Вычислительные технологии

УДК 519.862.6
DOI:10.25729/ESI.2023.31.3.013

Программное обеспечение для оценивания модульных линейных регрессий

Базилевский Михаил Павлович

Иркутский государственный университет путей сообщения,
Россия, Иркутск, *mik2178@yandex.ru*

Аннотация. Ранее были предложены модели модульной линейной регрессии, содержащие в качестве регрессоров модули отклонений значений объясняющих переменных от неизвестных коэффициентов. Известен алгоритм их точного оценивания с помощью метода наименьших модулей и алгоритм приближенного оценивания с помощью метода наименьших квадратов. Программных продуктов, реализующих эти алгоритмы, до сегодняшнего дня разработано не было. Данная статья посвящена описанию разработанного программного комплекса оценивания модульных линейных регрессий (ПК МОДУЛИР-1). В нём при оценивании модульной линейной регрессии с помощью метода наименьших модулей по заданным настройкам автоматически формируется задача частично-булевого линейного программирования для пакета LPSolve. В случае приближенного оценивания с помощью метода наименьших квадратов осуществляется полный перебор всех возможных вариантов моделей и выбирается лучшая по величине коэффициента детерминации модульная регрессия со всеми значимыми по t-критерию Стьюдента коэффициентами. С помощью ПК МОДУЛИР-1 решена задача моделирования грузооборота железнодорожного транспорта Забайкальского края. Коэффициент детерминации построенной с помощью метода наименьших квадратов модульной регрессии с пятью объясняющими переменными составил 0,94, что примерно в 4 раза выше, чем у традиционной линейной регрессии. При этом все коэффициенты модульной регрессии оказались значимы по t-критерию Стьюдента. Показано, как можно интерпретировать построенную модульную регрессию.

Ключевые слова: модульные регрессии, программное обеспечение, метод наименьших квадратов, метод наименьших модулей, коэффициент детерминации, t-критерий Стьюдента, грузооборот

Цитирование: Базилевский М.П. Программное обеспечение для оценивания модульных линейных регрессий / М.П. Базилевский // Информационные и математические технологии в науке и управлении. – 2023. – № 3(31). – С. 136-146. – DOI:10. 25729/ESI.2023.31.3.013.

Введение. В настоящее время методы машинного обучения [1, 2] находят широкое применение на практике и постоянно совершенствуются. Одним из наиболее известных методов машинного обучения является регрессионный анализ [3, 4], суть которого состоит в построении по имеющейся выборке статистических данных регрессионной модели зависимости выходной переменной от одной или нескольких входных переменных. Построенную математическую модель в дальнейшем можно использовать, например, для прогнозирования будущих значений выходной переменной. Простейшей моделью в регрессионном анализе считается модель множественной линейной регрессии следующего вида:

$$y_i = \alpha_0 + \sum_{j=1}^{l} \alpha_j x_{ij} + \varepsilon_i, \qquad i = \overline{1,n}, \tag{1}$$

где n – объем выборки; l – число входных (объясняющих) переменных; y_i, $i = \overline{1,n}$ – значения выходной (объясняемой) переменной; x_{ij}, $i = \overline{1,n}$, $j = \overline{1,l}$ – значения входных переменных; α_j, $j = \overline{0,l}$ – неизвестные параметры модели; ε_i, $i = \overline{1,n}$ – ошибки аппроксимации.

Для нахождения оценок неизвестных параметров модели (1) в регрессионном анализе разработан широкий арсенал методов. Наиболее эффективным из них принято считать метод наименьших квадратов (МНК), суть которого состоит в решении задачи минимизации суммы квадратов ошибок:

$$\sum_{i=1}^{n}\varepsilon_i^2=\sum_{i=1}^{n}\left(y_i-\alpha_0-\sum_{j=1}^{l}\alpha_j x_{ij}\right)^2\to\min.$$

Об аппроксимационном качестве оцененной с помощью МНК регрессии принято судить по величине коэффициента детерминации R^2, который связан с суммой квадратов остатков $\sum_{i=1}^{n}e_i^2$ (суммой квадратов разностей между фактическими и прогнозными значениями выходной переменной) следующим равенством:

$$R^2=1-\frac{\sum_{i=1}^{n}e_i^2}{\sum_{i=1}^{n}\left(y_i-\overline{y}\right)^2},$$

где $\overline{y}$ – среднее значение выходной переменной. Коэффициент детерминации R^2 принимает значения от 0 до 1. Чем ближе его значение к 1, тем выше аппроксимационное качество модели.

Проверка значимости коэффициентов оцененной с помощью МНК линейной регрессии (1) проверяется с помощью t-критерия Стьюдента. Для этого сначала определяется критическое значение критерия, зависящее от заданного уровня значимости α и числа степеней свободы $n-(l+1)$. Затем для каждого коэффициента регрессии определяются наблюдаемые значения критерия по формулам:

$$t_{\text{набл}}\left(\alpha_j\right)=\frac{\tilde{\alpha}_j}{se\left(\tilde{\alpha}_j\right)},\ \ j=\overline{0,l},$$

где $\tilde{\alpha}_j$, $j=\overline{0,l}$ – найденные с помощью МНК оценки линейной регрессии; $se\left(\tilde{\alpha}_j\right)$ – стандартные ошибки оценок коэффициентов модели. Потом наблюдаемые значения сравниваются с критическим значением критерия и делаются выводы о значимости коэффициентов регрессионной модели. Незначимые коэффициенты рекомендуется исключить, переоценив регрессию.

Реальные зависимости между исследуемыми факторами на практике зачастую носят нелинейный характер, поэтому в регрессионном анализе помимо линейных регрессий (1) разработаны и другие весьма эффективные спецификации. К ним относятся, например, логистические [5], полиномиальные [6], неэлементарные [7, 8], степенные [9], специфицированные на основе функций Леонтьева регрессии [9, 10].

В работе [11] впервые были предложены модели модульной линейной регрессии следующего вида:

$$y_i=\alpha_0+\sum_{j=1}^{l}\alpha_j\left|x_{ij}-\lambda_j\right|+\varepsilon_i,\qquad i=\overline{1,n},\tag{2}$$

где λ_j, $j=\overline{1,l}$ – неизвестные параметры.

Очевидно, что при $x_{ij}\ge 0$ и при $\lambda_1=\lambda_2=...=\lambda_l=0$ модульная регрессия (2) трансформируется в линейную регрессию (1). В той же работе [11] был предложен следующий алгоритм численного МНК-оценивания нелинейной модели (2).

1. Находятся области определения параметров $\lambda_j\in\left[x_{\min}^j,x_{\max}^j\right]$, $j=\overline{1,l}$, где $x_{\min}^j$, $x_{\max}^j$ – минимальное и максимальное значение j-й входной переменной.

2. На каждом отрезке $\left[x_{\min}^j,x_{\max}^j\right]$ равномерно выбирается p точек.

3. Полным перебором всех возможных $(p+2)^l$ комбинаций точек из отрезков $\left[x^j_{\min}, x^j_{\max}\right]$ с помощью МНК оцениваются уже линейные регрессии (2).

4. Выбирается та модель, у которой $\sum_{i=1}^{n} e_i^2 \to \min$.

5. В работе [12] задача точного оценивания модульной регрессии (2) с помощью метода наименьших модулей (МНМ) сведена к задаче частично-булевого линейного программирования.

Цель настоящей статьи состоит в описании разработанного программного комплекса, позволяющего оценивать модульные линейные регрессии с помощью МНК и МНМ.

1. Программный комплекс МОДУЛИР-1. Для идентификации неизвестных параметров модели (2) с помощью МНМ и МНК в среде программирования Delphi был разработан программный комплекс оценивания модульных линейных регрессий (ПК МОДУЛИР-1). Главное окно ПК МОДУЛИР-1 приведено на рис.1.

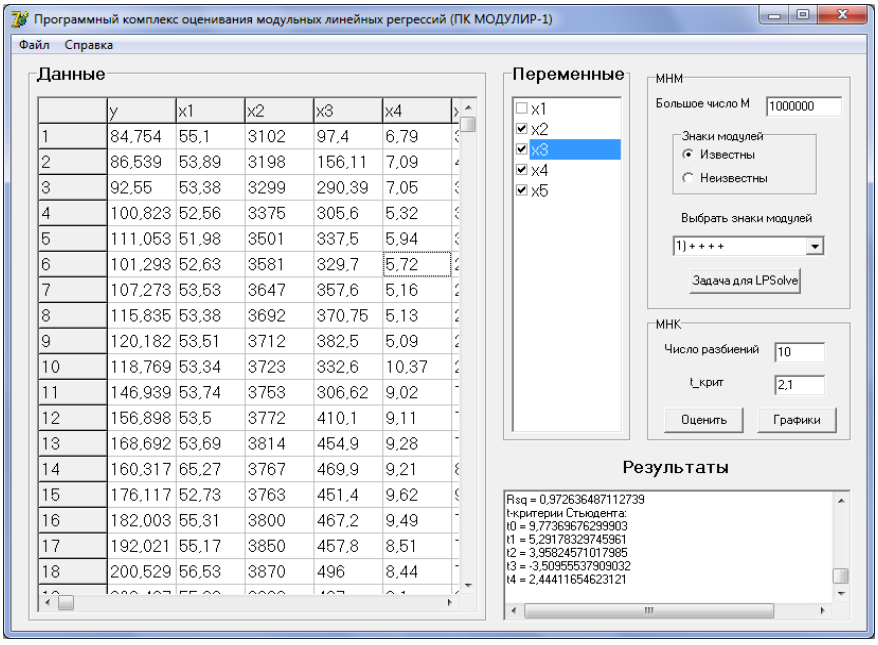

Рис. 1. Главное окно ПК МОДУЛИР-1

Для начала работы с ПК в его главном меню нужно воспользоваться командой Файл => Загрузка. Затем выбрать текстовый файл с расширением «.txt», хранящий исходную выборку данных. Этот файл должен содержать только прямоугольную таблицу с числами. Первый её столбец соответствует значениям объясняемой переменной y, второй – объясняющей переменной x_1 и т.д. Столбцы отделяются друг от друга символом «Tab». Целые и дробные части вещественных чисел отделяются друг от друга символом «,».

После выбора файла с данными они отобразятся на панели «Данные», а на панели «Переменные» появится список объясняющих переменных с переключателями. Используя эти переключатели, можно менять состав входящих в модель объясняющих переменных. Для дальнейшей работы хотя бы один из этих переключателей должен быть активирован.

Для оценивания модульной регрессии с помощью МНМ нужно на панели «МНМ» задать большое положительное число M, знаки коэффициентов, стоящих перед модулями, и нажать кнопку «Задача для LPSolve». После чего в поле «Результаты» будет сформирована задача частично-булевого линейного программирования для пакета LPSolve. Решение в нём сформированной задачи даёт точные МНМ-оценки выбранной модульной регрессии.

Для приближенного МНК-оценивания в ПК МОДУЛИР-1 модульной линейной регрессии нужно на панели «МНК» выбрать 2 параметра.

1. Число разбиений (razb). Этот параметр означает количество точек, разбивающих отрезки $\left[x_{\min}^{j}, x_{\max}^{j}\right]$. Чем больше этих точек задано, тем дольше будет происходить оценивание модульной регрессии, но тем ближе её МНК-оценки к оптимальным.

2. Критическое значение t-критерия Стьюдента (t_крит). Выбор этого параметра зависит от цели исследования. Если модульная регрессия нужна только для прогнозирования значений зависимой переменной, то t_крит нужно выбрать равным 0. Если модульную регрессию требуется интерпретировать, то t_крит выбирается в зависимости от уровня значимости α и числа степеней свободы $n-(l+1)$. В последнем случае нужно использовать таблицу критических значений t-критерия Стьюдента.

После выбора этих двух параметров на панели «МНК» нужно нажать кнопку «Оценить». Далее ПК МОДУЛИР-1 начинает работать по алгоритму, представленному на рис. 2. Рассмотрим подробнее его блоки.

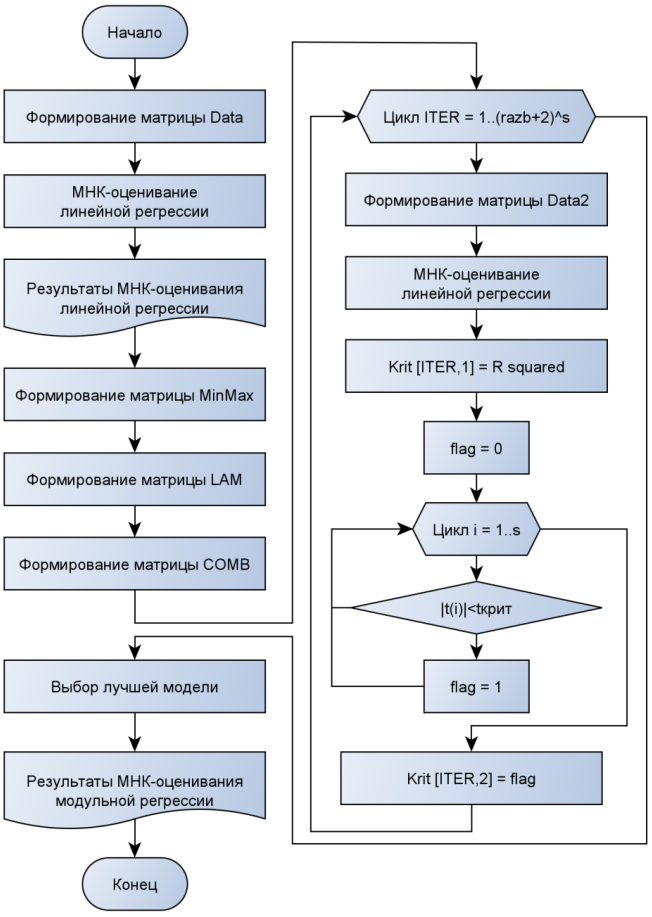

Рис. 2. Алгоритм работы ПК МОДУЛИР-1 при МНК-оценивании

Сначала формируется матрица Data размера n*(s+1), где n – объем выборки, s – число выбранных пользователем объясняющих переменных. Матрица Data содержит в первом столбце значения зависимой переменной y, а в оставшихся – значения s выбранных объясняющих переменных. Затем по матрице Data с помощью МНК оценивается линейная регрессия (1) и в поле «Результаты» главного окна ПК МОДУЛИР-1 выводится следующая информация: число наблюдений, список выбранных объясняющих переменных, МНК-оценки линейной регрессии, её коэффициент детерминации, наблюдаемые значения t-критерия Стьюдента для каждого коэффициента.

После чего формируется матрица MinMax размера 2*s, содержащая в первой строке минимальные значения выбранных объясняющих переменных, во второй – их максимальные значения. Потом с помощью MinMax формируется матрица лямбда-параметров LAM размера

(razb + 2)*s, содержащая по столбцам возможные значения этих параметров для каждой выбранной объясняющей переменной. Далее формируется матрица COMB размера ((razb + 2)^s)*s, содержащая по строкам в лексикографическом порядке все возможные комбинации размещений с повторениями из (razb +2) по s.

Затем создается критериальная матрица Krit размера ((razb+2)^s)*2, в первом столбце которой будут храниться значения коэффициентов детерминации моделей, а во втором – число 0, если все коэффициенты регрессии значимы по t-критерию, и 1, если имеется хотя бы один незначимый коэффициент. После чего запускается цикл с параметром ITER по строкам матрицы COMB, т.е. по всем возможным комбинациям значений параметров λ_j в модульной регрессии. Для каждой такой комбинации сначала формируется матрица Data2, содержащая в первом столбце значения зависимой переменной y, а в оставшихся столбцах – значения новых объясняющих переменных z_{ij}, преобразованных по правилу $z_{ij} = \left| x_{ij} - \lambda_j \right|$. Потом по матрице Data2 находятся МНК-оценки параметров α_j модульной регрессии (2), значение коэффициента детерминации сохраняется в первом столбце критериальной матрицы Krit и в результате проверки значимости коэффициентов по t-критерию Стьюдента во второй столбец матрицы Krit вносится число 0 или 1. По завершению работы цикла с параметром ITER критериальная матрица Krit будет полностью заполнена.

Далее с помощью критериальной матрицы Krit выбирается модель со всеми значимыми по t-критерию Стьюдента оценками и с наибольшим значением коэффициента детерминации R^2. Для найденной модульной регрессии в поле «Результаты» главного окна ПК МОДУЛИР-1 выводится следующая информация: общее количество оцененных регрессий ((razb+2)^s), количество моделей со всеми значимыми коэффициентами, МНК-оценки модульной регрессии, её коэффициент детерминации, наблюдаемые значения t-критерия Стьюдента для каждого коэффициента.

В ПК МОДУЛИР-1 также предусмотрена возможность вывода графиков наблюдаемых и прогнозных значений зависимой переменной для линейной и модульной линейной регрессий. Это можно сделать, нажав на панели «МНК» после оценивания модульной регрессии кнопку «Графики».

2. Решение прикладной задачи. Грузооборот – это основной показатель выполнения работы железной дороги. От него зависят такие важные экономические показатели, как доходы от перевозок и расходы по перевозкам.

Актуальной задачей, несомненно, можно считать моделирование грузооборота, являющегося одним из важнейших оперативных индикаторов экономической активности в реальном секторе, в частности экспортной активности [13]. Так, например, в [14] авторы моделировали прогнозирование грузооборота и объема перевезенных грузов.

Внешние и внутренние факторы оказывают существенное влияние на динамику и структуру грузооборота железнодорожного транспорта. Некоторые из них приведены на рис. 3.

Рассмотрим такой фактор, как наличие схем НТУ (непредусмотренные технические условия). Данная схема требуется для тех вагонов, в которых не предусмотрены техническими условиями способы размещения и крепления грузов для разовых, либо нерегулярных перевозок грузов, имеющих строго определённые размеры и массу. Схемы НТУ должны быть разработаны непосредственно к прибытию вагонов на станцию, но на практике наблюдается обратная ситуация – вагоны прибывают без схем, которые разрабатываются уже непосредственно после прибытия. На разработку и утверждение схемы затрачивается от трех дней (разработка схемы, создание реквизита крепления, работа бригады и т.д.). Соответственно возникают задержки при отправлении составов, которые затем суммируются и влияют на общее количество

отправленного груза в год. Чем больше вагонов, требующих разработки схем НТУ, тем меньше величина отправленного груза.

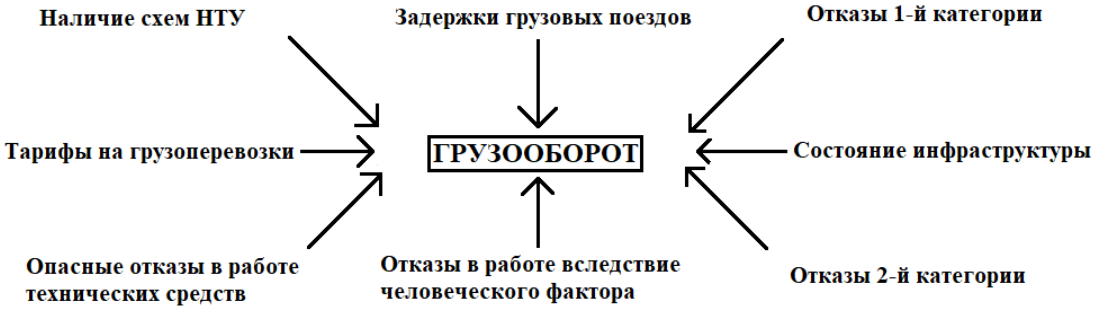

Рис. 3. Факторы, влияющие на грузооборот

В статье рассмотрим влияние на грузооборот задержек грузовых поездов, отказов 1-й и 2-й категорий, отказов в работе вследствие человеческого фактора и опасных отказов в работе технических средств. Именно эти факторы наиболее сильно влияют на грузооборот.

В зависимости от последствий отказов в работе технических средств вводится их следующая классификация по категориям:

– отказы 1-й категории – отказы, приводящие к задержке пассажирского, пригородного или грузового поезда на перегоне (станции) на 1 час и более, либо приводящие к транспортным происшествиям или событиям, связанным с нарушением правил безопасности движения и эксплуатации железнодорожного транспорта;

– отказы 2-й категории – отказы, вызвавшие задержку пассажирского, пригородного или грузового поезда на перегоне (станции) продолжительностью от 6 минут до 1 часа, либо ухудшение эксплуатационных показателей [15].

Опасные отказы перечислены в Правилах технической эксплуатации (ПТЭ). К некоторым из них относятся [16]: перевод стрелки под составом; ложная свободность участка; установление маршрута на занятый путь на станции и др.

Для построения модульной регрессии использовались ежемесячные статистические данные (табл. 1) за 2022 год, полученные от сотрудников Забайкальской железной дороги для следующих переменных: y – грузооборот, млрд тарифных тонно-км; x_1 – отказы 1-й категории, ед.; x_2 – отказы 2-й категории, ед.; x_3 – задержки грузовых поездов, ед; x_4 – отказы в работе технических средств вследствие человеческого фактора; x_5 – опасные отказы в работе технических средств.

Таблица 1. Статистические данные

переменная / месяц	y	x_1	x_2	x_3	x_4	x_5
январь	219,68	20	5	79	24	24
февраль	202,15	16	6	102	20	22
март	193,83	5	18	57	15	23
апрель	223,2	5	6	24	28	11
май	245,83	6	11	45	22	17
июнь	209,1	3	20	29	35	23
июль	253,43	6	16	34	19	22
август	188,24	11	30	67	33	41
сентябрь	175,88	9	23	53	20	28
октябрь	200,1	3	18	70	15	25
ноябрь	243,36	7	15	69	29	31
декабрь	280,95	10	26	115	27	36

Графики временных рядов для каждой переменной приведены на рис. 4 (а)-(е).

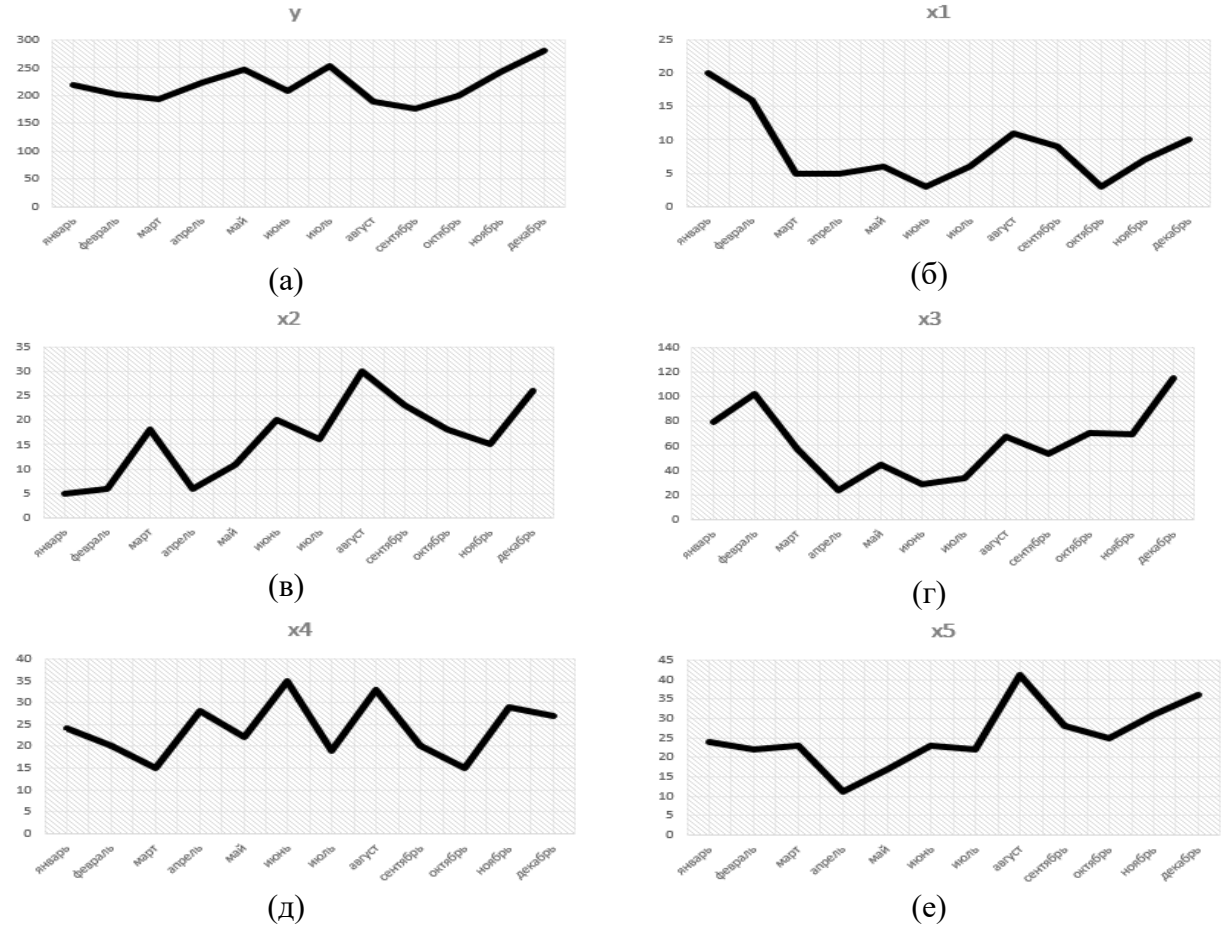

Рис. 4. Графики временных рядов

Анализируя графики, можно сделать вывод, что грузооборот слабо менялся в первом полугодии. Во втором полугодии с августа немного снизился и после октября начал возрастать. Связано это с всесторонним подходом к использованию возможностей Восточного полигона. Можно заметить, что отказов 1-й категории было меньше на протяжении года, чем отказов 2-й категории. Задержек грузовых поездов было больше в начале и в конце года, но больше, чем в середине года. Отказы работы технических средств вследствие человеческого фактора заметно снижались в марте (до 15 случаев) и в октябре (до 15 случаев), больше всего случаев отказа произошло в июне – всего 35. Самое большое количество случаев опасных отказов было зафиксировано в августе (41 случай), а самое низкое – в апреле (11 случаев). Тем не менее, итоги работы компании превзошли все ожидания, холдинг в прошлом году справился с масштабными вызовами и показал результаты, которые можно считать выдающимися [17].

По данным из таблицы 1 с помощью МНК получено уравнение линейной регрессии

$$\tilde{y} = 191,552 - \underset{(-0,882)}{3,334}\,x_1 - \underset{(-0,388)}{1,557}\,x_2 + \underset{(1,006)}{0,705}\,x_3 + \underset{(0,85)}{1,702}\,x_4 - \underset{(-0,027)}{0,124}\,x_5 . \qquad (3)$$

В скобках под коэффициентами этого уравнения указаны наблюдаемые значения t-критерия Стьюдента. Коэффициент детерминации R^2 регрессии (3) составил 0,242, что говорит о её низком аппроксимационном качестве. Причем, все коэффициенты модели (3) оказались незначимы по t-критерию Стьюдента.

Для оценивания модульной регрессии с помощью МНК был использован ПК МОДУЛИР-1. Число разбиений отрезков $\left[x^j_{\min}, x^j_{\max}\right]$, $j = \overline{1,5}$ было выбрано равным 10, а критическое значение t-критерия Стьюдента равным 2,45 [18]. В результате автоматически было оценено 248832 регрессий и выбрана лучшая из них по величине суммы квадратов остатков:

$$\tilde{y} = 163,450 + \underset{(7,356)}{11,645}\left|x_1 + 16,910\right| + \underset{(7,079)}{8,587}\left|x_2 + 25,454\right| +$$

$$+ \underset{(8,164)}{2,038}\left|x_3 + 73,636\right| - \underset{(-6,123)}{6,631}\left|x_4 + 22,273\right| - \underset{(-8,273)}{11,059}\left|x_5 + 38,273\right|. \qquad (4)$$

Для модульной регрессии (4) $R^2 = 0,94$. Таким образом, модель (4) обладает высоким качеством аппроксимации, превосходя линейную регрессию (3) по величине R^2 примерно в 4 раза. Все коэффициенты регрессии (4) значимы по t-критерию Стьюдента для уровня значимости $\alpha = 0,05$.

На рис. 5 приведены графики, автоматически построенные в ПК МОДУЛИР-1, демонстрирующие превосходство модульной регрессии (4) над линейной (3) по качеству.

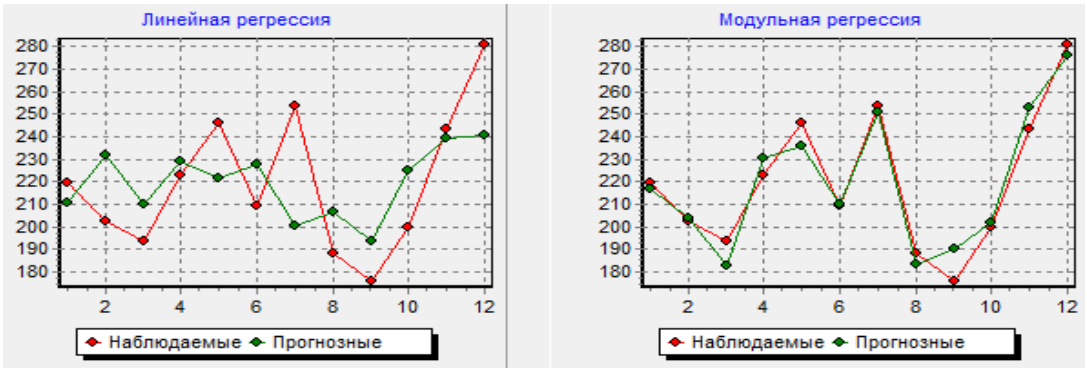

Рис. 5. Графики наблюдаемых и прогнозных значений зависимой переменной y для моделей (3) и (4)

Для интерпретации полученной модульной регрессии (4) необходимо представить её в кусочно-заданном виде, раскрыв знаки модулей. В нашем случае такая форма будет содержать 32 аналитических выражения, поэтому, чтобы не загромождать ею текст статьи, приведем вид модульной регрессии только при условиях $x_1 \geq 10,104$, $x_2 \geq 19,404$, $x_3 \geq 74,419$, $x_4 \geq 28,704$, $x_5 \geq 30,308$:

$$\tilde{y} = 1300,233 + 11,645x_1 + 8,587x_2 + 2,038x_3 - 6,631x_4 - 11,059x_5. \qquad (5)$$

Это выражение можно интерпретировать следующим образом: если отказов 1-й категории x_1 не меньше, чем 10,104, отказов 2-й категории x_2 не меньше, чем 19,404, задержек грузовых поездов x_3 не меньше, чем 74,419, отказов технических средств вследствие человеческого фактора x_4 не меньше, чем 28,704, а опасных отказов в работе технических средств x_5 не меньше, чем 30,308, то на размер грузооборота y независимые переменные x_1, x_2 и x_3 влияют со знаком «+», а x_4 и x_5 – со знаком «–». Причем с ростом x_1 на 1 единицу грузооборот y возрастает в среднем на 11,645 млрд тарифных тонно-км, с ростом x_2 на 1 единицу – на 8,587 млрд тарифных тонно-км, с ростом x_3 на 1 единицу – на 2,038 млрд тарифных тонно-км, с ростом x_4 на 1 единицу *y* убывает в среднем на 6,631 млрд тарифных тонно-км, а с ростом x_5 на 1 единицу *y* убывает в среднем на 11,059 млрд тарифных тонно-км.

Аналогично можно интерпретировать и другие аналитические выражения модульной регрессии (4).

Заключение. В данной статье приведено описание разработанного программного комплекса МОДУЛИР-1, позволяющего оценивать модульные линейные регрессии с помощью МНК и МНМ. ПК МОДУЛИР-1 относится к универсальным программным продуктам и может

быть использован для решения задач обработки и анализа данных из любых предметных областей. С помощью ПК МОДУЛИР-1 успешно решена задача моделирования грузооборота железнодорожного транспорта Забайкальского края. Построенная с помощью МНК модульная регрессия оказалась высокого качества, превзойдя линейную регрессию по величине коэффициента детерминации примерно в 4 раза. К тому же в модульной регрессии все коэффициенты установились значимыми по t-критерию Стьюдента. Полученный результат позволяет сделать предположение о том, что модульные регрессии являются весьма эффективным инструментом для прогнозирования поведения разного рода процессов и явлений. Безусловно, данное предположение подразумевает проведение в будущем специальных экспериментальных исследований. По их результатам можно будет судить об эффективности модульных регрессий и о том, какое место они занимают в области машинного обучения. К тому же вызывает научный интерес задача сопоставления вычислительных аспектов МНК и МНМ при оценивании модульных регрессий.

Список источников

1. Рашка С. Python и машинное обучение: пер. с англ / С. Рашка. – М.: Диалектика-Вильямс, 2022. – 848 с.
2. Флах П. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных: пер. с англ / П. Флах. – М.: ДМК Пресс, 2022. – 402 с.
3. Тищенко С.А. Методы машинного обучения в малом бизнесе: содержание и управление / С.А. Тищенко, М.А. Шахмурадян // Вестник Российского экономического университета им. ГВ Плеханова, 2019. – № 6 (108). – С. 83-95.
4. Montgomery D.C., Peck E.A., Vining, G.G. Introduction to linear regression analysis, John Wiley & Sons, 2021, 704 p.
5. Zabor E.C., Reddy C.A., Tendulkar R.D., Patil S. Logistic regression in clinical studies. International Journal of Radiation Oncology* Biology* Physics, 2021, v.112(2), pp.271-277.
6. Narayan V., Daniel A.K. Energy efficient protocol for lifetime prediction of wireless sensor network using multivariate polynomial regression model. Journal of Scientific & Industrial Research, 2022, v. 81(12), pp. 1297-1309.
7. Базилевский М.П. Метод построения неэлементарных линейных регрессий на основе аппарата математического программирования/ М.П. Базилевский // Проблемы управления, 2022. – № 4. – С. 3-14.
8. Базилевский М.П. Оценивание линейно-неэлементарных регрессионных моделей с помощью метода наименьших квадратов / М.П. Базилевский // Моделирование, оптимизация и информационные технологии, 2020. – Т. 8. – № 4 (31).
9. Клейнер Г.Б. Производственные функции: теория, методы, применение / Г.Б. Клейнер. – М.: Финансы и статистика, 1986. – 239 с.
10. Носков С.И. Программный комплекс построения некоторых типов кусочно-линейных регрессий / С.И. Носков, А.А. Хоняков // Информационные технологии и математическое моделирование в управлении сложными системами, 2019. – № 3 (4). – С. 47-55.
11. Базилевский М.П. Моделирование выбросов загрязняющих веществ в атмосферу Забайкальского края / М.П. Базилевский, А.Б. Ойдопова // Информационные технологии и математическое моделирование в управлении сложными системами, 2022. – № 2 (14). – С. 8-18.
12. Базилевский М.П. Оценивание модульных линейных регрессионных моделей с помощью метода наименьших модулей / М.П. Базилевский, А.Б. Ойдопова // Вестник Пермского национального исследовательского политехнического университета. Электротехника, информационные технологии, системы управления, 2023. – № 45. – С. 130-146.
13. Ворона А.А. Тенденции и перспективы грузооборота железнодорожного транспорта в России / А.А. Ворона // Таможенная политика России на Дальнем Востоке, – 2020. – № 3(92). – С. 93-99.
14. Краковский Ю.М. Разработка многофакторных моделей прогнозирования грузооборота и объема перевезенных грузов / Ю.М. Краковский, Т. Давааням // Современные технологии, системный анализ, моделирование, 2014. – №4 (44). – С. 110-113.
15. Об утверждении Положения об учете, расследовании и анализе отказов в работе технических средств на инфраструктуре ОАО «РЖД» с использованием автоматизированной системы КАСАНТ и положения об учете, расследовании и анализе технологических нарушений в перевозочном процессе на инфраструктуре ОАО «РЖД» с использованием автоматизированной системы КАСАТ: распоряжение ОАО «РЖД» N 2160/р от 1 октября 2018 г. – 69 с.

16. Классификатор опасных отказов технических средств хозяйства пути и сооружений утвержден и введен в действие Распоряжением ОАО «РЖД» от 28 апреля 2006 г. – 821 с.
17. Гусев С. Итоги работы ОАО «РЖД» в 2022 году превзошли ожидания / С. Гусев // «Гудок» Экономика. Электрон.журн., 2023. – URL: https://gudok.ru/news/util_avto (дата обращения: 01.04.2023).
18. Герасимов А.Н. Эконометрика (продвинутый уровень) / А.Н. Герасимов, Е.И. Громов, Ю.С. Скрипниченко – М.: Феникс, 2016. – 272 с. – EDN: XADINH.

Базилевский Михаил Павлович. *Доцент, кандидат технических наук, Иркутский государственный университет путей сообщения. AuthorID: 679277, SPIN: 4347-5028, ORCID: 0000-0002-3253-5697, mik2178@yandex.ru, Россия, Иркутск, Чернышевского д.15*

UDC 519.862.6
DOI:10.5729/ESI.2023.31.3.013

Software for estimating modular linear regressions

Mikhail P. Bazilevskiy

Irkutsk State Transport University,
Irkutsk, Russian Federation, *mik2178@yandex.ru*

Abstract. Previously, the author proposed modular linear regression models containing as regressors modules of deviations of the values of explanatory variables from unknown coefficients. An algorithm for their exact estimation using the least absolute deviation and an algorithm for approximate estimation using the least squares method is known. Software products implementing these algorithms have not been developed until today. This article is devoted to the description of the software package developed by the author for evaluating modular linear regressions (PC MODULIR-1). In it, when evaluating modular linear regression using the method of smallest modules according to the specified settings, a mixed-integer 0-1 linear programming problem for the LPSolve package is automatically generated. And in the case of approximate estimation using the least squares method, a complete search of all possible model variants is carried out and the best modular regression with all coefficients significant according to the Student's t-test is selected. The problem of modeling the freight turnover of railway transport in the Zabaykalsky krai was solved with the help of PC MODULAR-1. The coefficient of determination of the modular regression constructed using the least squares method with five explanatory variables was 0.94, which is about 4 times higher than that of traditional linear regression. At the same time, all the coefficients of modular regression turned out to be significant according to the Student's t-test. It is shown how the constructed modular regression can be interpreted.

Keywords: modular regressions, software, least absolute deviation, least squares method, coefficient of determination, Student's t-test, cargo turnover

References

1. Rashka S. Python i mashinnoye obucheniye [Python and machine learning]. – M.: Dialektika-Vil'yams [Dialectics-Williams], 2022, 848 p.
2. Flach P. Mashinnoye obucheniye. Nauka i iskusstvo postroyeniya algoritmov, kotoryye izvlekayut znaniya iz dannykh [Machine learning. The science and art of building algorithms that extract knowledge from data]. M., DMK Press, 2022, 402 p.
3. Tishchenko S.A., Shakhmuradyan M. A. Metody mashinnogo obycheniya v malom biznese: soderganie i upravlenie [Methods of machine-aided training in small business: content and management]. Vestnik Rossiyskogo ekonomicheskogo universiteta im. GV Plehanova [Vestnik of the Plekhanov Russian University of Economics], 2019, vol.6, pp. 83-95.
4. Montgomery D.C., Peck E.A., Vining, G. G. Introduction to linear regression analysis, John Wiley & Sons, 2021, p. 704.
5. Zabor E.C., Reddy C.A., Tendulkar R.D., Patil S. Logistic regression in clinical studies. International Journal of Radiation Oncology* Biology* Physics, 2021, v.112(2), pp.271-277.
6. Narayan V., Daniel A.K. Energy efficient protocol for lifetime prediction of wireless sensor network using multivariate polynomial regression model. Journal of Scientific & Industrial Research, 2022, v.81(12), pp.1297-1309.

7. Bazilevsky M.P. Metod postroeniya neelementarnykh lineinykh regressiy na osnove apparata matematicheskogo programmirovaniya [A method for constructing nonelementary linear regressions based on mathematical programming]. Problemy upravleniya [Control Sciences], 2022, no. 4, pp. 3-14.
8. Bazilevsky M.P. Otsenivanie lineyno-neelementarnykh regressionnykh modelei s pomoschyu metoda naimenshikh kvadratov [Estimation linear non-elementary regression models using ordinary least squares]. Modelirovaniye, optimizatsiya i informatsionnyye tekhnologii [The scientific journal modeling, optimization and information technology], 2020, vol. 8, no. 4 (31).
9. Kleiner G.B. Proizvodstvennie funktsii: teoriya, metody, primenenie [Production functions: theory, methods, application]. M.: Finansy i statistika [Moscow, Finance and statistics], 1986, 239 p.
10. Noskov S.I., Khonyakov A.A. Programnyi kompleks postroeniya nekotorykh tipov kusocho-lineynikh regressiy [Software complex for building some types pieces of linear regressions]. Informatsionnyye tekhnologii i matematicheskoye modelirovaniye v upravlenii slozhnymi sistemami [Information technology and mathematical modeling in the management of complex systems], 2019, no. 3(4), pp. 47-55.
11. Bazilevsky M.P., Oydopova A.B. Modelirovanie vibrosov zagryaznyuschikh veschestv v atmosphere Zabaykalskogo kraya [Modeling of emissions of pollutants into the atmosphere of the Zabaikalsky kray]. Informatsionnyye tekhnologii i matematicheskoye modelirovaniye v upravlenii slozhnymi sistemami [Information technology and mathematical modeling in the management of complex systems], 2022, no. 2 (14), pp. 8-18.
12. Bazilevsky M.P., Oydopova A.B. Otsenivanie modulnykh lineinykh regressionykh modelei s pomoschyu metoda naimenshikh modulei [Estimation of modular linear regression models using the least absolute deviations]. Vestnik Permskogo natsional'nogo issledovatel'skogo politekhnichesko-go universiteta. Elektrotekhnika, informatsionnyye tekhnologii, sistemy upravleniya [Perm National Research Polytechnic University Bulletin. electrotechnics, informational technologies, control systems], 2023, no. 45, pp. 130-146.
13. Vorona A.A. Tendetsii i perspektivi gruzooborota geleznodorognogo transporta v Rossii [Trends and prospects of freight turnover of railway transport in Russia]. Tamozhennaya politika Rossii na Dal'nem Vostoke [The scientific and practical journal Customs Policy of Russia in the Far East], 2020, no. 3(92), pp. 93-99.
14. Krakovskiy Y.M., Tamir D. Razrabotka mnogofactornykh modelei prognozirovaniya gruzooborota i obiema perevezennykh gruzov [Development of the multivariate model for goods turnover forecasting and transported cargo volume]. Sovremennyye tekhnologii, sistemnyy analiz, modelirovaniye [Modern technologies system analysis modeling], 2014, no. 4 (44), pp. 110-113.
15. Ob utverzhdenii Polozheniya ob uchete, rassledovanii i analize otkazov v rabote tekhnicheskikh sredstv na infrastrukture OAO "RZHD" s ispol'zovaniyem avtomatizirovannoy sistemy KASANT i polozheniya ob uchete, rassledovanii i analize tekhnologicheskikh narusheniy v perevozochnom protsesse na infrastruk-ture OAO "RZHD" s ispol'zovaniyem avtomatizirovannoy sistemy KASAT: rasporyazheniye OAO "RZHD" N 2160/r ot 1 oktyabrya [On approval of the Regulations on accounting, investigation and analysis of failures in the operation of technical means on the infrastructure of "The Russian Railways" using the automated system KASANT and regulations on accounting, investigation and analysis of technological violations in the transportation process on the infrastructure of JSC "Russian Railways" using the automated system KASAT: order of JSC "Russian Railways" N 2160/r, October, 1], 2018, 69 p.
16. Klassifikator opasnykh otkazov tekhnicheskikh sredstv khozyaystva puti i sooruzheniy utverzhden i vveden v deystviye Rasporyazheniyem OAO "RZHD" ot 28 aprelya [The classifier of dangerous failures of technical means of track management and structures was approved and put into effect by the Order of "The Russian Railways" dated April 28], 2006, 821 p.
17. Gusev S.I. Itogi raboty OAO «RZD» v 2022 gody prevzoshly ogidaniya [The results of the work of «The Russian Railways» in 2022 exceeded expectations]. "Gudok" Ekonomika Elektron. zhurn ["Gudok" Economy], 2023, available at: https://gudok.ru/news/util_avto (accessed: 04/01/2023).
18. Gerasimov A.N. Ekonometrika (prodvinytyi uroven) [Econometrics, advanced level], Moscow, Phoenix, 2016, 272 p., EDN: XADINH.

Mikhail Pavlovich Bazilevskiy. *Associate professor, candidate of technical sciences, Irkutsk state transport university. AuthorID: 679277, SPIN: 4347-5028, ORCID: 0000-0002-3253-5697, mik2178@yandex.ru, Russia, Irkutsk, 15 Chernyshevskogo St.*

Статья поступила в редакцию 28.06.2023; одобрена после рецензирования 11.07.2023; принята к публикации 17.08.2023.
The article was submitted 06/28/2023; approved after reviewing 07/11/2023; accepted for publication 08/17/2023.