

Носков С. И., Овсянников И. В.
S. I. Noskov, I. V. Ovsyannikov

**ОЦЕНИВАНИЕ ПАРАМЕТРОВ КУСОЧНО-ЛИНЕЙНОЙ ФУНКЦИИ РИСКА
С НЕЛИНЕЙНЫМИ ПРЕОБРАЗОВАНИЯМИ НЕЗАВИСИМЫХ ПЕРЕМЕННЫХ**

**ESTIMATION OF PARAMETERS OF A PIECEWISE LINEAR RISK FUNCTION
WITH NONLINEAR TRANSFORMATIONS OF INDEPENDENT VARIABLES**

Носков Сергей Иванович – доктор технических наук, профессор кафедры «Информационные системы и защита информации» Иркутского государственного университета путей сообщения (Россия, Иркутск); Россия, 664074, ул. Чернышевского, д. 15; тел. 8(914)902-24-94. E-mail: sergey.noskov.57@mail.ru.

Sergey I. Noskov – Doctor of Technical Sciences, Professor, Information Systems and Information Protection Department, Irkutsk State Transport University (Russia, Irkutsk); Russia, 664074, st. Chernyshevsky, 15; tel. 8(914)902-24-94. E-mail: sergey.noskov.57@mail.ru.

Овсянников Иван Владимирович – студент Иркутского государственного университета путей сообщения (Россия, Иркутск); Россия, 664074, ул. Чернышевского, д. 15; тел. 8(950)087-23-14. E-mail: bidanocka@gmail.com.

Ivan V. Ovsyannikov – Student, Irkutsk State Transport University (Russia, Irkutsk); Russia, 664074, st. Chernyshevsky, 15; tel. 8(950)087-23-14. E-mail: bidanocka@gmail.com.

Аннотация. В работе предложен алгоритмический способ построения кусочно-линейной функции риска с нелинейными преобразованиями независимых переменных. Этот способ сводится при задействовании метода наименьших модулей для идентификации неизвестных параметров к задаче линейно-булева программирования. Разработаны две версии функции риска для средних цен на жильё в России: обычная и с нелинейными преобразованиями предикторов. Вторая обладает значительно меньшей средней относительной ошибкой аппроксимации.

Summary. The paper proposes an algorithmic method of constructing a piecewise linear risk function with nonlinear transformations of independent variables. This method is reduced to a linear-Boolean programming problem when using the least moduli method to identify the unknown parameters. We have developed 2 versions of the risk function for average housing prices in Russia – the usual and with nonlinear transformations of predictors. The second one has a much smaller average relative approximation error.

Ключевые слова: регрессионная модель, кусочно-линейная функция риска, преобразования переменных, задача линейно-булева программирования, метод наименьших модулей, цены на жильё.

Key words: regression model, piecewise linear risk function, variable transformations, linear-Boolean programming problem, least modulus method, housing prices.

УДК 519.852

Введение. Исследование различных сложных систем, в том числе технических, с помощью методов математического моделирования, в частности регрессионного, часто сопровождается необходимостью анализа ситуаций, связанных с риском в самых разных его возможных трактовках, например оцениваемым уровнем опасности (безопасности). Так, в работе [13] с использованием моделей географически взвешенной логистической регрессии изучаются и анализируются пространственно-временные изменения вкладов факторов, вызывающих лесные пожары в Испании. Изучены несколько подмножеств возникновения пожаров, объединяющих причины (халатность/авария и поджог), сезон (лето-весна и зима-осень) и размер (<1 га, 1...100 га и >100 га). Полученные результаты свидетельствуют о том, что человеческий фактор теряет свою значимость, а климатические факторы выходят на первый план, что в конечном счёте может быть связано с

успехом проводимой в последнее время политики профилактики. Кроме того, обнаружены сильные различия в эффективности моделей возникновения пожаров в разных подмножествах, поэтому модели, основанные на долгосрочных исторических данных о пожарах, могут привести к ошибочным выводам. В [10] предложен метод построения нечёткой линейной регрессии, использующий данные реального времени для точного прогнозирования суточного пикового расхода воды для рек Боу и Локоть в Южной Альберте (Канада). В качестве независимой переменной использовался среднесуточный расход воды с задержкой в один, два, три или семь дней. При реализации модели был разработан уникальный гибридный подход, в котором пиковый расход рассматривался как вероятностный, а среднесуточный – как возможный. В статье [14] линейные регрессионные модели Блэка и Шоулза используются для анализа финансовых рисков морских перевозок.

Работа [12] посвящена применению методов регрессионного анализа для оценки риска подтопления городов из-за ливней с учётом шести пространственных факторов городских поверхностей: плотности населения, плотности дорог, расстояния до водных объектов, процента противодиффузионной поверхности, нормализованного индекса различий растительности и цифровой модели рельефа. Модель смогла объяснить риск подтопления в Ханое (Вьетнам) с уровнем достоверности 67,6 %. Этот результат очень важен при стратегическом планировании для предотвращения и контроля подтопления. Статья [6] основана на применении ранее разработанной авторами модели логистической регрессии для прогнозирования кредитного риска со стороны продавцов с целью его уменьшения в банковской сфере вследствие проявления мошеннических действий. В [7] модель логистической регрессии применяется к данным кредитного скоринга португальского финансового учреждения для оценки риска дефолта по потребительским кредитам. Было выявлено, что риск дефолта увеличивается в зависимости от размера кредита, срока кредита и возраста клиента, но снижается, если клиент владеет большим количеством кредитных карт. Кроме того, клиенты, получающие зарплату в том же банковском учреждении, где выдаётся кредит, имеют меньше шансов на неисполнение обязательств, чем клиенты, получающие зарплату в другом учреждении.

В статье [8] разработана упорядоченная модель логистической регрессии для обеспечения проверки качества оценок внутреннего контроля риска финансовых операций внутренними аудиторами и владельцами бизнеса. Результаты показывают, что предложенная модель может служить инструментом повышения эффективности аудита и обучения, который неспециалисты могут использовать для получения экспертных знаний. В [9] исследуются риски разрушения оползневых плотин. На основе японских данных в работе использовались метод логистической регрессии и техника складного ножа для определения важных геоморфических переменных, таких как максимальный расход (или площадь водосбора), высота плотины, её ширина и длина, влияющих на устойчивость оползневых плотин. В [11] рассматриваются нечёткие модели линейной регрессии с нечёткими входными данными. Разрабатываются модели нечёткой линейной регрессии, нейтральные к риску, не склонные к риску и стремящиеся к риску. Наконец, работа [15] посвящена разработке надёжной и эффективной оценке геологического риска и благоприятности залегания углеводородов для принятия значимого решения о разведке. В статье рассмотрены и обсуждены модели, имеющие отношение к оценке геологического риска и благоприятности, а также представлена процедура, основанная на логистической регрессии. В этом методе количественная логистическая регрессионная связь между наличием углеводородов и ключевыми геологическими факторами получается путём интеграции доступной геолого-геофизической информации и текущих результатов разведки. Это крайне важно для определения вероятности появления углеводородов в простой и понятной форме с более объективным и воспроизводящим результатом.

Расширение пространства переменных кусочно-линейной регрессионной функции риска. Традиционная форма этой функции имеет вид [1]:

$$y_k = \max(\alpha_1 x_{k1}, \alpha_2 x_{k2}, \dots, \alpha_m x_{km}) + \varepsilon_k, \quad k = \overline{1, n}, \quad (1)$$

где y – зависимая переменная; x_i – i -я независимая переменная; α_i – i -й подлежащий оцениванию параметр; ε_k – ошибки аппроксимации; k – номер наблюдения; n – число наблюдений (длина выборки). Будем считать все переменные модели (1) детерминированными и неотрицательными.

В модели (1) переменные имеют, как правило, негативный по отношению к исследуемому объекту смысл (уровень риска, величина убытка, уровень уязвимости, цены на ресурсы и т. д.). При этом в соответствии с её правой частью непропорциональное уменьшение значений одной независимой переменной (локального риска) не приведёт к уменьшению значения выходного фактора (глобального риска).

В [11] задача оценивания параметров модели (1) по методу наименьших модулей, т. е. путём минимизации функции потерь

$$J(\alpha) = \sum_{k=1}^n |\varepsilon_k|$$

сведена к задаче линейно-булева программирования (ЛБП).

В работе [12] для кусочно-линейной регрессионной функции Леонтьева

$$y_k = \min(\alpha_1 x_{k1}, \alpha_2 x_{k2}, \dots, \alpha_m x_{km}) + \varepsilon_k, \quad k = \overline{1, n} \quad (2)$$

сформулирована задача назначения из некоторой совокупности элементарных преобразований

$$F(x) = \{f_1(x), f_2(x), \dots, f_p(x)\}$$

лучшего преобразования для каждого предиктора в смысле минимума заданной функции потерь $J(\alpha)$. В качестве f_j могут выступать в том числе некоторые элементарные функции $\ln(x)$, $\exp(x)$, x^{-1} , x^2 , x^3 , $\sqrt{x}$ и т. д. При этом $f_1(x) = x$. Тем самым осуществляется переход от кусочно-линейной регрессии (2) к функции Леонтьева с нелинейными предикторами:

$$y_k = \min(\bar{\alpha}_1 f_{j_1}(x_{k1}), \bar{\alpha}_2 f_{j_2}(x_{k2}), \dots, \bar{\alpha}_m f_{j_m}(x_{km})) + \bar{\varepsilon}_k, \quad k = \overline{1, n}, \quad (3)$$

где f_{j_i} – преобразование с номером j , оказавшееся лучшим в смысле минимума функции потерь $J(\alpha)$ для i -й объясняющей переменной. Очевидно, что регрессия (3) будет характеризоваться меньшими ошибками аппроксимации в смысле манхэттенской метрики, чем кусочно-линейная зависимость (2).

Задача перехода от регрессии (2) к зависимости (3) в [2] также сведена к задаче ЛБП следующим образом.

Сформируем следующие матрицы:

$$Z^j = \left\| z_{ki}^j \right\|, \quad j = \overline{1, p} \quad k = \overline{1, n} \quad i = \overline{1, m},$$

где $z_{ki}^j = f_j(x_{ki})$. Тогда задача ЛБП для построения регрессии (3) примет вид

$$d_k + u_k - v_k = y_k, \quad k = \overline{1, n}, \quad (4)$$

$$d_k - \sum_{j=1}^g \beta_i^j z_{ki}^j \leq 0, \quad k = \overline{1, n}, \quad i = \overline{1, m}, \quad (5)$$

$$\sum_{j=1}^g \beta_i^j z_{ki}^j - d_k + M_1 \sigma_{ki} \leq M_1, \quad k = \overline{1, n}, \quad i = \overline{1, m}, \quad (6)$$

$$\sum_{i=1}^m \sigma_{ki} = 1, \quad k = \overline{1, n}, \quad (7)$$

$$\beta_i^j - M_2 \delta_i^j \leq 0, \quad j = \overline{1, g}, \quad i = \overline{1, m}, \quad (8)$$

$$\sum_{j=1}^g \delta_i^j = 1, \quad i = \overline{1, m}, \quad (9)$$

$$\beta_i^j \geq 0, \quad i = \overline{1, m}, \quad j = \overline{1, g}, \quad (10)$$

$$\sigma_{ki} \in \{0, 1\}, \quad k = \overline{1, n}, \quad i = \overline{1, m}, \quad (11)$$

$$\delta_i^j \in \{0, 1\}, \quad i = \overline{1, m}, \quad j = \overline{1, g}, \quad (12)$$

$$\sum_{k=1}^n (u_k + v_k) + l \sum_{j=1}^g \sum_{i=1}^m \beta_i^j \rightarrow \min, \quad (13)$$

где M_1 и M_2 – заранее заданные большие положительные числа; l – наперёд заданная малая положительная константа.

После решения задачи ЛБП (4) – (13) параметры модели (3) рассчитываются по формуле

$$\bar{\alpha}_i = \sum_{j=1}^g \beta_i^j,$$

при этом лучшее преобразование для i -й независимой переменной соответствует номеру j при $\delta_i^j = 1$.

Воспользуемся идеей преобразования модели (2) в модель с нелинейными предикторами (3) по отношению к кусочно-линейной функции риска с той же совокупностью преобразований $F(x)$, т. е. поставим задачу перехода от функции риска (1) к кусочно-линейной функции с нелинейными независимыми переменными:

$$y_k = \max(\bar{\alpha}_1 f_{j_1}(x_{k1}), \bar{\alpha}_2 f_{j_2}(x_{k2}), \dots, \bar{\alpha}_m f_{j_m}(x_{km})) + \bar{\varepsilon}_k, \quad k = \overline{1, n}. \quad (14)$$

Используем при этом некоторые приёмы построения простой формы вложенной кусочно-линейной регрессии [3].

Для сведения задачи построения кусочно-линейной функции риска с нелинейными независимыми переменными (14) к задаче ЛБП необходимо в задаче (4) – (13) заменить ограничения (5) и (6) соответственно на

$$d_k - \sum_{j=1}^g \beta_i^j z_{ki}^j \geq 0, \quad k = \overline{1, n}, \quad i = \overline{1, m}, \quad (15)$$

и

$$\sum_{j=1}^g \beta_i^j z_{ki}^j - d_k \geq (\sigma_{ki} - 1)M_1, \quad k = \overline{1, n}, \quad i = \overline{1, m}. \quad (16)$$

Таким образом, для построения кусочно-линейной функции риска с нелинейными независимыми переменными (14) необходимо решить задачу ЛБП с ограничениями (4), (15), (16), (7) – (12) и целевой функцией

$$\sum_{k=1}^n (u_k + v_k) - l \sum_{j=1}^g \sum_{i=1}^m \beta_i^j \rightarrow \min. \quad (17)$$

Заметим, что знак второго слагаемого в (17) противоположен (13), как это сделано в работе [4]. И вновь, как и ранее, после решения задачи ЛБП (4), (15), (16), (7) – (12), (17) параметры модели (14) рассчитываются по формуле

$$\bar{\alpha}_i = \sum_{j=1}^g \beta_i^j, \quad i = \overline{1, m},$$

а лучшее преобразование для i -й независимой переменной соответствует номеру j при $\delta_i^j = 1$.

Численный пример. Воспользуемся описанным выше способом построения кусочно-линейной функции риска с нелинейными независимыми переменными для построения регрессионной модели цен на жильё. В качестве исходных данных используем поквартальную статистическую информацию за период с первого квартала 2017 г. по четвёртый квартал 2020 г. [5].

Введём обозначения: y – средняя цена квадратного метра общей площади на рынке квартир в России, р./м²; x_1, x_2, x_3, x_4 – соответственно средние цены на силикатные блоки, железобетонные плиты перекрытий и песчано-гравийные смеси, р./м².

Обычная кусочно-линейная функция риска (1), построенная на основе ретроспективных данных, примет вид

$$y = \max(17.13x_1, 7.89x_2, 13.79x_3, 67.47x_4), \quad (18)$$

$$E = 3.85 \%, \lambda = (2, 2, 1, 2, 2, 2, 2, 4, 3, 3, 2, 2, 2, 4, 2, 2),$$

где E – средняя относительная ошибка аппроксимации; λ – вектор срабатываний, где λ_k указывает на номер независимой переменной, на которой максимум в модели (18) «сработал» в k -м наблюдении.

Теперь построим модель (14), включив в совокупность $F(x)$ четырнадцать степенных, показательных и логарифмических преобразований:

$$y = \max(16132.4 \log_{10} x_1, 0.089x_2^{1.5}, 13.79x_3, 67.47x_4), \quad (19)$$

$$E = 3.15 \%, \lambda = (2, 2, 1, 2, 3, 2, 2, 4, 3, 3, 2, 2, 2, 2, 2, 2).$$

Сравнение кусочно-линейных функций риска (18) и (19) показывает, что вторая существенно точнее первой по критерию E . При этом в обеих моделях, судя по значениям вектора срабатываний, цена квадратного метра жилья в основном зависит от цены на железобетонные плиты перекрытий – максимум в моделях (18) и (19) на переменной x_2 сработал в одиннадцати наблюдениях из шестнадцати.

Приведём ещё несколько пар кусочно-линейных функций риска (1) и (14) с разными наборами регрессоров среди $x_1, \dots, x_4$, подтверждающих тезис о большей точности второй из них:

$$y = \max(17.14x_1, 7.95x_2, 13.80x_3),$$

$$E = 4.35 \%, \lambda = (2, 2, 1, 2, 2, 2, 2, 2, 3, 3, 2, 2, 2, 2, 2, 2),$$

$$y = \max(0.32x_1^{1.5}, 14921.4 \log_{10} x_2, 0.22x_3^{1.5}),$$

$$E = 3.44 \%, \lambda = (2, 3, 1, 3, 3, 3, 3, 3, 3, 3, 3, 3, 1, 3, 3, 3),$$

$$y = \max(17.06x_1, 7.95x_2, 67.47x_4),$$

$$E = 4.08 \%, \lambda = (2, 2, 1, 2, 2, 2, 2, 3, 2, 2, 2, 2, 2, 3, 2, 2),$$

$$y = \max(16066.5 \log_{10} x_1, 0.087x_2^{1.5}, 20756.33 \log_{10} x_4),$$

$$E = 3.18 \%, \lambda = (2, 2, 1, 2, 3, 3, 3, 3, 3, 3, 2, 2, 2, 2, 2, 2),$$

$$y = \max(18.06x_1, 14.54x_3, 67.47x_4),$$

$$E = 4.42 \%, \lambda = (2, 2, 1, 2, 2, 2, 2, 3, 2, 2, 2, 2, 2, 3, 2, 2),$$

$$y = \max(16739.59 \log_{10} x_1, 0.223x_3^{1.5}, 2.045x_4^{1.5}),$$

$$E = 3.16 \%, \lambda = (2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 2, 2).$$

Заметим, что фактор x_1 в последней модели не «сработал» ни разу.

Заключение. В работе предложен способ построения кусочно-линейной функции риска с нелинейными независимыми переменными, сводящийся в случае применения метода наименьших модулей для вычисления оценок параметров к задаче линейно-булева программирования. Построены обычная функция риска и её аналог с нелинейными предикторами для средних цен на жильё в Российской Федерации.

ЛИТЕРАТУРА

1. Носков, С. И. Идентификация параметров кусочно-линейной функции риска / С. И. Носков // Транспортная инфраструктура Сибирского региона. – 2017. – Т. 1. – С. 417-421.
2. Носков, С. И. Подход к построению функции Леонтьева с нелинейными предикторами / С. И. Носков, Е. С. Попов // Известия Тульского государственного университета. Технические науки. – 2024. – № 2. – С. 105-108.
3. Носков, С. И. Идентификация параметров простой формы вложенной кусочно-линейной регрессии / С. И. Носков // Учёные записки Комсомольского-на-Амуре государственного технического университета. Науки о природе и технике. – 2023. – № III (67). – С. 57-61.
4. Уточнение способов идентификации параметров некоторых кусочно-линейных регрессий / С. И. Носков, М. С. Жукова, Т. К. Кириллова, Ю. О. Купитман, А. А. Хоняков // Научные труды КубГТУ. – 2023. – № 2. – С. 75-81.
5. Носков, С. И. Применение функции риска для модельного описания колебания цен на рынке недвижимости / С. И. Носков, А. А. Хоняков // Инженерно-строительный вестник Прикаспия. – 2021. – № 3 (37). – С. 77-82.
6. Agarwal, N.; Sharma, M. Vikalpa: The Journal for Decision Makers, 39 (3), 2014. – P. 67-76.
7. Costa, E.; Cristina, I.; Correia, A.; Fariac, S. Journal of Applied Statistics, 47, 2020. – P. 2879-2894.
8. Issa, H.; Kogan, A. Journal of Information Systems, 28 (2), 2014. – P. 209-229.
9. Jia-Jyun, D.; Yu-Hsiang, T.; Chien-Chih, C.; Jyh-Jong, L.; Yii-Wen, P. Engineering Geology 117 (1-2). 2011. – P. 52-61.
10. Khan, U.; Valeo, C. Journal of Environmental Informatics, 2016.
11. Modarres, M.; Nasrabadi, E.; Mehdi Nasrabadi, M. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems. Vol. 12 (5), 2004. – P. 635-649.
12. Tran, D.; Xu, D.; Dang, V.; Alwah, A. Water, 12 (3), 2020. – 879 p.
13. Rodríguez, M.; Jiménez-Ruano, A.; Peña-Angulo, D.; Fernández, J., Journal of Environmental Management, 225, 2018. P. 177-192.
14. Yazır, D.; Şahin, B. TransNav, the International Journal on Marine Navigation and Safety of Sea Transportation, 11 (4), 2017. – P. 635-640.
15. Zhaoqun, Z.; Chengyan, L.; Xianguo, Z.; Kai, W.; Jingjing, X.; Sudong, W. Marine and Petroleum Geology, 92, 2018. – P. 65-77.