НАУЧНЫЙ ОТДЕЛ

ИНФОРМАТИКА

Известия Саратовского университета. Новая серия. Серия: Математика. Механика. Информатика. 2021. Т. 21, вып. 1. С. 88–99
Izvestiya of Saratov University. New Series. Series: Mathematics. Mechanics. Informatics, 2021, vol. 21, iss. 1, pp. 88–99

Научная статья
УДК 519.862.6
https://doi.org/10.18500/1816-9791-2021-21-1-88-99

Многокритериальный подход к построению моделей парно-множественной линейной регрессии

М. П. Базилевский

Иркутский государственный университет путей сообщения, Россия, 664074, г. Иркутск, ул. Чернышевского, д. 15

Базилевский Михаил Павлович, кандидат технических наук, доцент кафедры математики, mik2178@yandex.ru, https://orcid.org/0000-0002-3253-5697

Аннотация. Рассматривается модель парно-множественной линейной регрессии, представляющая собой синтез регрессии Деминга и модели множественной линейной регрессии. Показано, что с изменением типа минимизируемого расстояния модель парно-множественной регрессии плавно «трансформируется» из модели парной в модель множественной линейной регрессии. При этом модели парно-множественной регрессии сохраняют возможности интерпретации коэффициентов и прогнозирования значений объясняемой переменной. Предложен агрегированный критерий качества регрессионных моделей, основанный на четырех известных показателях: коэффициенте детерминации, коэффициенте Дарбина – Уотсона, согласованности поведения и средней относительной ошибки аппроксимации. С помощью этого критерия задача многокритериального построения модели парно-множественной линейной регрессии формализована в виде задачи нелинейного программирования. Разработан алгоритм ее приближенного решения. Результаты данной работы могут быть использованы для улучшения суммарных качественных характеристик моделей множественной линейной регрессии.

Ключевые слова: регрессия Деминга, модель парно-множественной линейной регрессии, многокритериальный подход, агрегированный критерий, нелинейное программирование

Для цитирования: *Базилевский М. П.* Многокритериальный подход к построению моделей парно-множественной линейной регрессии // Известия Саратовского университета. Новая серия. Серия: Математика. Механика. Информатика. 2021. Т. 21, вып. 1. С. 88–99. https://doi.org/10.18500/1816-9791-2021-21-1-88-99

Article
https://doi.org/10.18500/1816-9791-2021-21-1-88-99

Multi-criteria approach to pair-multiple linear regression models constructing

M. P. Bazilevskiy

Irkutsk State Transport University, 15 Chernyshevskogo St., Irkutsk 664074, Russia

Mikhail P. Bazilevskiy, mik2178@yandex.ru, https://orcid.org/0000-0002-3253-5697

Abstract. A pair-multiple linear regression model which is a synthesis of Deming regression and multiple linear regression model is considered. It is shown that with a change in the type of minimized distance, the pair-multiple regression model transforms smoothly from the pair model into the multiple linear regression model. In this case, pair-multiple regression models retain the ability to interpret the coefficients and predict the values of the explained variable. An aggregated quality criterion of regression models based on four well-known indicators: the coefficient of determination, Darbin – Watson, the consistency of behaviour and the average relative error of approximation is proposed. Using this criterion, the problem of multi-criteria construction of a pair-multiple linear regression model is formalized as a nonlinear programming problem. An algorithm for its approximate solution is developed. The results of this work can be used to improve the overall qualitative characteristics of multiple linear regression models.
Keywords: Deming regression, pair-multiple linear regression model, multi-criteria approach, aggregate criterion, nonlinear programming
For citation: Bazilevskiy M. P. Multi-criteria approach to pair-multiple linear regression models constructing. *Izvestiya of Saratov University. New Series. Series: Mathematics. Mechanics. Informatics*, 2021, vol. 21, iss. 1, pp. 88–99 (in Russian). https://doi.org/10.18500/1816-9791-2021-21-1-88-99

ВВЕДЕНИЕ

Регрессионный анализ [1–4] является признанным инструментом исследования влияния одной или нескольких объясняющих переменных на объясняемую переменную. В большинстве случаев регрессионные модели оцениваются с помощью метода наименьших квадратов (МНК) в предположении, что объясняющие переменные не содержат ошибок. Если же в этих переменных содержатся ошибки, то такие модели в зарубежной литературе принято называть «Errors-In-Variables models» (EIV-модели) или «measurement error models» [5–7]. Для оценивания EIV-моделей к настоящему времени разработан весьма мощный математический аппарат [8–10]. Однако практического применения EIV-модели почти не находят, потому что они не пригодны для точечного прогнозирования и возникают проблемы с их

интерпретацией. Исключением является регрессия Деминга [11, 12], которая нашла широкое применение в клинической химии [13, 14] и связанных областях.

В работе [15] автор синтезировал регрессию Деминга и модель множественной линейной регрессии. Полученный в результате синтез моделей сохраняет способности интерпретации оценок параметров и прогнозирования значений объясняемой переменной. В работе [16] для разработанного синтеза исследованы зависимости некоторых критериев адекватности от соотношения дисперсий ошибок переменных. При этом экспериментально установлено, что применение предложенного синтеза позволяет существенно повысить некоторые важные характеристики классической модели множественной линейной регрессии за счет незначительного снижения ее аппроксимационного качества. Целью данной работы является формализация многокритериального подхода к построению разработанного синтеза моделей в виде задач нелинейного программирования и разработка приближенных методов их решения.

Стоит отметить, что работа выполнена в рамках логико-алгебраического подхода к обработке статистических данных, при котором предполагается, что никаких априорных сведений об их вероятностной природе нет, поэтому не изучаются традиционные свойства оценок параметров – несмещенность, состоятельность и эффективность.

1. МОДЕЛЬ ПАРНО-МНОЖЕСТВЕННОЙ ЛИНЕЙНОЙ РЕГРЕССИИ

Рассмотрим предложенный в работах [15, 16] синтез регрессии Деминга и модели множественной линейной регрессии.

Пусть y_i, x_{i1}, $i = \overline{1,n}$ — наблюдаемые значения объясняемой и объясняющей переменной y и x_1, а y_i^*, x_{i1}^*, $i = \overline{1,n}$ — их неизвестные «истинные» значения. Предположим, что между переменными y^* и x_1^* имеет место линейная функциональная зависимость

$$y_i^* = \alpha + \beta x_{i1}^*, \qquad i = \overline{1,n}, \tag{1}$$

где α, β — неизвестные параметры.

Наблюдаемые и «истинные» значения переменных y и x_1 связаны соотношениями

$$x_{i1} = x_{i1}^* + \varepsilon_i^{(x_1)}, \qquad y_i = y_i^* + \varepsilon_i^{(y)}, \qquad i = \overline{1,n}, \tag{2}$$

где $\varepsilon_i^{(y)}$, $\varepsilon_i^{(x_1)}$, $i = \overline{1,n}$ — ошибки переменных y и x_1, которые могут быть вызваны, например, неточностями при измерении значений переменных. Никаких априорных сведений об этих ошибках нет.

Совокупность уравнений (1), (2) представляет собой простейшую EIV-модель — регрессию Деминга. Для ее оценивания будем использовать метод наименьших полных квадратов (МНПК), состоящий в минимизации функционала:

$$F(\alpha, \beta, x_{11}^*, ..., x_{n1}^*) = \sum_{i=1}^{n} (y_i - \alpha - \beta x_{i1}^*)^2 + \frac{1}{\lambda} \sum_{i=1}^{n} (x_{i1} - x_{i1}^*)^2 \to \min, \tag{3}$$

где $\lambda = \sigma^2_{\varepsilon^{(x_1)}} / \sigma^2_{\varepsilon^{(y)}}$ — положительное число, задающее тип расстояния от точек (x_{i1}, y_i), $i = \overline{1,n}$ до линии регрессии (1) [15]. Так, при $\lambda \to 0$ будет минимизироваться сумма квадратов вертикальных расстояний, при $\lambda \to 1$ — евклидовых расстояний, а при $\lambda \to \infty$ — горизонтальных расстояний.

Задача (3) имеет следующее аналитическое решение:

$$\tilde{\beta} = \frac{D_y - \lambda^{-1} D_{x_1} + \sqrt{(D_y - \lambda^{-1} D_{x_1})^2 + 4\lambda^{-1} K_{x_1 y}^2}}{2K_{x_1 y}}, \qquad \tilde{\alpha} = \overline{y} - \tilde{\beta}\overline{x_1},$$

$$\tilde{x}^*_{i1} = \frac{-\tilde{\alpha}\tilde{\beta} + \tilde{\beta}y_i + \lambda^{-1}x_{i1}}{\lambda^{-1} + \tilde{\beta}^2}, \qquad i = \overline{1,n}, \tag{4}$$

где $\tilde{\alpha}$, $\tilde{\beta}$, $\tilde{x}^*_{i1}$, $i = \overline{1,n}$ — оценки параметров; D_{x_1}, D_y — дисперсии переменных; K_{x_1y} — ковариация.

Тогда оценки «истинных» значений объясняемой переменной y находятся по формулам

$$\tilde{y}^*_i = \tilde{\alpha} + \tilde{\beta}\tilde{x}^*_{i1}, \qquad i = \overline{1,n}. \tag{5}$$

Рассмотрим, как меняются оценки (5) при варьировании параметра λ.

1. В соответствии с (4), $\lim\limits_{\lambda\to 0}\tilde{\beta} = \tilde{\beta}^{\text{МНК}}$, $\lim\limits_{\lambda\to 0}\tilde{\alpha} = \tilde{\alpha}^{\text{МНК}}$, где $\tilde{\beta}^{\text{МНК}} = \frac{K_{x_1y}}{D_{x_1}}$, $\tilde{\alpha}^{\text{МНК}} = \overline{y} - \frac{K_{x_1y}}{D_{x_1}}\overline{x_1}$ — МНК-оценки модели парной линейной регрессии y от x_1, а $\lim\limits_{\lambda\to 0}\tilde{x}^*_{i1} = x_{i1}$, $i = \overline{1,n}$. Тогда из (5) следует, что $\lim\limits_{\lambda\to 0}\tilde{y}^*_i = \tilde{\alpha}^{\text{МНК}} + \tilde{\beta}^{\text{МНК}}x_{i1}$, $i = \overline{1,n}$, т.е. оценки «истинных» значений объясняемой переменной y при $\lambda \to 0$ стремятся к ее расчетным по модели парной линейной регрессии y от x_1 значениям.

2. Аналогично пределы $\lim\limits_{\lambda\to\infty}\tilde{\beta} = \frac{D_y}{K_{x_1y}}$, $\lim\limits_{\lambda\to\infty}\tilde{\alpha} = \overline{y} - \frac{D_y}{K_{x_1y}}\overline{x_1}$, $\lim\limits_{\lambda\to\infty}\tilde{x}^*_{i1} = -\frac{\tilde{\alpha}}{\tilde{\beta}} + \frac{1}{\tilde{\beta}}y_i$, $i = \overline{1,n}$. Тогда $\lim\limits_{\lambda\to\infty}\tilde{y}^*_i = y_i$, $i = \overline{1,n}$, т.е. оценки «истинных» значений объясняемой переменной y при $\lambda \to \infty$ стремятся к ее наблюдаемым значениям.

Таким образом, варьирование значений параметра λ от 0 до ∞ приводит к изменению оценок $\tilde{y}^*_i$ от $(\tilde{\alpha}^{\text{МНК}} + \tilde{\beta}^{\text{МНК}}x_{i1})$ до y_i, $i = \overline{1,n}$.

Дополним переменную x_1 совокупностью объясняющих переменных x_2, x_3, ..., x_m, наблюдаемые значения которых x_{ij}, $i = \overline{1,n}$, $j = \overline{2,m}$. Для исследования влияния переменных x_1, x_2, x_3, ..., x_m на переменную $\tilde{y}^*$ введем модель множественной линейной регрессии:

$$\tilde{y}^*_i = d_0 + \sum_{j=1}^{m} d_j x_{ij} + \varepsilon_i, \qquad i = \overline{1,n}, \tag{6}$$

где d_0, d_1, ..., d_m — неизвестные параметры; ε_i, $i = \overline{1,n}$ — ошибки модели, присутствие которых в уравнениях (6) означает, что данная связь описывает процесс не точно, а с некоторой погрешностью.

Для оценивания модели (6) с помощью МНК необходимо решить оптимизационную задачу:

$$G(d_0, ..., d_m) = \sum_{i=1}^{n}\varepsilon_i^2 = \sum_{i=1}^{n}(\tilde{y}^*_i - d_0 - \sum_{j=1}^{m} d_j x_{ij})^2 \to \min.$$

Поскольку $\lim\limits_{\lambda\to 0}\tilde{y}^*_i = \tilde{\alpha}^{\text{МНК}} + \tilde{\beta}^{\text{МНК}}x_{i1}$, $i = \overline{1,n}$, то

$$\lim_{\lambda\to 0} G(d_0, ..., d_m) = \sum_{i=1}^{n}(\tilde{\alpha}^{\text{МНК}} + \tilde{\beta}^{\text{МНК}}x_{i1} - d_0 - \sum_{j=1}^{m} d_j x_{ij})^2.$$

Из этого следует, что задача $\lim\limits_{\lambda\to 0} G(d_0, ..., d_m) \to \min$ имеет решение: $\tilde{d}_0 = \tilde{\alpha}^{\text{МНК}}$, $\tilde{d}_1 = \tilde{\beta}^{\text{МНК}}$, $\tilde{d}_j = 0$, $j = \overline{2,m}$. При этом $\sum\limits_{i=1}^{n}\varepsilon_i^2 = 0$.

Так как $\lim\limits_{\lambda\to\infty}\tilde{y}_i^* = y_i$, $i=\overline{1,n}$, то

$$\lim_{\lambda\to\infty} G(d_0,...,d_m) = \sum_{i=1}^{n}\left(y_i - d_0 - \sum_{j=1}^{m} d_j x_{ij}\right)^2.$$

Отсюда следует, что задача $\lim\limits_{\lambda\to\infty} G(d_0,...,d_m) \to \min$ имеет решение при $\tilde{d}_j = \tilde{\alpha}_j^{\text{МНК}}$, $j=\overline{0,m}$, где $\tilde{\alpha}_j^{\text{МНК}}$, $j=\overline{0,m}$ — МНК-оценки модели множественной линейной регрессии y от x_1, x_2, ..., x_m.

Пусть оцененная с помощью МНК модель (6) имеет вид

$$\tilde{\tilde{y}}^* = \tilde{d}_0 + \sum_{j=1}^{m}\tilde{d}_j x_j, \tag{7}$$

где $\tilde{\tilde{y}}_i^*$, $i=\overline{1,n}$ — расчетные значения переменной $\tilde{y}^*$. Тогда с учетом вышесказанного $\lim\limits_{\lambda\to 0}\tilde{\tilde{y}}_i^* = \tilde{\alpha}^{\text{МНК}} + \tilde{\beta}^{\text{МНК}} x_{i1}$, а $\lim\limits_{\lambda\to\infty}\tilde{\tilde{y}}_i^* = \tilde{\alpha}_0^{\text{МНК}} + \sum\limits_{j=1}^{m}\tilde{\alpha}_j^{\text{МНК}} x_{ij}$, $i=\overline{1,n}$, т.е. при $\lambda\to 0$ уравнение (7) принимает вид однофакторной, а при $\lambda\to\infty$ — многофакторной зависимости.

Используем переменную $\tilde{\tilde{y}}^*$ в качестве инструмента для получения прогнозных значений объясняемой переменной y. Для этого введем модель парной линейной регрессии:

$$y_i = a + b\tilde{\tilde{y}}_i^* + u_i, \qquad i=\overline{1,n}, \tag{8}$$

где a, b — неизвестные параметры; u_i, $i=\overline{1,n}$ — ошибки модели.

МНК-оценки модели (8) являются результатом решения оптимизационной задачи:

$$Q(a,b) = \sum_{i=1}^{n} u_i^2 = \sum_{i=1}^{n}(y_i - a - b\tilde{\tilde{y}}_i^*)^2 \to \min.$$

Рассмотрим МНК-оценки модели (8) при $\lambda\to 0$ и при $\lambda\to\infty$.

1. Если $\lambda\to 0$, то $\lim\limits_{\lambda\to 0}\tilde{\tilde{y}}_i^* = \tilde{\alpha}^{\text{МНК}} + \tilde{\beta}^{\text{МНК}} x_{i1}$, $i=\overline{1,n}$. Тогда

$$\lim_{\lambda\to 0} Q(a,b) = \sum_{i=1}^{n}(y_i - a - b(\tilde{\alpha}^{\text{МНК}} + \tilde{\beta}^{\text{МНК}} x_{i1}))^2.$$

Из этого следует, что задача $\lim\limits_{\lambda\to 0} Q(a,b) \to \min$ имеет решение при $b(\tilde{\beta}^{\text{МНК}}) = \tilde{\beta}^{\text{МНК}}$, $a + b(\tilde{\alpha}^{\text{МНК}}) = \tilde{\alpha}^{\text{МНК}}$, откуда $\tilde{b}=1$, $\tilde{a}=0$.

2. Если $\lambda\to\infty$, то $\lim\limits_{\lambda\to\infty}\tilde{\tilde{y}}_i^* = \tilde{\alpha}_0^{\text{МНК}} + \sum\limits_{j=1}^{m}\tilde{\alpha}_j^{\text{МНК}} x_{ij}$, $i=\overline{1,n}$. Тогда

$$\lim_{\lambda\to\infty} Q(a,b) = \sum_{i=1}^{n}(y_i - a - b\left(\tilde{\alpha}_0^{\text{МНК}} + \sum_{j=1}^{m}\tilde{\alpha}_j^{\text{МНК}} x_{ij}\right)^2.$$

Следовательно, задача $\lim\limits_{\lambda\to\infty} Q(a,b) \to \min$ имеет решение при $b(\tilde{\alpha}_j^{\text{МНК}}) = \tilde{\alpha}_j^{\text{МНК}}$, $a + b(\tilde{\alpha}_0^{\text{МНК}}) = \tilde{\alpha}_0^{\text{МНК}}$, поэтому $\tilde{b}=1$, $\tilde{a}=0$.

Таким образом, с увеличением параметра λ от 0 до ∞ оцениваемая с помощью МНК модель (8) плавно «трансформируется» из парной $\tilde{y} = \tilde{\alpha}^{\text{МНК}} + \tilde{\beta}^{\text{МНК}} x_1$ во множественную $\tilde{y} = \tilde{\alpha}_0^{\text{МНК}} + \sum_{j=1}^{m} \tilde{\alpha}_j^{\text{МНК}} x_j$ регрессию. Тогда (8) можно справедливо назвать моделью парно-множественной линейной регрессии. Варьируя значения параметра λ, для парно-множественной регрессии можно получить бесчисленное множество различных и не изучавшихся ранее оценок.

С учетом (7) модель (8) представляет собой множественную регрессию вида $y_i = a + b(\tilde{d}_0 + \sum_{j=1}^{m} \tilde{d}_j x_{ij}) + u_i$, $i = \overline{1,n}$, т.е. с ограничениями на параметры. Это означает, что сумма квадратов ее ошибок $\sum_{i=1}^{n} u_i^2$ для любого значения λ не меньше, чем сумма квадратов ошибок оцененной с помощью МНК модели множественной линейной регрессии y от x_1, x_2, ..., x_m.

Заметим, что в рассмотренном случае при построении (8) первым шагом было оценивание параметров регрессии Деминга зависимости переменной y от x_1. Однако вместо переменной x_1 можно использовать любую другую объясняющую переменную из набора x_2, x_3, ..., x_m. Естественно, что при этом будут получены абсолютно другие результаты оценивания парно-множественной линейной регрессии (8).

Отметим также, что модели парно-множественной регрессии для любого значения λ сохраняют возможности интерпретации коэффициентов и прогнозирования значений переменной y.

2. МНОГОКРИТЕРИАЛЬНЫЕ ЗАДАЧИ

Перейдем к формализации многокритериального подхода к построению моделей парно-множественной линейной регрессии (8). Для этого можно использовать, например, следующие известные критерии адекватности [17]:

R^2 — коэффициент детерминации, характеризующий аппроксимационное качество модели и принимающий значения от 0 до 1;

DW — критерий Дарбина – Уотсона, характеризующий степень автокорреляции ошибок модели, а также уровень коинтеграции во временных рядах и принимающий значения от 0 до 4;

SP — критерий согласованности поведения (СП-критерий) [17], характеризующий согласованность поведения фактических и расчетных траекторий изменения переменной y и принимающий значения от $(1-n)$ до $(n-1)$;

E — средняя относительная ошибка аппроксимации, так же, как и R^2, характеризующая аппроксимационное качество модели и принимающая значения от 0 до ∞.

Идеальными значениями для R^2, DW, SP и E являются 1, 2, $(n-1)$ и 0 соответственно.

Понятно, что для модели (8) эти четыре критерия зависят от параметра λ, поэтому будем обозначать их $R^2(\lambda)$, $DW(\lambda)$, $SP(\lambda)$ и $E(\lambda)$. Введем их нормированные аналоги:

$$K_1(\lambda) = 1 - R^2(\lambda) = \frac{\sum_{i=1}^{n} u_i^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2}, \tag{9}$$

$$K_2(\lambda) = 0.5|2 - DW(\lambda)| = 0.5\left|2 - \frac{\sum\limits_{i=2}^{n}(u_i - u_{i-1})^2}{\sum\limits_{i=1}^{n} u_i^2}\right|, \tag{10}$$

$$K_3(\lambda) = 0.5\left(1 - \frac{SP(\lambda)}{n-1}\right) = 0.5\left(1 - \frac{\sum\limits_{i=1}^{n-1} \operatorname{sign}(y_{i+1} - y_i)\operatorname{sign}(y_{i+1} - y_i + u_i - u_{i+1})}{n-1}\right), \tag{11}$$

$$K_4(\lambda) = 0.01E(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{u_i}{y_i}\right|, \tag{12}$$

где u_i, $i = \overline{1,n}$ — ошибки модели (8) в зависимости от λ.

Очевидно, что область значений каждого из критериев $K_1(\lambda)$, $K_2(\lambda)$, $K_3(\lambda)$ лежит в интервале от 0 до 1. При этом чем ближе значение $K_1(\lambda)$, $K_2(\lambda)$ или $K_3(\lambda)$ к 0, тем выше качество модели. Для критерия $K_4(\lambda)$ наилучшим значением также является 0, однако область его возможных значений не ограничена сверху.

Тогда об общем качестве модели парно-множественной линейной регрессии (8) в зависимости от параметра λ можно судить по значению агрегированного критерия:

$$S(\lambda) = w_1K_1(\lambda) + w_2K_2(\lambda) + w_3K_3(\lambda) + w_4K_4(\lambda),$$

где w_1, w_2, w_3, w_4 — некоторые положительные весовые коэффициенты, которые в случае отсутствия приоритетов критериев можно задать равными. Идеальным значением этого критерия является 0.

Задача 1. Пусть дана объясняемая переменная y и совокупность объясняющих переменных x_1, x_2, ..., x_m. Требуется выбрать такое значение параметра λ модели парно-множественной регрессии (8), оцениваемой на основе регрессии Деминга y от x_h, $h \in \{1, 2, ..., m\}$, для которого

$$S(\lambda) = w_1K_1(\lambda) + w_2K_2(\lambda) + w_3K_3(\lambda) + w_4K_4(\lambda) \to \min. \tag{13}$$

Формализуем поставленную задачу в виде задачи математического программирования. Первый этап оценивания парно-множественной регрессии предполагает оценивание регрессии Деминга y от x_h, поэтому по аналогии с формулами (4), (5) введем ограничения:

$$\tilde{\beta} = \frac{D_y - \lambda^{-1}D_{x_h} + \sqrt{(D_y - \lambda^{-1}D_{x_h})^2 + 4\lambda^{-1}K_{x_hy}^2}}{2K_{x_hy}}, \qquad \tilde{\alpha} = \overline{y} - \tilde{\beta}\overline{x_h}, \tag{14}$$

$$\tilde{x}_{i,h}^* = \frac{-\tilde{\alpha}\tilde{\beta} + \tilde{\beta}y_i + \lambda^{-1}x_{i,h}}{\lambda^{-1} + \tilde{\beta}^2}, \qquad \tilde{y}_i^* = \tilde{\alpha} + \tilde{\beta}\tilde{x}_{i,h}^*, \qquad i = \overline{1,n}. \tag{15}$$

Так как параметр $\lambda > 0$, то введем ограничение

$$Small \leqslant \lambda \leqslant Large, \tag{16}$$

где $Small$, $Large$ — малое и большое положительные числа.

На втором шаге с помощью МНК по формуле $(X^TX)^{-1}X^T\tilde{y}^*$, где X — матрица наблюдаемых значений объясняющих переменных, находятся оценки модели (6). Тогда справедливы ограничения

$$\sum_{j=0}^{m} z_{1,j+1}\tilde{d}_j = \sum_{i=1}^{n} \tilde{y}_i^*, \qquad \sum_{j=0}^{m} z_{k,j+1}\tilde{d}_j = \sum_{i=1}^{n} x_{i,k-1}\tilde{y}_i^*, \qquad k = \overline{2, m+1}, \tag{17}$$

где z_{ij}, $i = \overline{1, m+1}$, $j = \overline{1, m+1}$ — элементы матрицы $Z = X^TX$.

На третьем шаге с помощью МНК находятся оценки модели (8), удовлетворяющие условиям

$$n\tilde{\alpha} + \tilde{\beta}\sum_{i=1}^{n} \tilde{\tilde{y}}_i^* = \sum_{i=1}^{n} y_i, \qquad \tilde{\alpha}\sum_{i=1}^{n} \tilde{\tilde{y}}_i^* + \tilde{\beta}\sum_{i=1}^{n} \tilde{\tilde{y}}_i^2 = \sum_{i=1}^{n} y_i\tilde{\tilde{y}}_i^*. \tag{18}$$

Ошибки парно-множественной регрессии (8) находятся по формулам

$$u_i = y_i - (\tilde{\alpha} + \tilde{\beta}(\tilde{d}_0 + \sum_{j=1}^{m} \tilde{d}_j x_{ij})), \qquad i = \overline{1, n}. \tag{19}$$

Тогда решение задачи нелинейного программирования (13) с ограничениями (7), (9)–(12), (14)–(19) позволяет определить оптимальное по рассмотренным четырем критериям адекватности значение параметра λ модели парно-множественной линейной регрессии (8). При этом если в функционале (13) $w_1 = 1$, $w_2 = w_3 = w_4 = 0$, то решением данной задачи будут оценки модели множественной линейной регрессии при $\lambda = Large$.

Задача 2. Пусть исследователь не знает, какую именно переменную из набора x_1, x_2, ..., x_m следует включить в спецификацию регрессии Деминга, чтобы обеспечить наилучшее качество модели парно-множественной регрессии по критерию (13). Тогда требуется выбрать из этого набора такую переменную для регрессии Деминга и такое значение параметра λ, чтобы минимизировать критерий (13).

Введем бинарные переменные δ_j, $j = \overline{1, m}$, по правилу

$$\delta_j = \begin{cases} 0, & \text{если } j\text{-я переменная не входит в регрессию Деминга,} \\ 1, & \text{в противном случае.} \end{cases}$$

Тогда справедливы следующие ограничения:

$$\delta_j \in \{0, 1\}, \qquad \sum_{j=1}^{m} \delta_j = 1. \tag{20}$$

При такой постановке задачи константы D_{x_h}, $K_{x_h,y}$, $\overline{x_h}$ станут переменными, для которых

$$D_{x_h} = \sum_{j=1}^{m} \delta_j D_{x_j}, \qquad K_{x_h,y} = \sum_{j=1}^{m} \delta_j K_{x_j,y}, \qquad \overline{x_h} = \sum_{j=1}^{m} \delta_j \overline{x_j}. \tag{21}$$

Тогда решение задачи (13) с ограничениями (7), (9)–(12), (14)–(21) дает ответы сразу на два вопроса: какую переменную из набора x_1, x_2, ..., x_m нужно использовать на первом шаге построения парно-множественной регрессии, и каково при этом значение параметра λ.

3. АЛГОРИТМ ПРИБЛИЖЕННОГО РЕШЕНИЯ ЗАДАЧИ

Для точного решения задачи нелинейного программирования (13), (7), (9)–(12), (14)–(21) с булевыми переменными можно воспользоваться любым современным оптимизационным программным обеспечением, например программой APMonitor. Вместе с тем существует возможность получения приближенного решения данной задачи. Для этого разработан алгоритм, представленный на рисунке.

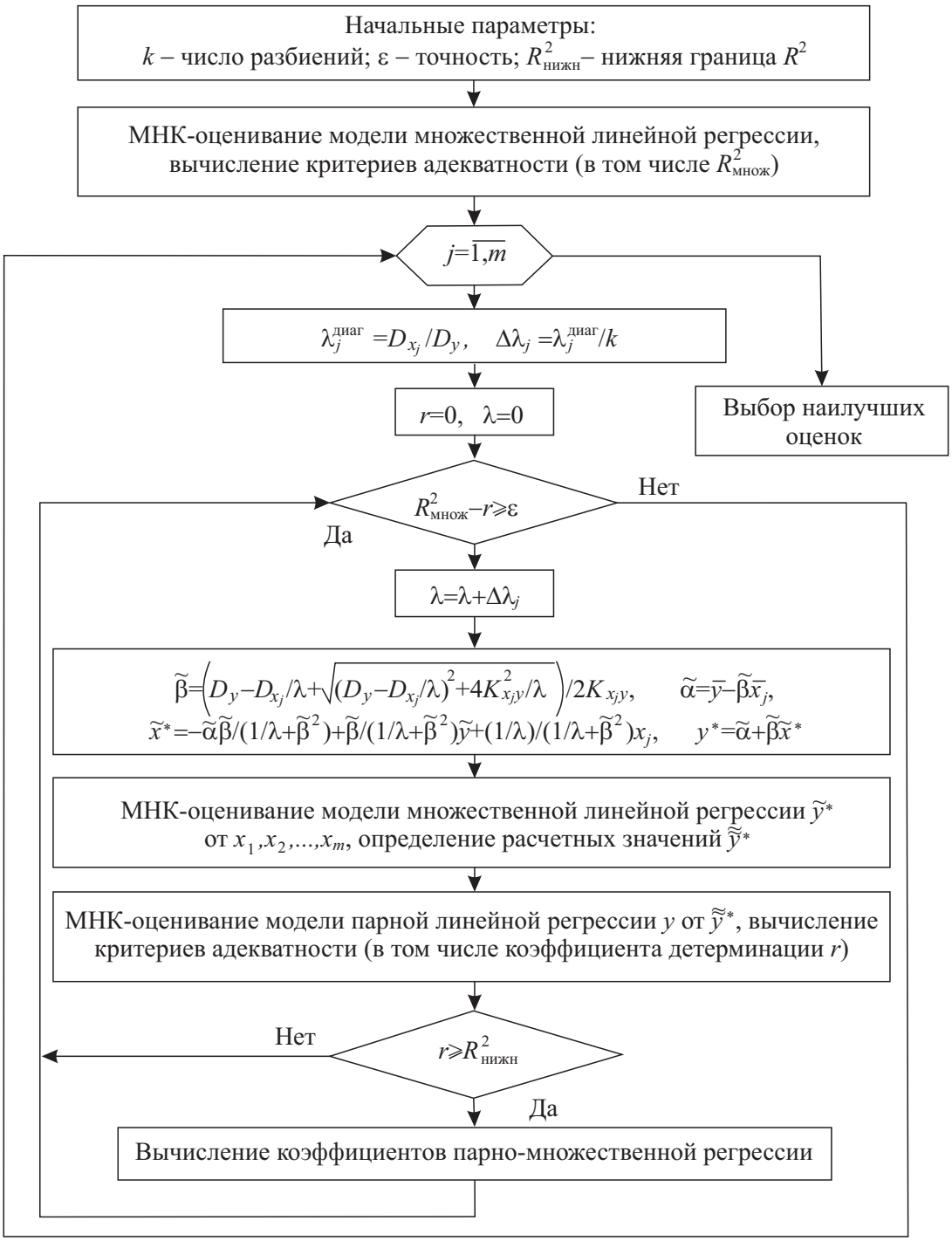

Рис. Алгоритм поиска приближенного решения задачи
Fig. Algorithm for finding an approximate solution to the problem

Суть алгоритма заключается в том, чтобы разбить интервал $\lambda \in [Small, Large]$ точками, а затем обычным перебором этих точек и переменных x_1, x_2, ..., x_m определить наилучшее решение задачи. Главной проблемой при разработке алгоритма являлось то, что было не понятно, как нужно разбивать интервал $\lambda \in [Small, Large]$. Если это делать равномерно, то, поскольку $Large$ — большое положительное число, в полученных точках оценки парно-множественной регрессии могут быть не представительными, т.е. существенно не отличаться от оценок множественной регрессии. Для решения этой проблемы в алгоритме была реализована следующая последовательность действий. Сначала для регрессии Деминга y от x_j, $j \in \{1, 2, ..., m\}$, значение параметра λ находится по формуле $\lambda_j = \frac{D_{x_j}}{D_y}$, т.е. как для диагональной регрессии [16]. В этой точке оценки парно-множественной регрессии существенно отличаются от оценок множественной регрессии. Затем отрезок $\lambda \in [Small, \lambda_j^{\text{диаг}}]$ равномерно разбивается k точками, находится шаг разбиения $\Delta\lambda_j = \frac{\lambda_j^{\text{диаг}}}{k}$. После чего на отрезке $\lambda \in [\lambda_j^{\text{диаг}}, Large]$ в цикле с шагом $\Delta\lambda_j$ продолжают генерироваться новые точки до тех пор, пока на очередной итерации разница между коэффициентами детерминации множественной $R^2_{\text{множ}}$ и парно-множественной регрессий r не станет меньше наперед заданной точности ε.

Стоит отметить, что в алгоритме на рисунке предусмотрено задание ограничения $R^2_{\text{нижн}}$ на коэффициент детерминации парно-множественной регрессии, т.е. при $r < R^2_{\text{нижн}}$ модель не будет принимать участия в процедуре выбора наилучших оценок.

ЗАКЛЮЧЕНИЕ

В данной работе рассмотрена модель парно-множественной линейной регрессии, представляющая собой синтез регрессии Деминга и модели множественной линейной регрессии. Задача многокритериального построения парно-множественной регрессии формализована в виде задачи нелинейного программирования. Разработан алгоритм для приближенного решения этой задачи.

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

Список литературы

1. *Montgomery D. C., Peck E. A., Vining G. G.* Introduction to Linear Regression Analysis. Wiley, 2012. 672 p.
2. *Kleinbaum D. G., Kupper L. L., Nizam A., Rosenberg E. S.* Applied Regression Analysis and Other Multivariable Methods. Cengage Learning, 2013. 1072 p.
3. *Harrell Jr., Frank E.* Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis. Springer Series in Statistics, 2015. 582 p.
4. *Kuhn M., Johnson K.* Applied Predictive Modeling. Springer, 2018. 600 p.
5. *Gillard J.* An overview of linear structural models in errors in variables regression // REVSTAT – Statistical Journal. 2010. Vol. 8, no. 1. P. 57–80.
6. *Xu K., Ma Y., Wang L.* Instrument assisted regression for errors in variables models with binary response // Scandinavian Journal of Statistics. 2015. Vol. 42, iss. 1. P. 104–117. https://doi.org/10.1111/sjos.12097

7. *Rudelson M., Zhou S.* Errors-in-variables models with dependent measurements // Electronic Journal of Statistics. 2017. Vol. 11, № 1. P. 1699–1797. https://doi.org/10.1214/17-EJS1234
8. *Gospodinov N., Komunjer I., Ng S.* Simulated minimum distance estimation of dynamic models with errors-in-variables // Journal of Econometrics. 2017. Vol. 200, iss. 2. P. 181–193. https://doi.org/10.1016/j.jeconom.2017.06.004
9. *Soderstrom T., Soverini U.* Errors-in-variables identification using maximum likelihood estimation in the frequency domain // Automatica. 2017. Vol. 79. P. 131–143. https://doi.org/10.1016/j.automatica.2017.01.016
10. *Bianco A. M., Spano P. M.* Robust estimation in partially linear errors-in-variables models // Computational Statistics & Data Analysis. 2017. Vol. 106. P. 46–64. https://doi.org/10.1016/j.csda.2016.09.002
11. *Deming W. E.* Statistical Adjustment of Data. Wiley, 1943. 273 p.
12. *Wu C., Yu J. Z.* Evaluation of linear regression techniques for atmospheric applications: The importance of appropriate weighting // Atmospheric Measurement Techniques. 2018. Vol. 11. P. 1233–1250. https://doi.org/10.5194/amt-11-1233-2018
13. *Henderson C. M., Shulman N. J., MacLean B., MacCoss M. J., Hoofnagle A. N.* Skyline performs as well as vendor software in the quantitative analysis of serum 25-hydroxy vitamin D and vitamin D binding globulin // Clinical Chemistry. 2018. Vol. 64, iss. 2. P. 408–410. https://doi.org/10.1373/clinchem.2017.282293
14. *Reverter-Branchat G., Bosch J., Vall J., Farre M., Papaseit E., Pichini S., Segura J.* Determination of recent growth hormone abuse using a single dried blood spot // Clinical Chemistry. 2016. Vol. 62, iss. 10. P. 1353–1360. https://doi.org/10.1373/clinchem.2016.257592
15. *Базилевский М. П.* Синтез модели множественной линейной регрессии и регрессии Деминга // Информационные технологии в моделировании и управлении: подходы, методы, решения : материалы II Всероссийской научной конференции с международным участием : в 2 ч. Тольятти, 2019. Ч. 1. С. 64–69.
16. *Базилевский М. П.* Синтез модели множественной линейной регрессии и регрессии Деминга: исследование зависимостей оценок параметров и критериев адекватности от соотношения дисперсий ошибок переменных // Информационные технологии и математическое моделирование в управлении сложными системами : электрон. науч. журнал. 2019. № 2. С. 18–25. URL: http://ismm-irgups.ru/toma/23-2019 (дата обращения: 19.06.2019).
17. *Носков С. И., Базилевский М. П.* Построение регрессионных моделей с использованием аппарата линейно-булевого программирования. Иркутск : ИрГУПС, 2018. 176 с.

References

1. Montgomery D. C., Peck E. A., Vining G. G. *Introduction to Linear Regression Analysis*. Wiley, 2012. 672 p.
2. Kleinbaum D. G., Kupper L. L., Nizam A., Rosenberg E. S. *Applied Regression Analysis and Other Multivariable Methods*. Cengage Learning, 2013. 1072 p.
3. Harrell Jr., Frank E. *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. Springer Series in Statistics, 2015. 582 p.
4. Kuhn M., Johnson K. *Applied Predictive Modeling*. Springer, 2018. 600 p.
5. Gillard J. An overview of linear structural models in errors in variables regression. *REVSTAT – Statistical Journal*, 2010, vol. 8, no. 1, pp. 57–80.
6. Xu K., Ma Y., Wang L. Instrument assisted regression for errors in variables models with binary response. *Scandinavian Journal of Statistics*, 2015, vol. 42, iss. 1, pp. 104–117. https://doi.org/10.1111/sjos.12097

7. Rudelson M., Zhou S. Errors-in-variables models with dependent measurements. *Electronic Journal of Statistics*, 2017, vol. 11, no. 1, pp. 1699–1797. https://doi.org/10.1214/17-EJS1234
8. Gospodinov N., Komunjer I., Ng S. Simulated minimum distance estimation of dynamic models with errors-in-variables. *Journal of Econometrics*, 2017, vol. 200, iss. 2, pp. 181–193. https://doi.org/10.1016/j.jeconom.2017.06.004
9. Soderstrom T., Soverini U. Errors-in-variables identification using maximum likelihood estimation in the frequency domain. *Automatica*, 2017, vol. 79, pp. 131–143. https://doi.org/10.1016/j.automatica.2017.01.016
10. Bianco A. M., Spano P. M. Robust estimation in partially linear errors-in-variables models. *Computational Statistics & Data Analysis*, 2017, vol. 106, pp. 46–64. https://doi.org/10.1016/j.csda.2016.09.002
11. Deming W. E. *Statistical Adjustment of Data*. Wiley, 1943. 273 p.
12. Wu C., Yu J. Z. Evaluation of linear regression techniques for atmospheric applications: The importance of appropriate weighting. *Atmospheric Measurement Techniques*, 2018, vol. 11, pp. 1233–1250. https://doi.org/10.5194/amt-11-1233-2018
13. Henderson C. M., Shulman N. J., MacLean B., MacCoss M. J., Hoofnagle A. N. Skyline performs as well as vendor software in the quantitative analysis of serum 25-hydroxy vitamin D and vitamin D binding globulin. *Clinical Chemistry*, 2018, vol. 64, iss. 2, pp. 408–410. https://doi.org/10.1373/clinchem.2017.282293
14. Reverter-Branchat G., Bosch J., Vall J., Farre M., Papaseit E., Pichini S., Segura J. Determination of recent growth hormone abuse using a single dried blood spot. *Clinical Chemistry*, 2016, vol. 62, iss. 10, pp. 1353–1360. https://doi.org/10.1373/clinchem.2016.257592
15. Bazilevskiy M. P. Synthesis of the multiple linear regression and deming regression model. *Informatsionnye tekhnologii v modelirovanii i upravlenii: podkhody, metody, resheniya: materialy II Vserossiiskoi nauchnoi konferentsii s mezhdunarodnym uchastiem* [Information Technologies in Modeling and Management: Approaches, Methods, Solutions: Materials of the II All-Russian Scientific Conference with International Participation: in 2 pt.]. Tolyatti, 2019, pt. 1, pp. 64–69 (in Russian).
16. Bazilevskiy M. P. Synthesis of multiple linear regression and Deming regression model's: investigation the dependences of parameter estimates and adequacy criteria on the ratio of variance error variables. *Informacionnye tehnologii i matematicheskoe modelirovanie v upravlenii slozhnymi sistemami: ehlektronnyj nauchnyj zhurnal* [Information technology and mathematical modeling in the management of complex systems: electronic scientific journal], 2019, no. 2, pp. 18–25 (in Russian). Available at: http://ismm-irgups.ru/toma/23-2019 (accessed 19 June 2019).
17. Noskov S. I., Bazilevskiy M. P. *Postroyenie regressionnykh modeley s ispol'zovaniem apparata lineino-bulevogo programmirovaniya* [Construction of Regression Models Using Linear Boolean Programming]. Irkutsk, IrGUPS, 2018. 176 p. (in Russian).

Поступила в редакцию / Received 11.11.2019
Принята к публикации / Accepted 07.10.2020
Опубликована / Published 01.03.2021