Минобрнауки России

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г.
ЧЕРНЫШЕВСКОГО»

Кафедра математической экономики

НЕЧЕТКАЯ ЛИНЕЙНАЯ РЕГРЕССИЯ ДЛЯ ОЦЕНКИ НЕДВИЖИМОСТИ В САРАТОВЕ

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студента 4 курса 451 группы
направления 38.03.05 Бизнес-информатика
Пилюгина Дмитрия Владимировича

Научный руководитель
доцент, к.ф.-м.н., доцент _____В.В.Новиков

Заведующий кафедрой
д.ф.-м.н., профессор _____С. И. Дудов

Саратов 2020

Введение

Известно немало методов оценки недвижимости. Разнообразие методов обусловлено целями оценки, экономическими моделями, лежащими в основе оценки, видом доступных данных и многими другими факторами. Типичным для большинства моделей является то, что в них так или иначе приходится учитывать фактор неопределенности. Оценки недвижимости разбиты на две большие группы. К первой (традиционные подходы) отнесены подходы, основанные на регрессионных моделях, сравнительный, затратный и доходный подходы. Во вторую группу (продвинутые, или новые подходы) включены подходы, основанные на «гедонистической» функции цены (оценка потребительских качеств, не имеющих непосредственно рыночной цены), подходы, связанные с пространственным анализом, а также подходы, основанные на методах теории нечетких множеств и искусственных нейронных сетях. В эту группу включены подходы, в которых для моделирования неопределенности используются методы, отличные от теоретико вероятностных. Тот факт, что неопределенность в экономических явлениях в общем и в оценке недвижимости в частности может не вполне адекватно описываться стохастическими методами, отмечался многими исследователями. Как альтернатива теории вероятностей в ряде работ при оценке недвижимости использовались нечеткие множества и мягкие вычисления. В большинстве посвященных этой тематике работ, в которых используются методы теории нечетких множеств, модели являются нелинейными. Это позволяет добиться высокой согласованности с исходными данными, но осложняет экономическую интерпретацию полученных результатов. В настоящейработе мы применяем для оценки нечеткую линейную регрессию. Обладая необходимой гибкостью, эта модель допускает достаточно прозрачную интерпретацию.

Применение линейной регрессии позволяет количественно оценить влияние учитываемых факторов x_1,...,x_n на результирующий показатель y. Классическая модель линейной регрессии описывается уравнением

y=a_0+a_1x_1+...+a_nx_n+ϵ

где a_0,a_1,...,a_n — числовые коэффициенты, отражающие влияние факторов, а значение результирующего показателя искажено случайной погрешностью ϵ. Зная распределение случайной велечины ϵ и имея достаточный запас наблюдений (y_j; x_{1j},..., x_{nj}), можно тем или иным способом восстановить зна-

чения коэффициентов a_0, a_1,..., a_n. Ключевую роль здесь играет предположение о вероятностном характере неопределенности. Применение вероятностных методов обосновано тогда, когда выполняется ряд условий, допускающих в определенном смысле физическую» проверку.

При моделировании экономических явлений эти условия выполняются далеко не всегда. Возможна ситуация, когда линейная зависимость носит приближенный характер, причем неточность обусловлена не случайными погрешностями наблюдений, а неопределенностью значений коэффициентов, тем, что не учтены некоторые факторы, и вообще сама зависимость не является стохастической. В этом случае адекватным решаемой задаче может оказаться поиск нечетких коэффициентов,наиболее точно соответствующих имеющимся данным.

В теории нечетких множеств не предполагается, что неопределенная величина подчиняется объективно проверяемым закономерностям (как в теории вероятностей), но в то же время на интервале еезначений задано некоторое распределение возможностей, отражающее большую или меньшую (субъективно оцениваемую) реалистичность в принципе возможных значений. Методы теории нечетких множеств дают полезную информацию для принятия обоснованных решений, хотя и не имеют такой точной интерпретации, как вероятностные.

Целью настоящей курсовой работы является изучение подхода к построению модели нечеткой линейной регрессии из, которая применялась авторами для исследования стоимости квадратного метра в частном доме и для оценки стоимости квартиры в Москве. В рамках достижения данной цели были поставлены следующие задачи: - изучение теории нечетких множеств; - изучение основ классического и нечеткого регрессионного анализа; - знакомство с методами построения нечетких регрессионных моделей; - на основе изученного теоретического материала написать на языке Python 3.6 программу, реализующую алгоритм нечеткой регрессии для оценки стоимости квартиры в Саратове.

Основное содержание работы

Работа состоит из 4 частей:

1)Необходимые сведение о нечетких данных;

2)Построение регрессии для четких данных;

3)Нечеткая линейная регрессия;

4)Нечеткая регрессия для оценки недвижимости с помощью языка программирования.

Причем в главе про построение регрессии для четких данных есть несколько разделов:

1)основные гипотезы;

2)дифференцирование функций по векторному аргументу;

3)метод наименьших квадратов. Теорема Гаусса-Маркова.

Целью настоящей выпускной квалификационной работы является изучение подхода к построению модели нечеткой линейной регрессии, которая применялась авторами для исследования стоимости квартиры в Москве. В рамках достижения данной цели были поставлены следующие задачи:

-изучение теории нечетких множеств;

-изучение основ классического и нечеткого регрессионного анализа;

-Знакомство с методами построения нечетких регрессионных моделей;

-на основе изученного материала написать на любом языке программирования программу, реализующую алгоритм нечеткой регрессии для оценки стоимости квартиры в Москве.

Определение №1

Нечеткая величина подмножество A задается своей функцией принадлежности μ_A:$\mathbb{R}\to[0;1]$. Множество всех x, для которых $\mu_A(x)\leq 1$, называется множеством уровня a. Мы будем обозначать его A^a.

Число a, для которого $\mu_A(\mathrm{a})=1$, называется модальным значением величины A. Замыкание множества $\{\mathrm{x}|\mu_A(\mathrm{x})>0\}$ называется носителем нечеткой величины A и обозначается supp A. Это множество считается множеством нулевого уровня и обозначатся также через A^0. «Обычное» четкое число a можно рассматривать как нечеткое с носителем, состоящим из единственной точки a, которая является модальным значением.

Нечеткие величины, описываемые выражениями типа «примерно a», обычно представляют так называемыми треугольными нечеткими числами. Треугольное нечеткое число A задается тройкой чисел $(a^L;a;a^R)$, такой, что

$a^L \le a \le a^R$. Отрезок $[a^L;a^R]$ является носителем множества A, числа a - модальным значением, а уровневые множества имеют следующий вид:

A^a=[(1-α)a^L + αa;(1-α)a^R+αa].

Для обозначения треугольного нечеткого числа мы будем использовать задающую его тройку (a^L;a;a^R). С учетом этого соглашения тройка (k; k; k) задает четкое число k. Треугольное нечеткое число A=(a^L;a;a^R) называется симметричным, если a-a^L=a^R-a. Симметричное треугольное нечеткое число задается тройкой вида (a-d;a;a+d), $d\ge 0$, число d будем называть мерой нечеткости треугольного симметричного нечеткого числа. Арифметические операции с нечеткими велечинами определяются на основе принципа обобщения. Сумма треугольных нечетких чисел A=(a^L;a;a^R) и B=(b^L;b;b^R). - это треугольное нечеткое число

A+B=(a^L;a;a^R)+(b^L;b;b^R)

Произведение треугольного нечеткого числа A=(a^L;a;a^R) на положительное четкое число k - это треугольное нечеткое число.

kA=(ka^L; ka;ka^R).

Как легко заметить, множество симметричных треугольных нечетких чисел замкнуто относительно сложения и умножения на четкое число.

Для сравнения нечетких величин используются различные методы, как правило, основанные на мерах возможности и необходимости. В общем случае возможность отношения $A \le B$ оценивается числом

Pos($A \le B$)=sup{min(μ_A(x),μ_B(y))|x$\le$y; x, $y \in R$}

В частности, если A=(a^L; a; a^R) - треугольник нечеткое число, а B=b - четкое число, имеем:

$$POS(A \le \text{B}) = \begin{cases} 1, \text{a} \le b \\ \mu_A(b),\ b \le a; \end{cases}$$

Отсюда

Pos(A=b)=min{Pos($A \le b$), Pos($A \ge b$)} = $\mu_A(b)$.

Таким образом, возможность того, что нечеткая величина A принимает значение b, оценивается значением функции принадлежности для этого значения.

Определение №2

Естественным обобщением линейной регрессионной модели с двумя переменными является многомерная регрессионная модель, или модель множественной регрессии [8]:

$y_t=\beta_1+\beta_2 x_{t2}+...+\beta_k x_{tk}+\epsilon_t$, $t=1,..., n$,

или

$y_t=\beta_1 x_{t1}+\beta_2 x_{t2}+...+\beta_k x_{tk}+\epsilon_t$, $t=1,..., n$,

где x_{tp} - значения регрессора x_p в наблюдении t, а $x_{t1}=1$, $t=1,.., n$. С учетом этого замечания мы не будем далее различать модели со свободным членом или без свободного члена.

Определение №3

Гипотезы, лежащие в основе модели множественной регрессии, являются естественным обобщением модели парной регрессии [8]:

1. $y_t=\beta_1 x_{t1}+\beta_2 x_{t2}+...+\beta_k x_{tk}+\epsilon_t$, $t=1,...,n$ - спецификация модели.

2. $x_{t1},...,x_{tk}$ - детерминированные величины. Векторы $x_s=(x_{1s},...,x_{ns})'$, $s=1,...,k$ линейно независимы в R^n.

3a. $E\epsilon_t=0$, $E(\epsilon_t^2)=V(\epsilon_t)=\sigma^2$ - не зависит от t.

3b. $E(\epsilon_t\epsilon_s)=0$ при $t\neq s$ - статистическая независимость (некоррелированность) ошибок для разных наблюдений. Часто добавляется следующее условие.

3c. Ошибки ϵ_t, $t=1,...,n$ имеют совместное нормальное распределение: $\epsilon_t\sim N(0,\sigma^2)$.

В этом случае модель называется нормальной линейной регрессионной (classical normal linear regression model).

Гипотезы, лежащие в основе множественной регрессии, удобно записать в матричной форме, которая главным образом и будет использоваться в дальнейшем.

Пусть y обозначает $n\times 1$ матрицу (вектор-столбец) $(y_1,...,y_n)'$, $\beta=(\beta_1,...,\beta_k)'$ - $k\times 1$ вектор коэффициентов; $\epsilon=(\epsilon_1,...,\epsilon_n)'$ - $n\times 1$ вектор ошибок;

Столбцами матрицы X являются $n\times 1$ векторы регрессоров $x_s = (x_{1s},...,x_{ns})'$, $s=1,...,k$. Условия 1-3 в матричной записи выглядят следующим образом:

1. y=$X\beta$+ϵ -спецификация модели;

2. X-детерминированная матриц, имеет максимальный ранг k;

3a,b. $E(\epsilon)$=0; $V(\epsilon)$=$E(\epsilon\epsilon')$=$\sigma^2 I_n$;

дополнительное условие:

3c. $\epsilon \sim N(0;\sigma^2 I_n)$, т.е. ϵ -нормально распределенный случайный вектор со средним 0 и матрицей ковариаций $\sigma^2 I_n$ (нормальная линейная регрессионная модель).

Определение №4

Как и в случае регрессионного уравнения с одной переменной, целью метода выбора вектора оценок $\widehat{\beta}$ [8], является минимизация суммы квадратов остатков ϵ_t (т.е. квадрата длины вектора остатков e):

e=y-$\widehat{y}$=y-$X\widehat{\beta}$,

Выразим ee' через X и β:

ee'=$(y\text{-}X\widehat{\beta})'(y\text{-}X\widehat{\beta})$=$y'y\text{-}y'X\widehat{\beta}\text{-}\widehat{\beta}X'y+\widehat{\beta}X'X\widehat{\beta}$=$y'y\text{-}2\widehat{\beta}X'y+\widehat{\beta}X'X\widehat{\beta}$.

Необходимые условия минимума ESS получаются дифференцированием по вектору $\widehat{\beta}$ (см. предыдущий раздел):

откуда, учитывая обратимость матрицы $X'X$ в силу условия, находим оценку метода наименьших квадратов:

$\widehat{\beta}_{OLS} = (X'X)^{-1}X'y$.

Покажем, что, как и в случае одного регрессора, означает, что вектор остатков e ортогонален всем независимым переменным x_1,...,x_k (столбцам матрицы X). Условие $x_1'e$=...=$x_k'e$=0 эквивалентно равенству $X'e$=0. Действительно,

$X'e$=$X'(y\text{-}X\widehat{\beta})$=$X'y\text{-}X'X\widehat{\beta}$ = $X'y\text{-}X'X(X'X)^{-1}X'y$=0

Получим полезную в дальнейшем формулу для суммы квадратов остатков

$e'e$=$y'y\text{-}2\widehat{\beta}'X'y+\widehat{\beta}'X'X\widehat{\beta}$=$y'y\text{-}\widehat{\beta}$'$(2X'y\text{-}X'X(X'X)^{-1}$X'y$)$=$y'y\text{-}\widehat{\beta}'X'y$.

Геометрическая интерпретация в основном совпадает с геометрической интерпритацией регрессионного уравнения с одной независимой переменной. Представим y,x_1,...,x_k как векторы в n-мерной евклидовом пространстве R^n. Векторы $x_1, ..., x_k$, порождают k-мерное пространство π.

Вектор $\widehat{y}$=$X\widehat{\beta}$ есть ортогональная проекция вектора y на гиперплоскость π.

Вектор остатков e=y-$\widehat{y}$ ортогонален подпространству π.

Как и в случае регрессионного уравнения с одной независимой переменной, можно показать, что оценка метода наименьших квадратов является оптимальной.

Теорема №1

ТЕОРЕМА ГАУССА-МАРКОВА. Предположим, что:

1. y=$X\beta$+ϵ

2. X - детерминированная $n\times k$ матрица, имеющая максимальный ранг k;

3. $E(\epsilon)$=0; $V(\epsilon)$=$E(\epsilon\epsilon')$=$\sigma^2 I_n$

Тогда оценка метода наименьших квадратов $\widehat{\beta}_{OLS} = (X'X)^{-1}X'y$ является наиболее эффективной (в смысле наименьшей дисперсии) оценкой в классе линейных (по y) несмещенных оценок (BEST LINEAR UNBIASED ESTIMATOR, BLUE)[8].

Доказательство. Обозначим $A = (X'X)^{-1}X'$, $\widehat{\beta}_{OLS}$=A_y. Любую другую линейную оценку вектора параметров β можно без ограничения общности представить в виде: b=$(A$+$C)y$, где C - некоторая $k\times n$ матрица.

1. Покажем, что МНК-оценка является несмешенной оценкой β:

$E\widehat{\beta}_{OLS} = E((X'X)^{-1}X'y)$=$(X'X)^{-1}X'E(y)$

=$(X'X)^{-1}X'E(X\beta$+$\epsilon)$

=$(X'X)^{-1}X'X\beta$+$(X'X)^{-1}X'E\epsilon$=β

Из условия несмещаемости оценки b получаем, что для всех β справедливо соотношение

β=Eb=$(A$+$C)Ey$=$(A$+$C)X\beta$=$(I$+$CX)\beta$,

откуда следует, что CX=0.

2.Подсчитаем матрицу ковариаций МНК- оценки:

$V(\widehat{\beta}_{OLS})$ =$V(Ay)$ =$A\,V(y)A$'= $A\delta^2 IA'$

=$\delta^2(X'X)^{-1}X'X(X'X)^{-1}$=$\delta^2(X'X)^{-1}$

(здесь мы использовали симметричность матрицы $X'X$ и свойство матрицы ковариаций)

3.Используя полученное выше равенство CX=0, получаем

b-β=$(A$+$C)y$ - β=$(A$+$C)X\beta$+$(A$+$C)\epsilon$-β

=$AX\beta$-β+$CX\beta$+$(A$+$C)\epsilon$=$(A$+$C)\epsilon$,

т.к. CX=0 и AX=I. вычислим теперь матрицу ковариаций вектора b:

$V(b)$=$E((b\text{-}\beta)(b\text{-}\beta)')$=$E((A+C)\epsilon\epsilon'(A+C)')$

=$(A+C)\delta^2 I(A+C)'$=$\delta^2(AA'+CA'+AC'+CC')$

=$\delta^2((X'X)^{-1}X'X(X'X)^{-1}+CX(X'X)^{-1}$

+$(X'X)^{-1}X'C'+CC')$=$\delta^2(X'X)^{-1}+\delta^2 CC'$.

Таким образом,

$V(b)$=$V(\widehat{\beta}_{OLS})+\delta^2 CC'$.

Матрица CC' неотрицательно определена, поэтому

$V(b)\geq V(\widehat{\beta}_{OLS})$.

Отсюда следует утверждение теоремы. В самом деле, i-й диагональный элемент матрицы $V(b)$ равен дисперсии i-й компоненты вектора коэффициентов $V(b_i)$. Поэтому следует соответствующее неравенство для дисперсий оценок коэффициентов регрессии

$V(b_i)\geq V(\widehat{\beta}_{OLS})$,

что и требовалось доказать.

Определение №5

Модель нечеткой регрессии к настоящему времени получила достаточно широкое применение.

Опишем общую схему нечеткой линейной регрессии.

В случае нечеткости уравнение приобретает вид

Y=A_0x_0+A_1x_1+...+A_nx_n,

где коэффициенты могут быть нечеткими числами.

В общем случае задача нечеткой линейной регрессии может быть поставлена следующим образом.

На основе m результатов наблюдений (y_j,x_j) требуется оптимальным образом определить, вообще говоря, нечеткие коэффициенты A_0, A_1,...,A_n. Оптимальность выражается двумя условиями:

(R1) для каждого j число y_j принадлежит носителю нечеткой величины

Y_j=A_0+A_1x_{1j}+A_2x_{2j}+...+A_nx_{nj}, j=1, 2,..., m.

(R2) суммарная мера нечеткости величин Y_j минимальна.

Иногда первое условие заменяют более сильным требованием:

(R1h) для каждого j выполняется неравенство

$\mu_{y_j}(y_j)\geq h$,

где h - некоторое заданное наперед пороговое значение.

При такой постановке будем говорить о задаче нечеткой линейной регрессии с пороговым значением h.

Если речь идет о поиске коэффициентов в виде симметричных треугольных нечетких чисел, задача нечеткой линейной регрессии сводится к задаче линейного программирования.

Будем искать коэффициентов A_j в виде

A_i=(a_i-d_i, a_i, a_i+d_i).

Тогда Y_j имеет следующий вид:

Y_j=(z_j-r_j, z_j, z_j+r_j).

Суммарная мера нечеткости вычисляется по формуле

$r=\sum_{j=1}^{m} r_j = \mathrm{m}d_0 + \sum_{j=1}^{m}\sum_{i=1}^{n} d_i|x_{ij}|$.

Задача нечеткой линейной регрессии с ограничениями (R1h) сводится к следующей задаче линейного программирования:

$r=\sum_{j=1}^{m} r_j = \mathrm{m}d_0 + \sum_{j=1}^{m}\sum_{i=1}^{n} d_j|x_{ij}| \to \min$;

$y_j \geq \sum_{i=0}^{n} a_i x_{ij} - (1-h)\sum_{i=0}^{n} d_i|x_{ij}|$, j=1,2,...,m;

$y_j \leq \sum_{i=0}^{n} a_i x_{ij} + (1-h)\sum_{i=0}^{n} d_i|x_{ij}|$, j=1,2,...,m;

$d_i \geq 0$, i=0,1,2,...,n.

При h=0 получается решение задачи нечеткой линейной регрессии с ограничениями (R1).

Для того чтобы сравнить нечеткую линейную регрессию с классической, рассмотрим случай n=1.

Пусть

$y_j = a + bx_j + \epsilon_j$, j=1,2,...,m.

Предположим, что погрешности ϵ_j не коррелированы и одинаково нормально распределены, $\epsilon_j \sim \mathrm{N}(0,\sigma^2)$. Пусть $\delta = \max|\epsilon_j|$, тогда

$a-\delta+\mathrm{b}x_j \leq y_j \leq \mathrm{a}+\delta+\mathrm{b}x_j$,

так как носитель нечеткой величины $(\mathrm{a}-\delta;\mathrm{a};\mathrm{a}+\delta)+\mathrm{b}x_j$ содержит число y_i.

Пусть далее значения a_0,d_0,a_1,d_1 найдены решения задачи нечеткой линейной регрессии. Тогда

$md_0 + d_1\sum_{j=1}^{m}|x_j| \leq \mathrm{m}\delta$.

Основываясь на этом неравенстве, можно оценить меру нечеткости найденного решения. Например, решим уравнение

$\mathrm{P}(d_0+d_1|\overline{x}|\leq\delta)>0{,}8$

относительно δ. Используя распределение максимумов нормального распределения, получаем $\delta\geq 2{,}68\delta$. Таким образом, с вероятностью 0,8 нечеткая полоса $Y=A_0+A_1x$ окажется достаточно «узкой».

Заключение

В работе рассмотрен подход к оценке объектов недвижимости с использованием нечеткой линейной регрессии. По итогам выполненного обзора литературы можно сделать вывод, что с использованием нечеткой линейной регрессии удается получить правдоподобный прогноз, а сам метод нечеткой линейной регрессии дает гибкий инструмент, который позволяет сочетать формальные расчеты с экспертными оценками. Так, экспертные оценки могут быть введены в список ограничений соответствующей задачи линейного программирования и тем самым учтены в окончательном результате. В практической части работы рассматривалась, согласно подходу из работы, реализация алгоритма нечеткой регрессии для изучения зависимости (с нечеткими коэффициентами) стоимости квадратного метра в частном доме и стоимости квартиры от ряда факторов.