# Оценивание с помощью метода наименьших модулей регрессионных моделей с целочисленными функциями пол и потолок

М. П. Базилевский

***Аннотация*—Среди актуальных научных задач машинного обучения можно выделить поиск новых структурных спецификаций регрессионных моделей, успешно справляющихся с обработкой самых разнородных статистических данных. При проведении регрессионного анализа часто возникают ситуации, когда на выбор спецификации регрессионной модели влияет тип зависимой переменной. Например, если зависимая переменная принимает значения 0 и 1, то целесообразно строить логистическую регрессию, а если положительные значения – регрессию Пуассона. Если зависимая переменная принимает целочисленные значения, то применять линейную регрессию нецелесообразно, поскольку прогнозные значения, вероятно, не окажутся целыми. Например, если зависимая переменная – численность работников в организации, то возможное прогнозное значение «1200,517 человек» вряд ли можно считать корректным. Настоящая статья посвящена поиску ответа на вопрос, как поступать в такой ситуации. Предложены регрессионные модели с известными целочисленными функциями пол и потолок. Задача оценивания предложенных регрессий с помощью метода наименьших модулей сведена к задаче частично целочисленного линейного программирования. Показано, как предложенные модели соотносятся друг с другом. Рассмотрена интеграция целочисленных функций пол и потолок в линейную регрессию. Разработанные регрессии с целочисленными функциями были использованы для моделирования численности исследователей с учеными степенями в Иркутской области. При этом объясняющие переменные дополнительно преобразовывались с помощью элементарной функции натуральный логарифм. Все модели с целочисленными функциями были построены в пакете LPSolve за приемлемое время и по величине суммы модулей остатков оказались лучше, чем соответствующие линейные регрессии.**

***Ключевые слова*—регрессионная модель, целочисленная функция, функция пол, функция потолок, метод наименьших модулей, задача частично целочисленного линейного программирования.**

## I. Введение

Методы машинного обучения [1,2] находят широкое применение и продолжают активно развиваться в настоящее время. Например, в [3] приведен современный обзор использования искусственного интеллекта (ИИ) в военной сфере, в [4] описаны возможности применения ИИ в сельском хозяйстве, в [5] подчеркивается важность машинного обучения для прогнозирования заболеваний. Новым направлением в машинном обучении, возникшем относительно недавно, можно считать интерпретируемое машинное обучение [6]. Оно характеризуется тем, что построенная модель не должна представлять собой «черный ящик», функционирующий непонятным образом, а должна быть объяснима в понятных для человека терминах, чтобы ему был отчетливо ясен механизм работы модели. В монографии [6] среди интерпретируемых моделей машинного обучения с позиции простоты автор выделяет некоторые виды регрессионных моделей.

Регрессионный анализ, как весомая составляющая машинного обучения, также эволюционирует. Например, в [7] исследована связь между двухсторонними фиксированными эффектами и двухсторонней регрессией Мундлака, в [8] рассмотрены современные рекомендации, направленные на смягчение последствий мультиколлинеарности, в [9] изучен метод наименьших квадратов (МНК) и модулей (МНМ) для оценивания моделей с неточными наблюдениями. Для количественного оценивания степени нелинейности регрессионных моделей в [10] разработаны специальные критерии. В [11] заложены основы полносвязного регрессионного моделирования. В [12] задача отбора информативных регрессоров в оцениваемой с помощью МНК линейной регрессии сведена к задаче целочисленного программирования. В [13,14] рассмотрены методы построения неэлементарных линейных регрессий.

При проведении регрессионного анализа часто возникают ситуации, когда на выбор спецификации регрессионной модели влияет тип зависимой переменной. Можно выделить как минимум три таких ситуации.

*Ситуация № 1.* Зависимая переменная – бинарная. Например, переменная «Пол человека», которая принимает значение «0» для мужчин, и «1» для женщин. В таком случае целесообразно строить логистическую регрессию [15].

*Ситуация № 2.* Значения зависимой переменной положительны. Например, при прогнозировании на основе линейной регрессии «Времени простоя поезда» могут получаться отрицательные значения, чего не должно быть. Поэтому в такой ситуации целесообразно строить регрессию Пуассона [6,16].

Статья получена 4 июля 2024.

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

*Ситуация № 3*. Значения зависимой переменной целые. Например, при прогнозировании на основе линейной регрессии «Численности работников» могут получаться вещественные значения, что, по нашему мнению, не совсем корректно. Рекомендации о том, как поступать в такой ситуации, отсутствуют.

Поэтому была поставлена цель – разработать новую спецификацию регрессионных моделей, предназначенную для прогнозирования целочисленных зависимых переменных.

## II. Регрессионные модели с целочисленными функциями пол и потолок

Предположим, что имеется выборка числовых статистических данных объема $n$, содержащая наблюдения для зависимой (объясняемой) переменной $y$ и для $l$ независимых (объясняющих) переменных $x_1$, $x_2$, …, $x_l$.

Введем в рассмотрение регрессионную модель вида

$$y_i = \left\lfloor \alpha_0 + \sum_{j=1}^{l} \alpha_j x_{ij} \right\rfloor + \varepsilon_i, \qquad i = \overline{1,n}, \qquad (1)$$

где $y_i$, $i = \overline{1,n}$ – значения зависимой переменной; $x_{ij}$, $i = \overline{1,n}$, $j = \overline{1,l}$ – значения независимых переменных; $\alpha_0$, $\alpha_1$, …, $\alpha_l$ – неизвестные параметры; $\varepsilon_i$, $i = \overline{1,n}$ – ошибки аппроксимации.

В регрессии (1) запись $\lfloor x \rfloor$ означает наибольшее целое, меньшее или равное $x$. Например, $\lfloor 2{,}9 \rfloor = 2$. Тем самым осуществляется округление числа $x$ до ближайшего целого в меньшую сторону. Как отмечено в [17], обозначение $\lfloor x \rfloor$ было введено в обиход Кеннетом Э. Айверсоном в начале 60-х годов и было названо «пол» (от англ. «floor»). Им же введена операция $\lceil x \rceil$, названная «потолок» (от англ. «ceiling»). Например, $\lceil 2{,}9 \rceil = 3$, т.е. происходит округление числа $x$ до ближайшего целого в большую сторону. Описание функций пол и потолок и их приложения можно найти в [17].

Будем называть регрессию (1) моделью с целочисленной функцией пол.

К сожалению, модель (1) относится к нелинейной по оцениваемым параметрам спецификации, поэтому обычный МНК для неё не работает. В этой связи будем использовать МНМ, предполагающий формализацию процесса идентификации оценок неизвестных параметров в терминах аппарата математического программирования.

МНМ для идентификации оценок модели (1) требует решения задачи минимизации следующей функции:

$$J(\alpha_0, \alpha_1, ..., \alpha_l) = \sum_{i=1}^{n} \left| y_i - \left\lfloor \alpha_0 + \sum_{j=1}^{l} \alpha_j x_{ij} \right\rfloor \right| \to \min. \qquad (2)$$

Ещё в далёком 1955 году ученые А. Чарнс, В.В. Купер и Р.О. Фергюсон в своей работе «Optimal estimation of executive compensation by linear programming» показали, как свести задачу (2) для линейной регрессии к задаче линейного программирования. Далее воспользуемся эти приёмом.

Введем неотрицательные переменные $u_i$, $v_i$, $i = \overline{1,n}$, следующим образом:

$$u_i = \begin{cases} y_i - \left\lfloor \alpha_0 + \sum_{j=1}^{l} \alpha_j x_{ij} \right\rfloor, \text{ если } y_i - \left\lfloor \alpha_0 + \sum_{j=1}^{l} \alpha_j x_{ij} \right\rfloor > 0, \\ 0, \text{ в противном случае,} \end{cases}$$

$$v_i = \begin{cases} -y_i + \left\lfloor \alpha_0 + \sum_{j=1}^{l} \alpha_j x_{ij} \right\rfloor, \text{ если } y_i - \left\lfloor \alpha_0 + \sum_{j=1}^{l} \alpha_j x_{ij} \right\rfloor < 0, \\ 0, \text{ в противном случае.} \end{cases}$$

Тогда оптимизационная задача (2) сводится к следующей задаче математического программирования:

$$\sum_{i=1}^{n} (u_i + v_i) \to \min, \qquad (3)$$

$$y_i = \left\lfloor \alpha_0 + \sum_{j=1}^{l} \alpha_j x_{ij} \right\rfloor + u_i - v_i, \qquad i = \overline{1,n}, \qquad (4)$$

$$u_i \ge 0, \; v_i \ge 0, \quad i = \overline{1,n}. \qquad (5)$$

К сожалению, задача (3) – (5) не относится к линейной, поскольку нелинейными остаются ограничения (4). Для их линеаризации воспользуемся справедливым для функции пол правилом [17]:

$$\lfloor x \rfloor = \theta \quad \Leftrightarrow \quad \theta \le x < \theta + 1, \qquad (6)$$

где $\theta$ – целое число, т.е. $\theta \in \square$.

Используя правило (6), можно переписать ограничения (4) в виде

$$y_i = \theta_i + u_i - v_i, \qquad i = \overline{1,n}, \qquad (7)$$

$$\theta_i \le \alpha_0 + \sum_{j=1}^{l} \alpha_j x_{ij} < \theta_i + 1, \qquad i = \overline{1,n}, \qquad (8)$$

$$\theta_i \in \square, \qquad i = \overline{1,n}. \qquad (9)$$

Трансформировать строгое неравенство в ограничениях (8) в нестрогое можно следующим образом:

$$\theta_i \le \alpha_0 + \sum_{j=1}^{l} \alpha_j x_{ij} \le \theta_i + 1 - \Delta, \quad i = \overline{1,n}, \qquad (10)$$

где $\Delta$ – близкое к нулю положительное число, например, 0.0001.

Таким образом, для оценивания с помощью МНМ регрессии (1) с целочисленной функцией пол требуется решить задачу частично целочисленного линейного программирования (ЧЦЛП) [18] с целевой функцией (3) и ограничениями (5), (7), (9), (10). При этом полученные по модели прогнозные значения зависимой переменной будут целыми.

Аналогичным образом введем регрессионную модель с целочисленной функцией потолок:

$$y_i = \left\lceil \alpha_0 + \sum_{j=1}^{l} \alpha_j x_{ij} \right\rceil + \varepsilon_i, \qquad i = \overline{1,n}. \qquad (11)$$

Для функции потолок справедливо следующее правило [17]:

$$\lceil x \rceil = \theta \quad \Leftrightarrow \quad \theta - 1 < x \le \theta. \qquad (12)$$

Тогда, используя правило (12), легко показать, что МНМ-оценивание регрессии (11) сводится к решению

задачи ЧЦЛП с целевой функцией (3), ограничениями (5), (7), (9) и

$$\theta_i - 1 + \Delta \le \alpha_0 + \sum_{j=1}^{l} \alpha_j x_{ij} \le \theta_i, \quad i = \overline{1,n}. \qquad (13)$$

Возникает вопрос, а есть ли смысл оценивать на практике обе модели (1) и (11), а потом выбирать из них лучшую по величине суммы модулей остатков? Известно, что равенство $\lfloor x+1 \rfloor = \lceil x \rceil$ несправедливо только тогда, когда $x$ целое число. В результате оценивания регрессий (1) и (11) по реальным данным маловероятно, что хотя бы одна из сумм $\alpha_0 + \sum_{j=1}^{l} \alpha_j x_{ij}$ примет целое значение, отсюда следует, что только в очень редких случаях оцененные модели будут различаться по величине суммы модулей остатков. В большинстве же случаев это будут равносильные модели, у которых свободные члены $\alpha_0$ отличаются на единицу.

Стоит отметить, что функцию пол (потолок) можно интегрировать в линейную регрессию следующим образом:

$$y_i = \beta_0 + \sum_{j=1}^{l} \beta_j x_{ij} + \left\lfloor \alpha_0 + \sum_{j=1}^{l} \alpha_j x_{ij} \right\rfloor + \varepsilon_i, \quad i = \overline{1,n}, \quad (14)$$

где $\beta_0$, $\beta_1$, …, $\beta_l$ – неизвестные параметры.

Для МНМ-оценивания регрессии (14) требуется решить следующую задачу оптимизации:

$$J(\alpha_0,...,\alpha_l,\beta_0,...,\beta_l) = \sum_{i=1}^{n} \left| y_i - \beta_0 - \sum_{j=1}^{l} \beta_j x_{ij} - \left\lfloor \alpha_0 + \sum_{j=1}^{l} \alpha_j x_{ij} \right\rfloor \right| \to \min. \qquad (15)$$

Решение задачи (15) легко формализуется в виде задачи ЧЦЛП с целевой функцией (3), ограничениями (5), (9), (10) и

$$y_i = \beta_0 + \sum_{j=1}^{l} \beta_j x_{ij} + \theta_i + u_i - v_i, \quad i = \overline{1,n}.$$

Понятно, что при прогнозировании по оцененной регрессии (14) значения зависимой переменной могут быть вещественными.

Также следует отметить, что в моделях (1) и (11) с целочисленными функциями пол и потолок, объясняющие переменные могут быть преобразованы с помощью элементарных функций. На предложенный алгоритм МНМ-оценивания такие трансформации влияния не оказывают.

## III. Пример

Для демонстрации достоинств предложенных моделей с целочисленными функциями было решено использовать статистические данные о научной деятельности в Иркутской области, представленные в таблице 1, по следующим переменным:

$y$ – численность исследователей с учеными степенями (человек);

$x_1$ – численность персонала (исследователи, техники, вспомогательный персонал и пр.), занятого научными исследованиями и разработками (человек);

$x_2$ – внутренние затраты на научные исследования и разработки (млн руб.).

Таблица 1 – Статистические данные и остатки

| Год | $y$ | $x_1$ | $x_2$ | $E_1$ | $E_2$ | $E_4$ | $E_5$ |
|---|---|---|---|---|---|---|---|
| 2000 | 1289 | 5295 | 391,068 | -50,712 | -53 | 14,979 | 13 |
| 2001 | 1289 | 5408 | 473,868 | -75,255 | -78 | -32,059 | -32 |
| 2002 | 1274 | 5387 | 669,578 | -91,445 | -94 | -78,843 | -78 |
| 2003 | 1280 | 5184 | 808,59 | -49,193 | -51 | -44,382 | -44 |
| 2004 | 1294 | 4983 | 971,962 | 0 | -1 | 0 | 0 |
| 2005 | 1326 | 4829 | 1195,2 | 56,293 | 55 | 49,680 | 49 |
| 2006 | 1299 | 4557 | 1547,158 | 73,352 | 73 | 68,462 | 67 |
| 2007 | 1375 | 4910 | 2484,817 | 54,186 | 52 | 0 | 3 |
| 2008 | 1395 | 4897 | 2874,765 | 66,127 | 64 | 7,874 | 11 |
| 2009 | 1407 | 4919 | 3106,125 | 67,485 | 65 | 6 | 10 |
| 2010 | 1382 | 4912 | 3493,9 | 33,302 | 31 | -29,689 | -25 |
| 2011 | 1484 | 5075 | 3785,8 | 95,201 | 92 | 22,588 | 28 |
| 2012 | 1581 | 5384 | 4897,7 | 100,967 | 96 | 17,652 | 26 |
| 2013 | 1394 | 5047 | 4684,0 | -13,744 | -17 | -83,025 | -77 |
| 2014 | 1370 | 4859 | 4659,6 | 0 | -2 | -58,555 | -53 |
| 2015 | 1308 | 4671 | 4333,6 | -16,040 | -17 | -63,050 | -59 |
| 2016 | 1277 | 4409 | 4042,9 | 12,554 | 12 | -13,817 | -12 |
| 2017 | 1259 | 4292 | 4210,8 | 13,057 | 13 | -2,188 | -1 |
| 2018 | 1234 | 4157 | 4749,8 | 0 | 0 | 0,367 | 0 |
| 2019 | 1212 | 4002 | 6087,117 | -27,860 | -27 | 0 | 0 |
| 2020 | 1238 | 4074 | 6126,2 | -17,125 | -17 | 2,814 | 3 |
| 2021 | 1165 | 3932 | 5914,6 | -56,354 | -55 | -21,675 | -22 |

Источник данных – сайт Федеральной службы государственной статистики.

Сначала по эти данным с помощью МНМ была оценена линейная регрессия:

$$\tilde{y}_1 = 284,713 + 0,197232x_1 + 0,0272415x_2, \quad (16)$$

для которой сумма модулей остатков составила 970,2499. Конкретные значения этих остатков приведены в столбце $E_1$ таблицы 1. Как и следовало ожидать, три остатка оказались нулевыми, а все остальные – не целыми.

Линейная регрессия (16) интерпретируется корректно: чем больше численность персонала, занятого научными исследованиями, и чем больше внутренние затраты на научные исследования, тем больше можно ожидать исследователей с учеными степенями.

Затем оценивалась регрессия (1) с целочисленной функцией пол. Для решения задачи ЧЦЛП (3), (5), (7), (9), (10) был использован пакет LPSolve. Значение переменной $\Delta$ в этой задаче выбиралось равным 0,0001.

На начальной этапе в LPSolve была написана программа, включающая в себя ограничения на расчетные значения зависимой переменной типа t1>=-Inf, t2>=-Inf, t3>=-Inf и т.д. Они означают, что расчетные значения могут быть любых знаков. В таком случае программа работала очень долго. Но по смыслу зависимая переменная *y* (численность исследователей) не может принимать отрицательные значения, поэтому ограничения типа t1>=-Inf были исключены. В результате решение на обычном персональном компьютере было получено всего за 19,806 сек. Оцененная регрессия с функцией пол имеет вид

$$\tilde{y}_2 = \lfloor 266,752 + 0,201169x_1 + 0,027595x_2 \rfloor. \quad (17)$$

Сумм модулей ошибок регрессии (17) составила 965, что меньше, чем у модели (16). Иными словами, качество регрессии (17) выше, чем у регрессии (17). Остатки регрессии (17) приведены в столбце $E_2$ таблицы 1. Все эти остатки, как и расчетные значения зависимой переменной, целые.

После чего оценивалась регрессия (11) с целочисленной функцией потолок. Для этого решалась задача ЧЦЛП (3), (5), (7), (9), (13). В результате была получена следующая зависимость:

$$\tilde{y}_3 = \lceil 265,752 + 0,201169x_1 + 0,027595x_2 \rceil. \quad (18)$$

Как и ожидалось, остатки регрессий (17) и (18) полностью совпадают. А свободные члены под знаками целочисленных функций отличаются ровно на 1 единицу.

Графики наблюдаемых и прогнозных по модели (17) значений зависимой переменной *y* представлены на рис. 1.

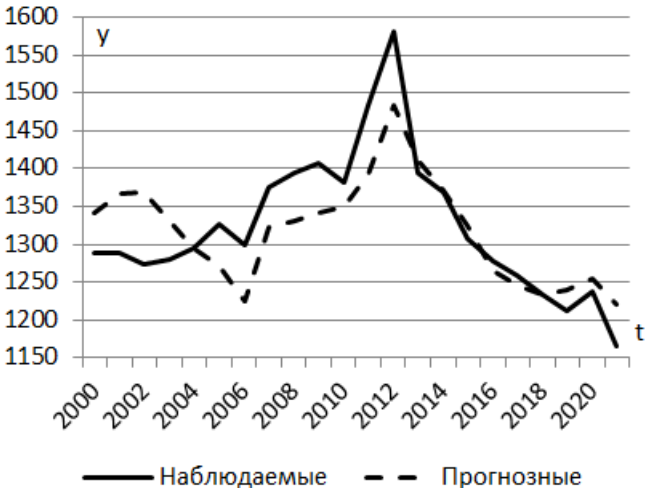


Рис. 1 – Наблюдаемые и прогнозные по модели (17) значения переменной *y*

По рис. 1 можно сделать вывод, что качество модели (17) в целом всё же неудовлетворительное. Поэтому было решено преобразовать объясняющие переменные в моделях с помощью элементарной функции натуральный логарифм, т.е. построить регрессии

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

$$y_i = \lfloor \alpha_0 + \alpha_1 \ln x_{i1} + \alpha_2 \ln x_{i2} \rfloor + \varepsilon_i, \quad i = \overline{1,n}.$$

Оцененные с помощью МНМ уравнения этих моделей имеют вид:

$$\tilde{y}_4 = -10182,5 + 1262,23 \ln x_1 + 106,141 \ln x_2, \quad (19)$$

$$\tilde{y}_5 = \lfloor -9878,168 + 1229,76194 \ln x_1 + 102,1867 \ln x_2 \rfloor. \quad (20)$$

Сумма модулей остатков регрессии (20) с целочисленной функцией пол снова оказалась ниже, чем у линейной регрессии (19). Для модели (19) эта сумма составляет 617,6991, а для (20) – 613. Остатки регрессии (19) приведены в столбце $E_4$ таблицы 1, а регрессии (20) – в столбце $E_5$ той же таблицы.

Графики наблюдаемых и прогнозных по модели (20) значений зависимой переменной *y* представлены на рис. 2.

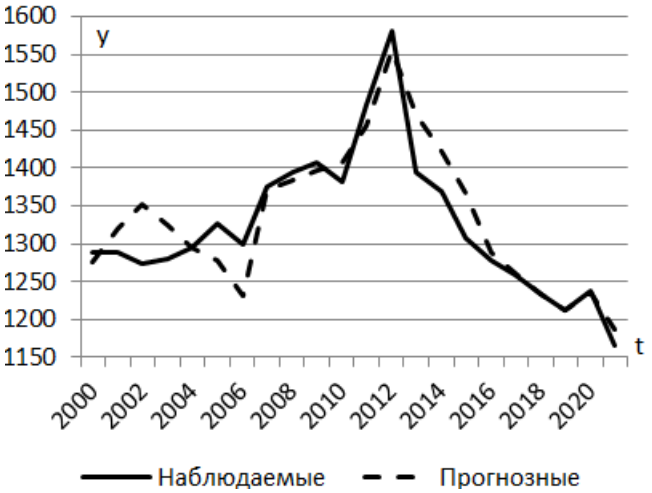


Рис. 2 – Наблюдаемые и прогнозные по модели (20) значения переменной *y*

Как видно по рис. 1 и рис. 2, качество модели (20) оказалось существенно выше, чем у регрессии (17). К

тому же сумма модулей остатков модели (20) примерно на 36,5% ниже, чем у (17). Поэтому регрессию (20) в целом можно признать адекватной и использовать, например, для прогнозирования численности исследователей с учеными степенями.

## IV. Заключение

В статье предложены регрессионные модели с целочисленными функциями пол и потолок. Разработан подход к их оцениванию с помощью метода наименьших модулей. Решена задача моделирования численности исследователей с учеными степенями в Иркутской области. Все построенные модели с целочисленными функциями оказались лучше по качеству, чем стандартные линейные регрессии. В дальнейшем планируется синтезировать предложенные модели с неэлементарными линейными регрессиями [13,14], содержащими бинарные операции min, max, а также модуль.

## Библиография


[1] Mahesh B. Machine learning algorithms-a review // International Journal of Science and Research. 2020. Vol. 9. No. 1. P. 381-386.

[2] Janiesch C., Zschech P., Heinrich K. Machine learning and deep learning // Electronic Markets. 2021. Vol. 31. No. 3. P. 685-695.

[3] Намиот Д.Е., Ильюшин Е.А., Чижов И.В. Военные применения машинного обучения // International Journal of Open Information Technologies. 2022. Т. 10. № 1. С. 69-76.

[4] Свецкий А.В. Применение искусственного интеллекта в сельском хозяйстве // Сельское хозяйство. 2022. № 3. С. 1-12.

[5] Гусев А.В., Новицкий Р.Э., Ившин А.А., Алексеев А.А. Машинное обучение на лабораторных данных для прогнозирования заболеваний // Фармакоэкономика. Современная фармакоэкономика и фармакоэпидемиология. 2021. Т. 14. № 4. С. 581-592.

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

[7] Wooldridge J.M. Two-way fixed effects, the two-way mundlak regression, and difference-in-differences estimators. 2021. Available at SSRN 3906345.

[8] Chan J.Y.L., Leow S.M.H., Bea K.T., Cheng W.K., Phoong S.W., Hong Z.W., Chen Y.L. Mitigating the multicollinearity problem and its machine learning approach: a review // Mathematics. 2022. Vol. 10. No. 8. P. 1283.

[9] Liu Z., Yang Y. Least absolute deviations estimation for uncertain regression with imprecise observations // Fuzzy Optimization and Decision Making. 2020. Vol. 19. P. 33-52.

[10] Базилевский М.П. Критерии нелинейности квазилинейных регрессионных моделей // Моделирование, оптимизация и информационные технологии. 2018. Т. 6. № 4 (23). С. 185-195.

[11] Базилевский М.П. Исследование двухфакторной модели полносвязной линейной регрессии // Моделирование, оптимизация и информационные технологии. 2019. Т. 7. № 2 (25). С. 80-96.

[12] Базилевский М.П. Отбор информативных регрессоров с учётом мультиколлинеарности между ними в регрессионных моделях как задача частично-булевого линейного программирования // Моделирование, оптимизация и информационные технологии. 2018. Т. 6. № 2 (21). С. 104-118.

[13] Базилевский М.П. Отбор информативных операций при построении линейно-неэлементарных регрессионных моделей // International Journal of Open Information Technologies. 2021. Т. 9. № 5. С. 30-35.

[14] Базилевский М.П. Метод построения неэлементарных линейных регрессий на основе аппарата математического программирования // Проблемы управления. 2022. № 4. С. 3-14.

[15] Schober P., Vetter T.R. Logistic regression in medical research // Anesthesia & Analgesia. 2021. Vol. 132. No. 2. P. 365-366.

[16] Amin M., Akram M.N., Kibria B.M.G. A new adjusted Liu estimator for the Poisson regression model // Concurrency and computation: Practice and experience. 2021. Vol. 33. No. 20. P. e6340.

[17] Грэхем Р., Кнут Д., Паташник О. Конкретная математика. Основание информатики: Пер. с англ. М. : Мир, 1998. 703 с.

[18] Wolsey L.A. Integer programming. John Wiley & Sons, 2020.


**Базилевский Михаил Павлович**, к.т.н., доцент кафедры математики Иркутского государственного университета путей сообщения, Иркутск, Россия; ORCID 0000-0002-3253-5697 (e-mail: mik2178@yandex.ru)

# Estimation using Least Absolute Deviations Method of Regression Models with Integer Floor and Ceiling Functions

M. P. Bazilevskiy

***Abstract*—Among the current scientific problems of machine learning, one can highlight the search for new structural specifications of regression models that successfully cope with the processing of the most heterogeneous statistical data. When conducting regression analysis, situations often arise when the type of dependent variable influences the choice of regression model specification. For example, if the dependent variable takes values 0 and 1, then it is advisable to construct a logistic regression, and if the values are positive, it is advisable to construct a Poisson regression. If the dependent variable is an integer, then linear regression is not appropriate because the predicted values will likely not be an integer. For example, if the dependent variable is the number of employees in the organization, then the possible forecast value of «1200.517 people» can hardly be considered correct. This article is devoted to finding an answer to the question of what to do in such a situation. Regression models with known integer functions floor and ceiling are proposed. The problem of estimating the proposed regressions using the least absolute deviations method is reduced to a mixed integer linear programming problem. It is shown how the proposed models relate to each other. The integration of integer functions floor and ceiling into linear regression is considered. The developed regressions with integer functions were used to model the number of researchers with academic degrees in the Irkutsk region. In this case, the explanatory variables were additionally transformed using the elementary function natural logarithm. All models with integer functions were construct in the LPSolve package in an acceptable time and, in terms of the sum of residual modules, turned out to be better than the corresponding linear regressions.**

***Keywords*—regression model, integer function, floor function, ceiling function, least absolute deviations method, mixed integer linear programming.**

## REFERENCES

[1] Mahesh B. Machine learning algorithms-a review // International Journal of Science and Research. 2020. Vol. 9. No. 1. P. 381-386.

[2] Janiesch C., Zschech P., Heinrich K. Machine learning and deep learning // Electronic Markets. 2021. Vol. 31. No. 3. P. 685-695.

[3] Namiot D.E., Il'yushin E.A., Chizhov I.V. Voennye primeneniya mashinnogo obucheniya // International Journal of Open Information Technologies. 2022. Vol. 10. No. 1. P. 69-76.

[4] Svetskiy A.V. Primenenie iskusstvennogo intellekta v sel'skom khozyaystve // Sel'skoe khozyaystvo. 2022. No. 3. P. 1-12.

[5] Gusev A.V., Novitskiy R.E., Ivshin A.A., Alekseev A.A. Mashinnoe obuchenie na laboratornykh dannykh dlya prognozirovaniya zabolevaniy // Farmakoekonomika. Sovremennaya farmakoekonomika i farmakoepidemiologiya. 2021. Vol. 14. No. 4. P. 581-592.

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

[7] Wooldridge J.M. Two-way fixed effects, the two-way mundlak regression, and difference-in-differences estimators. 2021. Available at SSRN 3906345.

[8] Chan J.Y.L., Leow S.M.H., Bea K.T., Cheng W.K., Phoong S.W., Hong Z.W., Chen Y.L. Mitigating the multicollinearity problem and its machine learning approach: a review // Mathematics. 2022. Vol. 10. No. 8. P. 1283.

[9] Liu Z., Yang Y. Least absolute deviations estimation for uncertain regression with imprecise observations // Fuzzy Optimization and Decision Making. 2020. Vol. 19. P. 33-52.

[10] Bazilevskiy M.P. Kriterii nelineynosti kvazilineynykh regressionnykh modeley // Modelirovanie, optimizatsiya i informatsionnye tekhnologii. 2018. Vol. 6. No. 4 (23). P. 185-195.

[11] Bazilevskiy M.P. Issledovanie dvukhfaktornoy modeli polnosvyaznoy lineynoy regressii // Modelirovanie, optimizatsiya i informatsionnye tekhnologii. 2019. Vol. 7. No. 2 (25). P. 80-96.

[12] Bazilevskiy M.P. Otbor informativnykh regressorov s uchetom mul'tikollinearnosti mezhdu nimi v regressionnykh modelyakh kak zadacha chastichno-bulevogo lineynogo programmirovaniya // Modelirovanie, optimizatsiya i informatsionnye tekhnologii. 2018. Vol. 6. No. 2 (21). P. 104-118.

[13] Bazilevskiy M.P. Otbor informativnykh operatsiy pri postroenii lineyno-neelementarnykh regressionnykh modeley // International Journal of Open Information Technologies. 2021. Vol. 9. No. 5. P. 30-35.

[14] Bazilevskiy M.P. Metod postroeniya neelementarnykh lineynykh regressiy na osnove apparata matematicheskogo programmirovaniya // Problemy upravleniya. 2022. No. 4. P. 3-14.

[15] Schober P., Vetter T.R. Logistic regression in medical research // Anesthesia & Analgesia. 2021. Vol. 132. No. 2. P. 365-366.

[16] Amin M., Akram M.N., Kibria B.M.G. A new adjusted Liu estimator for the Poisson regression model // Concurrency and computation: Practice and experience. 2021. Vol. 33. No. 20. P. e6340.

[17] Grekhem R., Knut D., Patashnik O. Konkretnaya matematika. Osnovanie informatiki: Per. s angl. Moscow : Mir, 1998. 703 p.

[18] Wolsey L.A. Integer programming. John Wiley & Sons, 2020.

**Bazilevskiy Mikhail Pavlovich**, Ph.D., Associate Professor of the Department of Mathematics, Irkutsk State Transport University, Irkutsk, Russia; ORCID 0000-0002-3253-5697 (e-mail: mik2178@yandex.ru)