана теорема, согласно которой одновременное увеличение в оцененной модели полносвязной линейной регрессии значений наблюдаемых переменных на определенные величины приводит к увеличению оценок истинных значений переменных на те же самые величины. С использованием этого факта можно интерпретировать любую модель полносвязной линейной регрессии, что продемонстрировано на примере моделирования таких макроэкономических показателей Иркутской области, как оборот оптовой и розничной торговли, а также продукция сельского хозяйства.**

***Ключевые слова*—регрессионный анализ, модель полносвязной линейной регрессии, ошибки в переменных, интерпретация, метод максимального правдоподобия.**

I. Введение

При проведении регрессионного анализа игнорирование ошибок в объясняющих (входных) переменных может приводить к получению неверных выводов, основанных на вычисленных оценках параметров модели [1]. Поэтому вместо так называемой наивной схемы, когда, например, модель множественной линейной регрессии без ошибок в объясняющих переменных оценивается с помощью метода наименьших квадратов или модулей, следует использовать методы оценивания моделей с ошибками во всех переменных (errors-in-variables models, EIV-модели) [2,3]. Зачастую под спецификацией таких моделей понимают зависимость объясняемой (выходной) переменной от одной или нескольких объясняющих переменных. В работах [4,5] автором предложена новая форма EIV-моделей – модель полносвязной линейной регрессии (МПЛР) вида:

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

$$x_j^* = a_j + b_j x_m^*, \quad j = \overline{1,m-1}, \tag{2}$$

где n – объем выборки; m – количество взаимосвязанных переменных; x_{ij} – i-е значение j-й наблюдаемой переменной; x_{ij}^* – неизвестное i-е значение j-й истинной переменной; a_j, b_j, $j = \overline{1,m-1}$ – неизвестные параметры; $\varepsilon_i^{(x_j)}$ – i-я ошибка j-й наблюдаемой переменной. МПЛР обобщает часто применяемую в клинической химии регрессию Деминга [6,7].

Таким образом, в полносвязной регрессии (1), (2) все взаимосвязанные переменные содержат ошибки, а все пары истинных переменных связаны между собой линейными функциональными зависимостями. Из этого следует, что модели (1), (2), в отличие от моделей множественной линейной регрессии, целесообразно применять при сильной корреляции всех переменных в выборке.

Возникает вопрос, как оценивать и интерпретировать МПЛР? Можно ли проводить интерпретацию таким же образом, как и в моделях множественной линейной регрессии? Данная статья в первую очередь посвящена поиску ответов на эти вопросы.

II. Оценка МПЛР методом максимального правдоподобия

В [4,5] для оценки МПЛР применяется взвешенный метод наименьших полных квадратов. В данной работе применим для оценки МПЛР метод максимального правдоподобия [8,9].

Пусть $\varepsilon_i^{(x_j)}$, $i = \overline{1,n}$, $j = \overline{1,m}$ – случайные величины, распределенные по нормальному закону с нулевыми математическими ожиданиями и постоянными дисперсиями, т.е. $\varepsilon_i^{(x_j)} \sim N\left(0, \sigma^2_{\varepsilon^{(x_j)}}\right)$, $i = \overline{1,n}$, $j = \overline{1,m}$. Тогда функция правдоподобия L будет иметь следующий вид:

$$L = \prod_{i=1}^{n}\left(\prod_{j=1}^{m-1}\frac{1}{\sigma_{\varepsilon^{(x_j)}}\sqrt{2\pi}}\exp\left(-\frac{\left(x_{ij} - a_j - b_j x_{im}^*\right)^2}{2\sigma^2_{\varepsilon^{(x_j)}}}\right)\times\right.$$

$$\left.\times\frac{1}{\sigma_{\varepsilon^{(x_m)}}\sqrt{2\pi}}\exp\left(-\frac{\left(x_{im} - x_{im}^*\right)^2}{2\sigma^2_{\varepsilon^{(x_m)}}}\right)\right).$$

Статья получена 14 июля 2023.

Базилевский Михаил Павлович, Иркутский государственный университет путей сообщения, Иркутск, Российская Федерация (e-mail: mik2178@yandex.ru).

Следовательно, можно записать логарифмическую функцию правдоподобия:

$$l = \ln L = -\frac{nm}{2}\ln 2\pi - \frac{n}{2}\ln\left(\sigma^2_{\varepsilon^{(x_1)}} \cdot ... \cdot \sigma^2_{\varepsilon^{(x_m)}}\right) -$$

$$-\sum_{i=1}^{n}\sum_{j=1}^{m-1}\frac{\left(x_{ij} - a_j - b_j x^*_{im}\right)^2}{2\sigma^2_{\varepsilon^{(x_j)}}} - \sum_{i=1}^{n}\frac{\left(x_{im} - x^*_{im}\right)^2}{2\sigma^2_{\varepsilon^{(x_m)}}} \to \max. \quad (3)$$

Предположим, что известны весовые коэффициенты $\lambda_j > 0$, $j = \overline{1, m-1}$, для взаимосвязанных переменных, причем, $\sigma^2_{\varepsilon^{(x_j)}} = \frac{1}{\lambda_j}\sigma^2_{\varepsilon^{(x_m)}}$, $j = \overline{1, m-1}$. Тогда целевая функция (3) будет иметь вид:

$$l = -\frac{nm}{2}\ln 2\pi - \frac{n}{2}\left(m\ln\sigma^2_{\varepsilon^{(x_m)}} + \ln\prod_{j=1}^{m-1}\frac{1}{\lambda_j}\right) -$$

$$-\frac{1}{2\sigma^2_{\varepsilon^{(x_m)}}}\sum_{i=1}^{n}\sum_{j=1}^{m-1}\lambda_j\left(x_{ij} - a_j - b_j x^*_{im}\right)^2 -$$

$$-\frac{1}{2\sigma^2_{\varepsilon^{(x_m)}}}\sum_{i=1}^{n}\left(x_{im} - x^*_{im}\right)^2 \to \max. \quad (4)$$

Вычислив частную производную функции (4) по переменной $\sigma^2_{\varepsilon^{(x_m)}}$ и приравнивая её к нулю, получим оценку параметра $\sigma^2_{\varepsilon^{(x_m)}}$:

$$\tilde{\sigma}^2_{\varepsilon^{(x_m)}} = \frac{1}{nm}\left(\sum_{i=1}^{n}\sum_{j=1}^{m-1}\lambda_j\left(x_{ij} - a_j - b_j x^*_{im}\right)^2 + \sum_{i=1}^{n}\left(x_{im} - x^*_{im}\right)^2\right). \quad (5)$$

Подставляя оценку (5) в выражение (4), отбрасывая постоянные члены и логарифмы и меняя знак целевой функции, получим:

$$\sum_{i=1}^{n}\sum_{j=1}^{m-1}\lambda_j\left(x_{ij} - a_j - b_j x^*_{im}\right)^2 + \sum_{i=1}^{n}\left(x_{im} - x^*_{im}\right)^2 \to \min. \quad (6)$$

Таким образом, для МПЛР оценки взвешенного метода наименьших полных квадратов, применяемого в [4,5], совпадают с оценками метода максимального правдоподобия.

III. Интерпретация МПЛР

Переменная x^*_m в правой части равенств (2) называется связующей. В [10] доказана теорема, согласно которой выбор связующей переменной в полносвязной регрессии (1), (2) не влияет на решение оптимизационной задачи (6).

Если в задаче (6) коэффициенты λ_j известны, то оценки полносвязной регрессии находятся по следующему алгоритму [4].

1. Численно решается нелинейная система

$$b_p\left(D_{x_m} + \sum_{j=1}^{m-1}\lambda_j^2 b_j^2 D_{x_j} + \right.$$

$$\left. + 2\sum_{j_1=1}^{m-2}\sum_{j_2=j_1+1}^{m-1}\lambda_{j_1}\lambda_{j_2}b_{j_1}b_{j_2}K_{x_{j_1}x_{j_2}} + 2\sum_{j=1}^{m-1}\lambda_j b_j K_{x_j x_m}\right) =$$

$$= \left(1 + \sum_{j=1}^{m-1}\lambda_j b_j^2\right)\left(\sum_{j=1}^{m-1}\lambda_j b_j K_{x_j x_p} + K_{x_m x_p}\right), \quad p = \overline{1, m-1}, \quad (7)$$

где символом D обозначены дисперсии, а K – ковариации переменных. Численный метод решения системы (7) подробно описан в [11]. В результате решения будут найдены оценки $\tilde{b}_1$, $\tilde{b}_2$, ..., $\tilde{b}_{m-1}$ параметров b_1, b_2, ..., b_{m-1}.

2. По формулам $\tilde{a}_j = \overline{x_j} - \tilde{b}_j \cdot \overline{x_m}$, $j = \overline{1, m-1}$, определяются оценки параметров a_1, a_2, ..., a_{m-1}.

3. Вычисляются оценки истинных значений переменной x_m по формулам

$$\tilde{x}^*_{im} = A_0 + A_1 x_{i1} + A_2 x_{i2} + ... + A_m x_{im}, \quad i = \overline{1, n}, \quad (8)$$

где $A_0 = \frac{-\sum_{j=1}^{m-1}\lambda_j \tilde{a}_j \tilde{b}_j}{1 + \sum_{j=1}^{m-1}\lambda_j \tilde{b}_j^2}$, $A_m = \frac{1}{1 + \sum_{j=1}^{m-1}\lambda_j \tilde{b}_j^2}$, $A_j = \frac{\lambda_j \tilde{b}_j}{1 + \sum_{j=1}^{m-1}\lambda_j \tilde{b}_j^2}$, $j = \overline{1, m-1}$.

Тогда оцененная полносвязная регрессия представляет собой множество взаимосвязей между всеми возможными парами переменных:

$$\tilde{x}^*_j = \tilde{a}_j + \tilde{b}_j \tilde{x}^*_m, \quad j = \overline{1, m-1}, \quad (9)$$

где значения переменной $\tilde{x}^*_m$ находятся по формуле (8).

Выражения (9) можно представить в виде:

$$\frac{\tilde{x}^*_1 - \tilde{a}_1}{\tilde{b}_1} = \frac{\tilde{x}^*_2 - \tilde{a}_2}{\tilde{b}_2} = ... = \frac{\tilde{x}^*_{m-1} - \tilde{a}_{m-1}}{\tilde{b}_{m-1}} = \tilde{x}^*_m. \quad (10)$$

По выражению (10) можно сделать вывод, что при оценивании полносвязной регрессии строится уравнение прямой в пространстве, в отличие от множественной регрессии, при оценивании которой строится гиперплоскость в пространстве. Этим же объясняется и то, что для построения полносвязной регрессии достаточно всего двух разных точек m-мерного пространства.

В [11] для измерения суммарного аппроксимационного качества полносвязной регрессии введен аддитивный коэффициент детерминации

$$R^2_{\text{add}} = \sum_{j=1}^{m} R^2_{x_j}, \quad (11)$$

где $R^2_{x_j}$ – коэффициент детерминации парной линейной регрессии $\tilde{x}^*_j$ от x_j. Чем ближе значение R^2_{add} к m, тем меньше разница между всеми наблюдаемыми и оцененными истинными переменными. Если $R^2_{\text{add}} \to m$, то это означает, что все пары наблюдаемых переменных x_1, x_2, ..., x_m связаны между собой практически линейными функциональными зависимостями.

В той же работе установлено, что значение аддитивного коэффициента детерминации (11) будет максимальным тогда, когда коэффициенты λ_j в задаче (6) будут назначены как отношения дисперсий переменных, т.е. $\lambda_1 = D_{x_m} / D_{x_1}$, $\lambda_2 = D_{x_m} / D_{x_2}$, ..., $\lambda_{m-1} = D_{x_m} / D_{x_{m-1}}$.

Очевидно, что при сильной корреляции всех переменных оценки $\tilde{b}_1$, $\tilde{b}_2$, ..., $\tilde{b}_{m-1}$ будут близки к МНК-оценкам соответствующих парных регрессий, а значит, знаки этих оценок будут согласованы со знаками соответствующих коэффициентов корреляции, т.е. будут

справедливы неравенства $A_j \cdot r_{x_j x_m} > 0$, $j = \overline{1, m-1}$. На этой основе в [11] разработан метод выпрямления искаженных коэффициентов (МВИК). Его суть состоит в том, чтобы сначала с использованием сильно коррелирующих переменных x_1, x_2, ..., x_m оценить полносвязную регрессию, а потом с помощью МНК оценить парную регрессию:

$$y_i = c_0 + c_1 \tilde{x}_m^* + \xi_i, \qquad i = \overline{1, n}, \tag{12}$$

где c_0, c_1 – неизвестные параметры; ξ_i – i-я ошибка аппроксимации.

Пусть оцененная модель (12) имеет вид $\tilde{y} = \tilde{c}_0 + \tilde{c}_1 \tilde{x}_m^*$. Тогда, используя (8), перепишем это уравнение в виде

$$\tilde{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + ... + \theta_m x_m, \tag{13}$$

где $\theta_0 = \tilde{c}_0 + \tilde{c}_1 A_0$, $\theta_j = \tilde{c}_1 A_j$, $j = \overline{1, m}$.

В [11] показано, что при сильной корреляции переменных x_1, x_2, ..., x_m с y знаки коэффициентов θ_j, $j = \overline{1, m}$, в уравнении (13) согласуются со знаками коэффициентов корреляции r_{yx_j}, т.е. справедливы неравенства $\theta_j \cdot r_{yx_j} > 0$, $j = \overline{1, m}$.

С помощью МВИК в работах [10, 11] успешно решены прикладные задачи и построены регрессионные уравнения (13). Но при этом их интерпретация никогда не проводилась.

Оценки $\tilde{\alpha}_j$, $j = \overline{1, m}$, множественной регрессии традиционно интерпретируются следующим образом: при увеличении переменной x_j на 1 единицу (при неизменных значениях остальных переменных) значение переменной y увеличится в среднем на $\tilde{\alpha}_j$ единиц. Использовать такую интерпретацию для коэффициентов θ_j, $j = \overline{1, m}$, уравнения (13) сомнительно, поскольку оно строилось в предположении, что все пары переменных x_1, x_2, ..., x_m сильно коррелируют между собой. Из-за полносвязности всех переменных изменение любой из них должно приводить к изменениям всех остальных. Используя взаимосвязи (9), можно дать такую интерпретацию: с увеличением переменной $\tilde{x}_m^*$ на 1 единицу переменная $\tilde{x}_1^*$ увеличится на $\tilde{b}_1$ единиц, переменная $\tilde{x}_2^*$ – на $\tilde{b}_2$ единиц, ..., переменная $\tilde{y}$ – на $\tilde{c}_1$ единиц. Однако такая интерпретация не может быть в полной степени удовлетворительной, поскольку основана на манипуляциях с латентными переменными $\tilde{x}_1^*$, $\tilde{x}_2^*$, ..., $\tilde{x}_m^*$, которые изначально и вовсе не наблюдались. Больший интерес естественно вызывает интерпретация влияния исходных, наблюдаемых переменных x_1, x_2, ..., x_m на истинные. Ответ на вопрос, как можно провести такую интерпретацию, дает следующая теорема.

Теорема. Одновременное увеличение в оцененной модели полносвязной линейной регрессии (2), (3) значений исходных переменных x_1, x_2, ..., x_{m-1} и x_m на $k \cdot \tilde{b}_1$, $k \cdot \tilde{b}_2$, ..., $k \cdot \tilde{b}_{m-1}$ и k единиц соответственно, где k – любое отличное от нуля число, приводит к увеличению оценок истинных значений переменных $\tilde{x}_1^*$, $\tilde{x}_2^*$, ..., $\tilde{x}_{m-1}^*$, $\tilde{x}_m^*$ на те же самые величины.

Доказательство. Пусть переменные x_1, x_2, ..., x_m принимают значения x_1^0, x_2^0, ..., x_m^0 соответственно. Тогда в этой точке оценка истинного значения переменной x_m по формуле (8):

$$\tilde{x}_m^{*0} = A_0 + A_1 x_1^0 + A_2 x_2^0 + ... + A_m x_m^0.$$

Теперь увеличим значение переменной x_1 с x_1^0 до $\left(x_1^0 + k\tilde{b}_1\right)$, x_2 – с x_2^0 до $\left(x_2^0 + k\tilde{b}_2\right)$, ..., x_{m-1} – с x_{m-1}^0 до $\left(x_{m-1}^0 + k\tilde{b}_{m-1}\right)$, x_m – с x_m^0 до $\left(x_m^0 + k\right)$. Тогда в этой точке оценка истинного значения переменной x_m по формуле (6):

$$\tilde{x}_m^{*1} = A_0 + A_1\left(x_1^0 + k\tilde{b}_1\right) + A_2\left(x_2^0 + k\tilde{b}_2\right) + ... + A_m\left(x_m^0 + k\right) =$$

$$= A_0 + A_1 x_1^0 + A_2 x_2^0 + ... + A_m x_m^0 + A_1 k\tilde{b}_1 + A_2 k\tilde{b}_2 + ... + A_m k =$$

$$= \tilde{x}_m^{*0} + k\left(\frac{\lambda_1 \tilde{b}_1}{1 + \sum_{j=1}^{m-1} \lambda_j \tilde{b}_j^2}\tilde{b}_1 + \frac{\lambda_2 \tilde{b}_2}{1 + \sum_{j=1}^{m-1} \lambda_j \tilde{b}_j^2}\tilde{b}_2 + ... + \right.$$

$$\left. + \frac{\lambda_{m-1} \tilde{b}_{m-1}}{1 + \sum_{j=1}^{m-1} \lambda_j \tilde{b}_j^2}\tilde{b}_{m-1} + \frac{1}{1 + \sum_{j=1}^{m-1} \lambda_j \tilde{b}_j^2}\right) = \tilde{x}_m^{*0} + k.$$

Это означает, что с увеличением значений переменных x_1, x_2, ..., x_{m-1} и x_m на $k\tilde{b}_1$, $k\tilde{b}_2$, ..., $k\tilde{b}_{m-1}$ и k единиц соответственно оцененная истинная переменная $\tilde{x}_m^*$ увеличится на k единиц. Как следует из (9), такое увеличение приведет к увеличению переменных $\tilde{x}_1^*$, $\tilde{x}_2^*$, ..., $\tilde{x}_{m-1}^*$ на $k\tilde{b}_1$, $k\tilde{b}_2$, ..., $k\tilde{b}_{m-1}$ единиц соответственно.

Теорема доказана.

Из теоремы следует, что если, например, значения исходных переменных x_1, x_2, ..., x_{m-1} и x_m увеличатся одновременно на $\tilde{b}_1$, $\tilde{b}_2$, ..., $\tilde{b}_{m-1}$ и 1 единицу соответственно, то значение переменной y увеличится в среднем на $\tilde{c}_1$ единиц.

IV. Пример

Для демонстрации предложенного способа интерпретации полносвязных регрессий были использованы ежегодные статистические данные за период с 2000 по 2021 годы по следующим показателям Иркутской области:

x_1 – продукция сельского хозяйства (млн руб.);

x_2 – оборот розничной торговли (млн руб.);

x_3 – оборот оптовой торговли (млн руб.).

Коэффициенты корреляции этих переменных составляют $r_{x_1 x_2} = 0,989$, $r_{x_1 x_3} = 0,9835$, $r_{x_2 x_3} = 0,9868$, что говорит о наличии весьма тесной линейной зависимости между ними. Это значит, что не выполняется условие применимости модели

множественной линейной регрессии, но выполняется условие применимости полносвязной регрессии.

Найденные по формулам $\lambda_1 = D_{x_3} / D_{x_1}$, $\lambda_2 = D_{x_3} / D_{x_2}$, соотношения дисперсий ошибок переменных составили $\lambda_1 = 208,032$, $\lambda_2 = 5,479$. С их помощью численно была оценена МПЛР:

$$\tilde{x}_1^* = 12701,131 + 0,0694\tilde{x}_3^*, \quad (14)$$

$$\tilde{x}_2^* = 54768,88 + 0,428\tilde{x}_3^*, \quad (15)$$

$$\tilde{x}_3^* = -103743 + 4,8x_1 + 0,78x_2 + 0,333x_3. \quad (16)$$

Таким образом, произошла связка всех пар переменных, чего невозможно добиться при использовании независимых моделей парной линейной регрессии.

Тогда оцененной модели (14) – (16) можно дать следующую интерпретацию: в регионе одновременно с увеличением оборота оптовой торговли x_3 на 1 млн руб. продукция сельского хозяйства x_1 увеличивается в среднем на 69400 руб., а оборот розничной торговли x_2 – на 428000 руб. Интерпретировать оценки уравнения (16) так, как это делается для уравнения множественной регрессии, нельзя. Даже несмотря на то, что знаки этих оценок согласуются со знаками соответствующих коэффициентов корреляции.

V. Заключение

Статья посвящена актуальному научному направлению – интерпретируемому машинному обучению [12]. В работе показано, что оценки моделей полносвязной линейной регрессии, полученные взвешенным методом наименьших квадратов, совпадают с оценками метода максимального правдоподобия. Установлено, что полносвязные регрессии нельзя интерпретировать по аналогии с их множественными аналогами. Доказано, что одновременное увеличение в оцененной модели полносвязной линейной регрессии значений наблюдаемых переменных на $k \cdot \tilde{b}_1$, $k \cdot \tilde{b}_2$, ..., $k \cdot \tilde{b}_{m-1}$ и k единиц соответственно приводит к увеличению оценок истинных значений переменных на те же самые величины. Тем самым можно интерпретировать любую полносвязную регрессию.

Библиография

[1] Carrasco J.M., Figueroa-Zuñiga J.I., Leiva V., Riquelme M., Aykroyd R.G. An errors-in-variables model based on the Birnbaum–Saunders distribution and its diagnostics with an application to earthquake data // Stochastic Environmental Research and Risk Assessment. 2020. Vol. 34. No. 2. P. 369-380.

[2] Nghiem L.H., Byrd M.C., Potgieter C.J. Estimation in linear errors-in-variables models with unknown error distribution // Biometrika. 2020. Vol. 107. No. 4. P. 841-856.

[3] Han J., Zhang S., Li Y., Zhang X. A general partial errors-in-variables model and a corresponding weighted total least-squares algorithm // Survey Review. 2020. Vol.52. No. 371. P. 126-133.

[4] Базилевский М.П. Методы построения регрессионных моделей с ошибками во всех переменных. Иркутск : ИрГУПС, 2019. 208 с.

[5] 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.

[6] Ciccione L., Dehaene S. Can humans perform mental regression on a graph? Accuracy and bias in the perception of scatterplots // Cognitive Psychology. 2021. Vol. 128. P. 101406.

[7] Vicente F.B., Lin D.C., Haymond S. Automation of chromatographic peak review and order to result data transfer in a clinical mass spectrometry laboratory // Clinica Chimica Acta. 2019. Vol. 498. P. 84-89.

[8] Karvonen T., Oates C.J. Maximum likelihood estimation in Gaussian process regression is ill-posed // Journal of Machine Learning Research. 2023. Vol. 24. No. 120. P. 1-47.

[9] Correia S., Guimarães P., Zylkin T. Verifying the existence of maximum likelihood estimates for generalized linear models // arXiv preprint arXiv:1903.01633. 2019.

[10] 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.

[11] Базилевский М.П. Метод выпрямления искаженных из-за мультиколлинеарности коэффициентов в регрессионных моделях // Информатика и её применения. 2021. Т. 15. № 2. С. 60-65.

[12] Molnar C. Interpretable machine learning. – Lulu.com, 2020.

Базилевский Михаил Павлович, к.т.н., доцент кафедры математики Иркутского государственного университета путей сообщения, Иркутск, Россия; (e-mail: mik2178@yandex.ru)

Interpretation of Parameter Estimates for Fully connected Linear Regression Models

M. P. Bazilevskiy

***Abstract*— This article is devoted to the study of interpretation questions of parameter estimates for fully connected linear regression models. In such models, all observed variables contain errors, and true variables are interconnected by linear functional dependencies. A special case of fully connected regression is the well-studied Deming regression. Previously, a weighted total least squares method was used to estimate fully connected regressions. In this article, it is established that the estimates of fully connected linear regression obtained by this method coincide with the estimates of the maximum likelihood method. It was found that it is impossible to interpret fully connected regressions by analogy with multiple regressions, since the former are construct on the assumption that all variables are strongly correlated with each other. A theorem is proved according to which a simultaneous increase in the values of the observed variables in the estimated model of a fully connected linear regression by certain values leads to an increase in the estimates of the true values of the variables by the same values. Using this fact, any model of fully connected linear regression can be interpreted, which is demonstrated by the example of modeling such macroeconomic indicators of the Irkutsk region as the turnover of wholesale and retail trade, as well as agricultural products.**

***Keywords*—regression analysis, fully connected linear regression model, errors in variables, interpretation, maximum likelihood method.**

REFERENCES

[1] Carrasco J.M., Figueroa-Zuñiga J.I., Leiva V., Riquelme M., Aykroyd R.G. An errors-in-variables model based on the Birnbaum–Saunders distribution and its diagnostics with an application to earthquake data // Stochastic Environmental Research and Risk Assessment. 2020. Vol. 34. No. 2. P. 369-380.

[2] Nghiem L.H., Byrd M.C., Potgieter C.J. Estimation in linear errors-in-variables models with unknown error distribution // Biometrika. 2020. Vol. 107. No. 4. P. 841-856.

[3] Han J., Zhang S., Li Y., Zhang X. A general partial errors-in-variables model and a corresponding weighted total least-squares algorithm // Survey Review. 2020. Vol.52. No. 371. P. 126-133.

[4] Bazilevskiy M.P. Metody postroeniya regressionnykh modeley s oshibkami vo vsekh peremennykh. Irkutsk : IrGUPS, 2019. 208 p.

[5] Bazilevskiy M.P. Mnogofaktornye modeli polnosvyaznoy lineynoy regressii bez ogranicheniy na sootnosheniya dispersiy oshibok peremennykh // Informatika i ee primeneniya. 2020. Vol. 14. No. 2. P. 92-97.

[6] Ciccione L., Dehaene S. Can humans perform mental regression on a graph? Accuracy and bias in the perception of scatterplots // Cognitive Psychology. 2021. Vol. 128. P. 101406.

[7] Vicente F.B., Lin D.C., Haymond S. Automation of chromatographic peak review and order to result data transfer in a clinical mass spectrometry laboratory // Clinica Chimica Acta. 2019. Vol. 498. P. 84-89.

[8] Karvonen T., Oates C.J. Maximum likelihood estimation in Gaussian process regression is ill-posed // Journal of Machine Learning Research. 2023. Vol. 24. No. 120. P. 1-47.

[9] Correia S., Guimarães P., Zylkin T. Verifying the existence of maximum likelihood estimates for generalized linear models // arXiv preprint arXiv:1903.01633. 2019.

[10] Bazilevskiy M.P. Issledovanie povedeniya otnositel'nykh vkladov peremennykh v obshchuyu determinatsiyu v otsenennom na osnove metoda vypryamleniya iskazhennykh koeffitsientov regressionnom uravnenii // Vestnik SibGUTI. 2022. No. 1 (57). P. 89-96.

[11] Bazilevskiy M.P. Metod vypryamleniya iskazhennykh iz-za mul'tikollinearnosti koeffitsientov v regressionnykh modelyakh // Informatika i ee primeneniya. 2021. Vol. 15. No. 2. P. 60-65.

[12] Molnar C. Interpretable machine learning. – Lulu.com, 2020.

Bazilevskiy Mikhail Pavlovich, Ph.D., Associate Professor of the Department of Mathematics, Irkutsk State Transport University, Irkutsk, Russia; (e-mail: mik2178@yandex.ru)