УДК 519.23

ЗАДАЧИ НЕЧЕТКО-ВЕРОЯТНОСТНОЙ ОПТИМИЗАЦИИ: РЕГРЕССИЯ С НЕЧЕТКИМИ ДАННЫМИ

А.С. Шведов
Национальный исследовательский университет Высшая школа экономики
Россия, 101000, Москва, Мясницкая ул., 20
E-mail: ashvedov@hse.ru

Ключевые слова: нечеткое множество, нечетко-случайная величина, линейная регрессия, несмещенность оценок, состоятельность оценок.

Аннотация: Построение зависимостей типа линейной регрессии между нечеткими множествами представляет интерес не только с теоретической, но и с практической точки зрения. Обычно при этом рассматривается частный вид нечетких множеств – нечеткие числа. С одной стороны, данные могут быть известны лишь приближенно. Тогда нечеткие множества могут использоваться, чтобы передать эту неопределенность. С другой стороны, данные могут быть объединены в некоторые группы. Тогда значение для группы может задаваться в виде нечеткого множества. При нечеткой регрессии рассматриваются те же задачи, что и при обычной регрессии: несмещенность и состоятельность оценок, доверительные интервалы для параметров, проверка гипотез. Для решения этих задач при обычной регрессии необходимо использовать инструментарий теории вероятностей. При нечеткой регрессии это обуславливает комбинированный нечетко-вероятностный подход. В докладе обсуждается понятие нечетко-случайной величины, рассматриваются различные подходы к построению линейных регрессионных моделей с включением нечеткости, приводятся новые результаты о несмещенности и состоятельности оценок.

1. Введение

И методы теории вероятностей, и методы теории нечетких множеств широко применяются в различных задачах оптимизации и управления. Однако, если применение методов теории вероятностей в оптимизации и в управлении является общепринятым, то использование теории нечетких множеств в этих областях – более новый подход (см., например, [1, 2]). И теория вероятностей, и теория нечетких множеств применяются для включения неопределенности в математическую модель. Но речь идет о двух разных видах неопределенности. С одной стороны, разным значениям какой-то величины (или группам значений) могут приписываться некоторые вероятности. С другой стороны, сами понятия могут быть расплывчатыми. Для включения неопределенности может применяться и чисто вероятностный подход, и чисто нечеткий подход, и комбинированный нечетко-вероятностный подход. Причем существуют различные подходы к комбинированию методов теории вероятностей и теории нечетких множеств (см., например, [3]). Оптимизационные задачи очень разнообразны. В дальнейшем, в докладе рассматривается та область, за которой закрепилось название «нечеткая регрессия».

При построении обычных регрессионных зависимостей первый шаг состоит в применении метода наименьших квадратов, и этот шаг не требует использования математического аппарата теории вероятностей, задача решается чисто аналитически. Однако при построении доверительных интервалов для параметров регрессии и при проверке различных гипотез, касающихся этих параметров, применение теории вероятностей становится необходимым. При этом регрессионная зависимость превращается в зависимость между случайными величинами. Однако данные, по которым строится регрессия, могут включать некоторую неоднозначность. Например, в качестве дневной цены акции может приниматься средняя за день цена. Но тогда при построении регрессионной зависимости теряется информация о максимальной и о минимальной цене акции за день, о цене закрытия. Вся эта информация может учитываться, если дневная цена акции рассматривается как нечеткое число. Такой подход доказал свою эффективность в задаче оценивания акций. При использовании нечетких данных регрессионная зависимость становится зависимостью между нечетко-случайными величинами. Либо между нечеткими числами, которые рассматриваются как реализации нечетко-случайных величин. Требуют изучения и методы построения зависимостей между нечеткими числами, и такие свойства оценок параметров как несмещенность и состоятельность, и вопросы, касающиеся построения доверительных интервалов и проверки гипотез.

В первой части доклада дается обзор некоторых публикаций по нечеткой регрессии. Во второй части доклада достаточно подробно рассматривается один из подходов к построению регрессионных зависимостей для нечетких данных. Здесь термины «нечеткое число», «нечетко-случайная величина» используются так же, как в статьях [4, 5]. Однако в современной научной литературе при определении этих понятий устойчиво сохраняется некоторая вариативность. При общем согласии, что нечеткое число – это частный вид нечеткого множества, когда универсальным множеством является множество действительных чисел, и функция принадлежности сначала монотонно не убывает, затем – монотонно не возрастает, а нечетко-случайная величина – это некоторое обобщение случайной величины, детали разнятся от работы к работе. (Некоторые авторы различают понятия нечетко-случайной величины и случайно-нечеткой величины. В первом случае речь идет об измеримой функции, значениями которой являются не обычные действительные числа, а нечеткие числа. Во втором случае рассматриваются какие-то распределения вероятностей, например, нормальное или гамма-распределение, но параметрами являются нечеткие числа.) В первой, обзорной, части доклада, разумеется, не даются детально все определения из всех работ.

2. Регрессионные зависимости, включающие нечеткие множества (краткий обзор публикаций)

Во-первых, в настоящем докладе мы не касаемся подхода, предложенного в работе [6], где рассматриваются системы нечетких правил, и каждое нечеткое правило содержит свою регрессионную зависимость. При этом разные регрессионные зависимости строятся, например, для малых, средних и больших значений некоторого показателя. При конкретном значении показателя каждая из построенных регрессионных зависимостей используется с некоторым весом, который определяется тем, каковы степени принадлежности данного значения к нечетким множествам малых, средних и больших значений. Исходные данные, используемые в каждом из нечетких правил, могут быть нечеткими множествами. Но в докладе рассматриваются только задачи, где строится одна регрессионная зависимость для всех значений.

Число публикаций и по развитию методов нечеткой линейной регрессии, и по применению соответствующих моделей в различных прикладных областях очень велико. Рассматриваются три разных типа линейных регрессионных зависимостей:

(1) $\tilde{y} = \tilde{a}_0 + \tilde{a}_1 x_1 + ... + \tilde{a}_m x_m$,

(2) $\tilde{y} = \tilde{a}_0 + a_1 \tilde{x}_1 + ... + a_m \tilde{x}_m$,

(3) $\tilde{y} = \tilde{a}_0 + \tilde{a}_1 \tilde{x}_1 + ... + \tilde{a}_m \tilde{x}_m$,

тильда обозначает нечеткое число. К моделям с нечеткими данными относятся модели типа (2) и (3). Модели типа (1) также широко используются, см. [7] и последующие работы; обзор этого направления, включая и ряд прикладных работ, дается, например, в [8]. Оказывается, очень существенно, используется в моделях типа (2) нечеткий свободный член $\tilde{a}_0$ или четкий свободный член a_0. Подробнее об этом говорится во второй части доклада.

Другая классификация нечетких линейных регрессионных моделей может быть дана по типам используемых нечетких чисел. Многие авторы ограничиваются рассмотрением треугольных или трапецоидальных нечетких чисел, когда функция принадлежности является непрерывной на всей прямой и равняется 0 до некоторого значения, затем линейно возрастает от 0 до 1, для трапецоидальных нечетких чисел может быть участок, на котором функция принадлежности равняется 1, затем эта функция линейно убывает от 1 до 0 и затем равняется 0 после некоторого значения. Значительно более общей конструкцией, но все же не самой общей, являются так называемые *LR* нечеткие числа. От трапецоидальных нечетких чисел они отличаются тем, что вместо линейных функций используются две произвольные непрерывные монотонные функции, неубывающая от 0 до 1 функция L (левая) и невозрастающая от 1 до 0 функция R (правая). Функции L и R вместе со своими областями определения задают функцию принадлежности нечеткого числа. Наконец, наиболее общей конструкцией является задание нечеткого числа не с помощью одной функции принадлежности, а с помощью двух функций, определенных на отрезке $[0,1]$. Такие нечеткие числа используются во второй части доклада. Отметим также, что как о частном случае трапецоидальных нечетких чисел можно было бы говорить о прямоугольных нечетких числах, хотя этот термин используется редко. Это соответствует тому, что интервальная арифметика является частным случаем нечеткой арифметики.

В работе [7] метод наименьших квадратов не используется, вместо этого решается некоторая задача математического программирования. Однако в [9, 10] и во многих последующих работах именно метод наименьших квадратов стал основной «идеологией» при построении нечетких линейных регрессионных моделей; обзор этого направления дается, например, в работе [11]. Упомянем также про существующие исследования по робастной нечеткой регрессии (см., например, [12, 13]) и по нечеткой регрессии, основанной на методе опорных векторов (см., например, [14]).

Наконец, работы по нечеткой линейной регрессии можно разделить на две группы еще по одному признаку. К первой группе относятся работы, где основное внимание уделяется алгоритмам для определения параметров моделей, и на примерах проводится численное сравнение различных подходов. В работах второй группы изучаются статистические свойства. В качестве примеров работ, относящихся ко второй группе, можно назвать [15-21].

3. Использование методов вариационного исчисления в задачах нечеткой регрессии

Для метода, предложенного в работе [22] и получившего свое развитие в работах [23, 24], рассмотрим вопрос, какие преимущества дает включение нечеткого свободного члена в модели (2) по сравнению с четким свободным членом, и какие трудности при таком обобщении возникают. В работе [22] рассматриваются только модели с четким свободным членом, в работах [23, 24] – модели с нечетким свободным членом. Поскольку во втором случае требуется искать некоторую функцию (или функции), применяется вариационное исчисление. Насколько нам известно, в работах [23, 24] вариационное исчисление применяется для решения задач нечеткой регрессии впервые.

В соответствии с подходом, использовавшемся, например, в работе [4], нечеткое число $\widetilde{c}$ определяется двумя функциями $c^L : [0,1] \to R$, $c^R : [0,1] \to R$, где R – множество действительных чисел. Обе эти функции являются непрерывными слева, функция c^L монотонно неубывающая, функция c^R монотонно невозрастающая, $c^L(1) \le c^R(1)$. Нечеткое число – это множество точек $(\xi, \eta) \in R^2$ таких, что $0 \le \eta \le 1$, $c^L(\eta) \le \xi \le c^R(\eta)$ при всех η. При таком определении нечеткое число $\widetilde{c}$ является компактным множеством. Нечеткие числа $\widetilde{c}$, для которых функции c^L и c^R совпадают, отождествляются с действительными числами. Операции умножения действительного числа на нечеткое число, результатом которой является нечеткое число, и сложения нечетких чисел описываются в [4]. Пусть для нечетких чисел $\widetilde{c}$ и $\widetilde{d}$

$$\rho^2\left(\widetilde{c}, \widetilde{d}\right) = \int_0^1 \left(c^L(\eta) - d^L(\eta)\right)^2 d\eta + \int_0^1 \left(c^R(\eta) - d^R(\eta)\right)^2 d\eta .$$

Задача состоит в нахождении нечеткого числа $\widetilde{a}_0$ и действительных чисел $a_1, \ldots, a_m$, доставляющих минимум функционалу

$$\sum_{i=1}^{n} \rho^2\left(\widetilde{y}_i, \widetilde{a}_0 + a_1 \widetilde{x}_{1i} + \ldots + a_m \widetilde{x}_{mi}\right). \tag{4}$$

Как это предложено в [23], минимизация функционала (4) проводится в два этапа. На первом этапе при фиксированных $a_1, \ldots, a_m$ ищутся функции a_0^L, a_0^R (разумеется, зависящие от $a_1, \ldots, a_m$). Для этого решаются задачи вариационного исчисления. На втором этапе ищется минимум функции m переменных $a_1, \ldots, a_m$ и таким образом строится оценка $\hat{a}_1, \ldots, \hat{a}_m$.

Оказывается, что построенная оценка $\hat{a}_1, \ldots, \hat{a}_m$ отличается от оценки для параметров регрессии из работы [22], где свободный член четкий. При этом оценка для модели с нечетким свободным членом обладает свойством близким к несмещенности. (Для случая $m = 1$ и трапецоидальных нечетких чисел этот результат доказывается в работе [24]. Но используя теорему типа Фубини, доказанную в [4], результат можно получить и в общем случае.) Оценка из работы [22] подобным свойством не обладает, и наши попытки использовать эту оценку для построения доверительных интервалов и проверки гипотез о значениях параметров регрессионных моделей к успеху не привели. Также в докладе приводится результат о состоятельности оценки $\hat{a}_1, \ldots, \hat{a}_m$ для модели с нечетким свободным членом. Результаты численного сравнения моделей с четким и с нечетким свободными членами, приведенные в [23], показывают, что точность модели с нечетким свободным членом выше.

4. Заключение

В докладе рассматривается необходимый математический инструментарий, и сопоставляются некоторые подходы к построению регрессионных зависимостей, когда входными и выходными данными являются нечеткие числа. Приводятся результаты о несмещенности и состоятельности оценок.

В ряде задач управления оказывается, что использование нечетких множеств типа 2 дает лучшие результаты, чем использование нечетких множеств типа 1. Одним из преимуществ подхода к построению регрессионных зависимостей из работ [23,24] является то, что основные результаты обобщаются для случая, когда входными и выходными данными являются нечеткие числа типа 2.

Список литературы

1. Lilly J.H. Fuzzy Control and Identification. Hoboken (NJ): Wiley, 2010. 231 p.
2. Zhang H., Liu D. Fuzzy Modeling and Fuzzy Control. Boston (MA): Birkhäuser, 2006. 416 p.
3. Лю Б. Теория и практика неопределенного программирования. М.: БИНОМ. Лаборатория знаний, 2005. 416 с.
4. Шведов А.С. Оценивание средних и ковариаций нечетко-случайных величин // Прикладная эконометрика. 2016. Т. 42, № 1. С. 121-138.
5. Шведов А.С. Нечеткое математическое программирование: краткий обзор // Проблемы управления, 2017. № 3. С. 2-10.
6. Takagi T., Sugeno M. Fuzzy Identifications of Systems and Its Applications to Modeling and Control // IEEE Transactions on Systems, Man, and Cybernetics. 1985. Vol. 15, No. 1. P. 116-132.
7. Tanaka H., Uejima S., Asai K. Linear Regression Analysis and Fuzzy Model // IEEE Trans. Systems Man Cybernetics. 1982. Vol. 12, No. 6. P. 903–907.
8. Kahraman C., Beşkese A., Bozbura F.T. Fuzzy Regression Approaches and Applications // Fuzzy Applications in Industrial Engineering. Berlin: Springer, 2006. P. 589-615.
9. Celmiņš A. Least Squares Model Fitting to Fuzzy Vector Data // Fuzzy Sets and Systems. 1987. Vol. 22, No. 3. P. 245-269.
10. Diamond P. Fuzzy Least Squares // Information Sciences. 1988. Vol. 46, No. 3. P. 141–157.
11. Chachi J., Taheri S.M. Multiple Fuzzy Regression Model for Fuzzy Input-Output Data // Iranian J. of Fuzzy Systems. 2016. Vol. 13, No. 4. P. 63-78.
12. Chen Y.-S. Outliers Detection and Confidence Interval Modification in Fuzzy Regression // Fuzzy Sets and Systems. 2001. Vol. 119, No. 2. P. 259-272.
13. D'Urso P., Massari R., Santoro A. Robust Fuzzy Regression Analysis // Information Sciences. 2011. Vol. 181, No. 19. P. 4154–4174.
14. Hong D.H., Hwang C. Support Vector Fuzzy Regression Machines // Fuzzy Sets and Systems. 2003. Vol. 138, No. 2. P. 271-281.
15. Akbari M.G., Sadegh M.K. Estimators Based on Fuzzy Random Variables and Their Mathematical Properties // Iranian J. of Fuzzy Systems. 2012. Vol. 9, No. 1. P. 79-95.
16. Arnold B.F., Gerke O. Testing Fuzzy Linear Hypotheses in Linear Regression Models // Metrika. 2003. Vol. 57, No. 1. P. 81-95.
17. González-Rodríguez G., Blanco Á., Colubi A., Lubiano M.A. Estimation of a Simple Linear Regression Model for Fuzzy Random Variables // Fuzzy Sets and Systems. 2009. Vol. 160, No. 3. P. 357-370.
18. Lee W.-J., Jung H.Y., Yoon J.H., Choi S.H. The Statistical Inferences of Fuzzy Regression Based on Bootstrap Techniques // Soft Comput. 2015. Vol. 19, No. 4. P. 883-890.
19. Lin J.-G., Zhuang Q.-Y., Huang C. Fuzzy Statistical Analysis of Multiple Regression with Crisp and Fuzzy Covariates and Applications in Analyzing Economic Data of China // Computational Economics. 2012. Vol. 39, No. 1. P. 29-49.
20. Näther W. Regression with Fuzzy Random Data // Computational Statistics and Data Analysis. 2006. Vol. 51, No. 1. P. 235–252.
21. Watada J., Wang S., Pedrycz W. Formulation of Fuzzy Random Regression Model // New Advances in Intelligent Signal Processing. Berlin: Springer, 2011. P. 1-20.
22. Bargiela A., Pedrycz W., Nakashima T. Multiple Regression with Fuzzy Data // Fuzzy Sets and Systems. 2007. Vol. 158, No. 19. P. 2169-2188.
23. Вельдяксов В.Н., Шведов А.С. О методе наименьших квадратов при регрессии с нечеткими данными // Экономический журнал ВШЭ. 2014. Т. 18, № 2. С. 328–344.
24. Вельдяксов В.Н., Шведов А.С. Проверка гипотез при регрессии с нечеткими данными // Экономический журнал ВШЭ. 2014. Т. 18, № 3. С. 508-521.