УДК 519.25`

МЕТОДОЛОГИЧЕСКИЕ АСПЕКТЫ ПРАКТИЧЕСКОГО РЕГРЕССИОННОГО ОЦЕНИВАНИЯ

Е.К. Корноушенко

Институт проблем управления им. В.А. Трапезникова РАН, г. Москва

Предложен подход, ориентированный на случай неоднородных выборок, когда исходная выборка наблюдений (объектов) «разваливается» на классы, отличающиеся существенно разными значениями зависимой переменной (но не регрессоров) и достаточно представительные для построения в каждом из них какой-либо регрессионной модели. На практическом примере показано, как предложенный подход позволяет улучшить качество оценивания по сравнению с традиционным регрессионным подходом.

ВВЕДЕНИЕ

Широкое применение регрессионного оценивания в практических задачах обусловлено чисто экономическими причинами: при наличии больших совокупностей оцениваемых объектов дешевле потратить средства на оценку объектов, образующих обучающую выборку, и построение регрессионных моделей, которые далее используются для оценки остальных объектов, нежели оценивать индивидуально каждый объект исходной совокупности. На практике такой подход получил название *массовой оценки*. Благоприятным полем для широкого применения методов массовой оценки оказались разнообразные объекты недвижимости (квартиры, дома, земельные участки и др.). Различные аспекты применения регрессионных моделей для оценки объектов недвижимости подробно освещены в книге Дж. Эккерта [1]. Американской Ассоциацией налоговых оценщиков (A*I*AA) выпущен ряд стандартов по различным вопросам оценки объектов недвижимости, в частности, стандарт по построению регрессионных моделей для массовой оценки объектов недвижимости [2]. Весьма показательно в этом плане решение Правительства Москвы о проведении в ближайшие годы массовой оценки всех объектов недвижимости г. Москвы в рамках перехода на новую систему налогообложения.

В настоящее время имеется ряд статистических и вычислительных пакетов, содержащих компьютерные программы построения регрессионных моделей различного вида (упомянем лишь общеизвестные пакеты STATISTICA, SPSS, NCSS, MATLAB). Однако для практического применении регрессионного подхода к массовой оценке тех или иных объектов оказывается полезным ряд приемов, не отраженных в этих пакетах и являющихся следствием накопления практического опыта построения и применения регрессионных моделей. Цель настоящей статьи состоит в обсуждении этих приемов и целесообразности их применения в практических задачах массовой оценки.

Предлагаемый в работе подход к регрессионному оцениванию ориентирован на работу с неоднородными выборками, когда вариационный ряд по значениям зависимой переменной, построенный по исходной выборке наблюдений, имеет существенно нелинейный характер. В таких случаях исходную выборку наблюдений можно представить в виде объединения нескольких классов, отличающихся существенно разными значениями зависимой переменной (но не регрессоров) и достаточно представительных для построения в каждом из них какой-либо регрессионной модели. Процедура оценивания в этом случае включает в себя этап классификации (отнесения оцениваемого объекта к тому или иному классу выборки) и этап собственно оценивания с использованием соответствующей регрессионной модели для этого класса.

С методологической точки зрения такой подход к оцениванию можно рассматривать как одну из конкретных реализаций подхода с применением композиций алгоритмов [3]. «Сочленение» построенных моделей для классов в единую модель названо составной моделью. На примере массовой оценки земельных участков промышленного на-

значения Подмосковья показано, что использование составных моделей позволяет улучшить качество оценивания по сравнению с традиционным регрессионным подходом.

1. ИСХОДНЫЕ ДОПУЩЕНИЯ

Будем считать, что предметом рассмотрения является некоторый класс объектов, описываемых одной и той же совокупностью признаков с разными значениями этих признаков для разных объектов. Под признаками понимаются характеристики (количественные или качественные) объектов (это могут быть описания производимых или продаваемых товаров, характеристики объектов недвижимости и т. п.). Совокупность признаков рассматривается как совокупность «входных» переменных (регрессоров, предикторов) объектов. Для объектов рассматриваемого класса задается «выходная» (зависимая) количественная переменная, в оценке возможных значений которой состоит цель процедуры оценивания. В роли зависимой переменной может выступать прибыль от производства или продажи той или иной партии товара, стоимость объектов недвижимости и т. п. Пары «совокупность известных значений регрессоров — соответствующее значение зависимой переменной» называются *наблюдениями*. Исходной информацией служит некоторая совокупность наблюдений, называемая *выборкой*.

Сложность задачи массовой оценки объектов состоит в том, что:

— функциональная зависимость зависимой переменной от наблюдаемых признаков (регрессоров) аналитику неизвестна;

— при оценке объектов из различных областей некоторого ареала внутри рассматриваемого класса эта зависимость может быть существенно различной по характеру;

— характер случайных добавок к значениям зависимой переменной, обусловленных влиянием неизмеримых признаков, также неизвестен.

В такой ситуации при массовой оценке объектов аналитику необходимо последовательно пройти ряд этапов, связанных со сбором и предварительной обработкой исходной информации об оцениваемых объектах, формированием обучающей и контрольной выборок, выбором типа модели и ее калибровкой, оценкой качества построенной модели.

Ряд этапов из этой последовательности типичны для построения регрессионных моделей и в комментариях не нуждаются. Как уже было сказано, предмет нашего внимания — практические приемы, применяемые на некоторых этапах для повышения качества строящихся моделей оценки.

2. ОБЕСПЕЧЕНИЕ ОДНОНАПРАВЛЕННОСТИ «В СРЕДНЕМ» ВЛИЯНИЙ РЕГРЕССОРОВ НА ЗАВИСИМУЮ ПЕРЕМЕННУЮ НА ОБУЧАЮЩЕЙ ВЫБОРКЕ

Будем считать, что исходная выборка подверглась проверке на полноту и репрезентативность входящих в нее наблюдений и предварительному статистическому анализу и из нее выделены обучающая выборка (на объектах которой строится модель) и контрольная выборка, на объектах которой проверяется качество оценки построенной модели. Один из первых описываемых далее практических приемов заключается в обеспечении *однонаправленности «в среднем» влияний* регрессоров на зависимую переменную на обучающей выборке. Обеспечение такой однонаправленности влияний регрессоров на зависимую переменную уменьшает взаимную компенсацию таких влияний, что приводит к увеличению чувствительности зависимой переменной к совокупным изменениям значений регрессоров и, соответственно, к улучшению качества модели.

Решение этой задачи начинается с рассмотрения качественных регрессоров, значения которых определены либо в числовом виде как некоторые служебные коды (например, тип водоснабжения, тип фундамента и т. п.), либо словесно (например, тип стен — панельные, кирпичные и др.). Процедуру оцифровки значений качественных регрессоров, при которой обеспечивается однонаправленность «в среднем» влияний регрессоров на зависимую переменную, называют процедурой *приписывания меток*. У американских оценщиков подобная (в принципе, существенно нелинейная, как показано далее) процедура получила название *процедуры линеаризации*.

Суть процедуры приписывания меток состоит в том, что для каждого значения (числового или словесного) выбранного качественного регрессора формируется совокупность значений зависимой переменной, соответствующих тем объектам обучающей выборки, у которых этот регрессор принимает рассматриваемое значение. Для каждой такой совокупности значений зависимой переменной находится среднее значение, которое затем нормируется путем деления на медианное (или среднее) значение множества этих средних. Результирующие значения («метки») приписываются соответствующим значениям рассматриваемого качественного регрессора. Из того факта, что большему значению среднего для некоторой совокупности значений зависимой переменной соответствует большее значение метки, следует положительность парных коэффициентов корреляции «помеченных» качественных регрессоров с зависимой переменной (отсюда и термин — однонаправленность «в среднем»).

Для количественных регрессоров, имеющих положительные коэффициенты корреляции зависимой переменной, значения количественного регрессора делятся на медиану (среднее) его значений в обучающей выборке, и в качестве меток рассматриваются полученные значения этого регрессора.

Для количественных регрессоров, имеющих отрицательные коэффициенты корреляции с зависимой переменной, исходные значения X_{ks} регрессора X_k переопределяются каким-либо образом для обеспечения положительности этих коэффициентов. В частности, следующим образом:

$$X^*_{ks} = \max(X_k) - X_{ks} + C, \qquad (1)$$

где $\max(X_k)$ — максимальное значение регрессора X_k в обучающей выборке, а положительная константа C добавляется, чтобы избежать нулевого значения X^*_{ks}, что недопустимо при построении мультипликативной модели. Далее полученные значения регрессора X^*_k нормируются аналогично предыдущему. Нормированные значения рассматриваются как метки соответствующих исходных значений этого регрессора.

Нормировка «помеченных» качественных, а также количественных регрессоров делается для того, чтобы обеспечить сравнимые диапазоны значений качественных и количественных регрессоров перед построением модели.

Для двоичных регрессоров может оказаться целесообразной следующая процедура приписывания меток.

1. Изначально состояния 0,1 приписываются этим регрессорам таким образом, чтобы коэффициент корреляции каждого двоичного регрессора с зависимой переменной сделать положительным.

2. Двоичные регрессоры с приписанными значениями 0,1 упорядочиваются по возрастанию их коэффициентов корреляции с зависимой переменной.

3. Первый регрессор в этом упорядочении рассматривается как двоичный разряд при 2^0, второй регрессор — как двоичный разряд при 2^1, ..., k-й регрессор — как двоичный разряд при 2^k. Таким образом, совокупности k двоичных регрессоров ставится в соответствие так называемый разрядный регрессор с множеством 2^{k+1} значений. Поскольку в исходной выборке могут присутствовать не все комбинации значений двоичных факторов, «реальное» число значений разрядного регрессора может быть меньше числа 2^{k+1}. Так как наименьшим «реальным» значением этого регрессора может быть 0 (что недопустимо при использовании мультипликативных моделей), то ко всем значениям разрядного регрессора добавляется 1.

4. Разрядный регрессор рассматривается как качественный регрессор, и его значениям приписываются метки в соответствии с процедурой приписывания меток значениям качественных регрессоров. ♦

Смысл введения разрядного регрессора в том, что его значения соответствуют существующим в исходной выборке *комбинациям значений* двоичных регрессоров и именно эти комбинации (а не сами двоичные регрессоры) оказывают существенное влияние на зависимую переменную.

Регрессионная модель выбранного типа строится не на исходных значениях регрессоров, а на метках, приписанных этим значениям. В данной работе рассматриваются простейшие регрессионные модели — линейные и приводимые к линейным, параметры которых вычисляются с помощью обычного метода наименьших квадратов (МНК). Поскольку приписывание меток — существенно нелинейная процедура, она деформирует минимизируемую функцию таким образом, что получаемые МНК-решения определяют модели с улучшенными показатели качества.

3. ОЦЕНКА ОБЪЕКТОВ В УСЛОВИЯХ НЕОДНОРОДНЫХ («ПЛОХИХ») ВЫБОРОК

3.1. Понятие неоднородной выборки

Основная проблема, с которой сталкивается аналитик при построении той или иной модели при массовой оценке объектов, скажем, экономического характера в условиях слабо развитого рынка, заключается в неоднородности выборок, содержащих рыночные данные об объектах, причем эта неоднородность сохраняется после удаления из выборок выбросов и неполных описаний объектов. Нестрого говоря, выборка считается *неоднородной*, если в ней много объектов с одними значениями регрессоров и (или) зависимой переменной и мало объектов с другими значениями. Важно подчеркнуть, что подобной неоднородностью может обладать и генеральная совокупность объектов, обращающихся на данном слабо развитом рынке.

Анализ неоднородной выборки удобно начинать с построения для нее вариационного ряда по возрастанию (или убыванию) зависимой переменной. Для неоднородных выборок график вариационного ряда носит существенно нелинейный характер. Для простоты рассмотрим случай, когда на таком графике явно выражен участок с малым наклоном (пологий участок), участок с большим наклоном (крутой участок). Чтобы избежать резкого скачка в получающихся оценках при переходе от пологого участка к крутому, будем считать, что эти участки пересекаются на некоторой «зоне перекрытия».

В силу существенной нелинейности графика значений зависимой переменной применяемый при построении традиционных регрессионных моделей обычный МНК «сталкивается» с проблемой *одновременной минимизации* квадратов невязок на пологом и на крутом участках графика. При этом результирующее МНК-решение задачи минимизации, оптимальное для всего графика в целом, не оптимально для каждого из этих участков в отдельности. Этот факт послужил отправным моментом для рассматриваемого далее подхода к построению так называемых *составных* моделей оценки, при построении которых производится раздельная минимизация сумм квадратов невязок на пологом и на крутом участках вариационного ряда значений зависимой переменной.

Замечание 1. В принципе, график вариационного ряда может содержать несколько участков с разными наклонами и, соответственно, несколько зон перекрытия. Возможность анализа всех таких участков с целью построения моделей оценки на каждом из участков определяется представительностью (репрезентативностью) соответствующей каждому участку группы рассматриваемых объектов (т. е. возможностью построения на каждом участке статистически значимой модели оценки), что далеко не всегда можно сделать, особенно для слабо развитых рынков с небогатой рыночной информацией.♦

3.2. Процедура построения составной модели по неоднородной обучающей выборке

Процедура построения составной модели по неоднородной обучающей выборке включает в себя следующие этапы (для простоты рассматриваем два участка).

1. Построение по этой выборке вариационного ряда значений зависимой переменной и выделение на графике вариационного ряда пологого и крутого участков.

2. Переупорядочение объектов обучающей выборки в соответствии со значениями зависимой переменной этих объектов в вариационном ряду.

3. Независимое приписывание меток значениям регрессоров объектов пологого и крутого участков. Это означает, что для регрессоров объектов, принадлежащих пологому (крутому) участку, приписывание меток производится с использованием выборки, составленной из объектов, принадлежащих лишь *рассматриваемому участку*. При этом значениям регрессоров объектов из зоны перекрытия будут приписаны две метки.

4. Построение традиционной регрессионной модели (моделей) для каждого из участков. Обозначим через МП модель оценки, построенную на пологом участке, а через МК — модель оценки, построенную на крутом участке. Составная модель определяется как «сочленение» моделей-компонентов МП и МК, реализуемое по следующему правилу. Пусть i — текущий номер объекта в исходном вариационном ряду, а МП(i) (или МК(i)) — модельная оценка зависимой переменной i-го объекта, полученная с использованием модели МП (или МК), а МС(i) — соответствующая модельная оценка с использованием составной модели. Тогда:

— если i-й объект принадлежит пологому участку, то МС(i) = МП(i);

— если i-й объект принадлежит зоне перекрытия участков, то в простейшем случае можно считать, что МС(i) = (МП(i) + МК(i))/2. В принципе, возможны и другие правила «сочленения»;

— если i-й объект принадлежит крутому участку, то МС(i) = МК(i).♦

3.3. Оценка объектов контрольной выборки

При оценке объектов контрольной выборки с помощью составной модели необходимо разрешить две проблемы:

— проблему приписывания меток значениям регрессоров объектов контрольной выборки (поскольку модели, входящие в составную модель, определены на метках значений регрессоров);

— проблему выбора модели (МП или МК, или обеих) для оценки рассматриваемого объекта контрольной выборки. Эта проблема решается путем соотнесения каждому из объектов контрольной выборки некоторого объекта из обучающей выборки, положение которого в упорядоченной совокупности объектов согласно вариационному ряду сразу указывает, какая модель должна быть выбрана.♦

3.3.1. Приписывание меток значениям регрессоров объектов контрольной выборки

Замечание 2. Для корректности приведенной далее процедуры приписывания меток необходимо выполнение для объектов контрольной выборки следующих условий:

— каждое значение (числовое или словесное) всякого качественного регрессора должно присутствовать у каких-либо объектов обучающей выборки;

— каждое значение всякого количественного регрессора должно принадлежать диапазону значений соответствующего количественного регрессора в обучающей выборке.♦

Качественные регрессоры

Для каждого значения качественного регрессора объектов контрольной выборки находится совпадающее с ним исходное значение соответствующего регрессора из объектов обучающей выборки, и метка этого значения переносится на рассматриваемое значение качественного регрессора. Если такое совпадающее значение принадлежит и пологому, и крутому участкам (так что оно имеет две метки), то соответствующему значению регрессора

в контрольной выборке приписываются обе метки. При невыполнении требований замечания 2 для рассматриваемого значения качественного регрессора необходимо привлечение дополнительной информации для обоснования ближайшего к нему исходного значения соответствующего регрессора из объектов обучающей выборки, и метка (метки) этого ближайшего значения переносится на рассматриваемое значение аналогично предыдущему.

Количественные регрессоры

Если данный количественный регрессор без приписанных меток положительно коррелирует с зависимой переменной на обучающей выборке, то для каждого значения этого регрессора у объектов контрольной выборки находится ближайшее к нему исходное значение соответствующего регрессора из объектов обучающей выборки, и метка (метки) этого значения переносится на рассматриваемое значение количественного регрессора.

Если данный количественный регрессор без приписанных меток отрицательно коррелирует с зависимой переменной на обучающей выборке, то исходные значения данного регрессора на контрольной выборке переопределяются по аналогии с формулой (1), причем значения $\max(X_k)$ берется то же, что и для объектов обучающей выборки. Переопределенным значениям этого регрессора метки приписываются так, как указано в предыдущем пункте.

3.3.2. Соотнесение объектам контрольной выборки объектов из обучающей выборки

Ключевой момент при использовании составных моделей состоит в выборе для каждого объекта контрольной выборки модели МП или МК, с помощью которой будет оцениваться зависимая переменная этого объекта. Поскольку при этом каждый объект контрольной выборки необходимо отнести либо к пологому, либо к крутому участку вариационного ряда, то данная задача представляет собой, фактически, задачу классификации (для корректности постановки задачи классификации зона перекрытия относится к пологому участку).

Очевидным решением представляется нахождение для каждого объекта контрольной выборки ближайшего (по исходным значениям факторов или по меткам) объекта из обучающей выборки, т. е. применение в качестве алгоритма классификации правила наименьшего расстояния между объектами (см. например, работу [4]). В данном случае такой алгоритм дает плохие результаты, поскольку все регрессоры здесь выступают «на равных», в то время как значения качественных регрессоров и двоичных регрессоров, присутствующие у объектов и пологого, и крутого участков, существенно ухудшают результаты такой классификации. По этой же причине и логистическая модель [5] может не обеспечить приемлемое качество классификации. Именно этот факт указывает на то, что при разработке алгоритма классификации необходимо использовать то свойство, что разные значения регрессоров в упорядоченной совокупности объектов вариационного ряда ведут себя по-разному на пологом и крутом участках. В формальном плане понятие «по-разному» означает несимметричность распределения значений рассматриваемого регрессора на объектах пологого и крутого участков вариационного ряда. Для количественной оценки такой несимметричности можно воспользоваться соотношениями между числами вхождений того или иного значения рассматриваемого регрессора в описания объектов пологого и крутого участков. Чем больше разность между этими числами для некоторого значения рассматриваемого регрессора, тем больше *степень информативности* этого значения.

При этом процедура соотнесения объектам контрольной выборки объектов из обучающей выборки выглядит следующим образом.

1. Выбирается первый объект из контрольной выборки. Значения всех его регрессоров упорядочиваются по убыванию их степени информативности (исходное упорядочение для выбранного объекта).

2. Выбирается первый регрессор в этом упорядочении и формируется группа объектов из обучающей выборки с тем же (или ближайшим) значением этого регрессора, что и у рассматриваемого объекта.

3. Из этой группы выделяется подгруппа объектов с тем же (или ближайшим) значением регрессора, следующего в исходном упорядочении.

4. Процесс выделения следующей подгруппы из предыдущей с учетом значения очередного в исходном упорядочении регрессора продолжается до выполнения какого-либо из следующих условий:

— следующая подгруппа содержит единственный объект из обучающей выборки, принадлежность которого к какому-либо участку (или к зоне перекрытия) известна априори в силу единственности отнесения объектов вариационного ряда к тому или иному участку;

— завершен перебор всех факторов стоимости в исходном упорядочении. Если при этом последняя рассматриваемая подгруппа содержит объекты, относящиеся к пологому участку, и объекты, относящиеся к крутому участку, то процесс классификации считается незавершенным, а исходная совокупность регрессоров — недостаточной для требуемой классификации. В этом случае требуется привлечение дополнительных признаков (регрессоров) для описания объектов исходной выборки.

При выполнении первого из этих условий — переход к следующему объекту контрольной выборки и повторение указанной процедуры. ♦

Формальные аспекты этой процедуры требуют отдельного рассмотрения.

3.3.3. Выбор моделей для оценки объектов контрольной выборки

При завершении процесса классификации каждому объекту контрольной выборки ставится в соответствие единственный номер соответствующего объекта обучающей выборки. Обозначим через $\Phi(i)$ такой номер для i-го объекта контрольной выборки. По аналогии с правилом выбора моделей для объектов обучающей выборки определяется правило выбора моделей для объектов контрольной выборки:

— если объект с номером $\Phi(i)$ принадлежит пологому участку, то $МС(i) = МП(i)$;

— если объект с номером $\Phi(i)$ принадлежит зоне перекрытия участков, то в простейшем случае можно считать, что $МС(i) = (МП(i) + МК(i))/2$. В принципе, возможны и другие правила «сочленения»;

— если объект с номером $\Phi(i)$ принадлежит крутому участку, то $МС(i) = МК(i)$.

Напомним, что в ту или иную выбираемую модель подставляются метки, приписанные соответствующим значениям регрессоров i-го объекта контрольной выборки. Как показано далее на конкретном примере, использование составных моделей для оценки объектов неоднородных выборок позволяет существенно улучшить показатели качества оценки.

4. ПРИМЕР. МАССОВАЯ ОЦЕНКА ЗЕМЕЛЬНЫХ УЧАСТКОВ ПРОМЫШЛЕННОГО НАЗНАЧЕНИЯ ПОДМОСКОВЬЯ

4.1. Описание исходной задачи

При оценке объектов недвижимости, к которым относятся земельные участки разного функционального назначения, в роли зависимой переменной выступает тот или иной стоимостный показатель (полная цена, удельная цена, цена аренды и т. п.), а в роли регрессоров, которые в случае объектов недвижимости принято называть *факторами стоимости*, — различные характеристики объектов недвижимости (факторы местоположения, факторы собственно объекта недвижимости, факторы его непосредственного окружения и т. д.). Исходной информацией о земельных участках в данном примере является выборка, содержащая 136 участков из 30-ти районов Подмосковья. Описания этих участков в исходной выборке содержали следующие факторы стоимости.

Факторы местоположения

Оценочная зона — фактор *Зона*. Согласно постановлению Правительства Московской области, принятому в 1995 г., вся территория Московской области разделена на 6 оценочных зон. Фактор *Зона* — качественный, имеющий 6 числовых значений.

Административный район Московской области — фактор *Район*. Присутствующие в исходной выборке земельные участки принадлежали к 30-ти административным районам Подмосковья. Фактор *Район* — качественный, имеющий 30 словесных значений.

Расстояние от МКАД — Московской кольцевой автомобильной дороги — фактор *МКАД*, является следующим фактором, уточняющим местоположение каждого участка. Фактор *МКАД* — количественный, диапазон его значений в исходной выборке от 0,2 до 101 км.

Факторы, описывающие собственно земельный участок

Площадь земельного участка — фактор *ПлУч*, — количественный, диапазон его значений в исходной выборке от 1,5 до 500 тыс. кв. м.

Наличие коммуникаций (электроэнергия + + связь) на участке (есть—нет) — фактор *Комм*.

Наличие железнодорожных подъездных путей к участку (есть—нет) — фактор *Подъезд*;

Находится ли участок на территории населенного пункта поселкового типа с числом жителей, большим 1 тыс. чел. (да—нет) — фактор *НП*.

Исходная выборка содержала также данные о полных рыночных стоимостях земельных участков от 56 до 11 250 тыс. долл. Далее для краткости вместо слова «земельный участок» будем использовать слово «объект».

4.2. Процедура построения составной модели по обучающей выборке

Процедура построения составной модели содержала следующие этапы.

1. *Формирование обучающей и контрольной выборок*. Для исходной выборки был построен вариационный ряд по возрастанию полной стоимости объектов. Объекты исходной выборки переупорядочивались в соответствии с положением значения стоимости каждого объекта в вариационном ряду. Обучающая и контрольная выборки формировались путем поочередного отнесения объектов этой упорядоченной совокупности то к одной, то к другой выборке, так что обучающая выборка содержала объекты с четными номерами, а контрольная — с нечетными. Такое поочередное отнесение объектов к обучающей и контрольной выборке позволяет в значительной степени обеспечить выполнение требований замечания 2, § 3. В результате обучающая и контрольная выборки содержали по 68 объектов.

2. *Выделение пологого и крутого участков на графике вариационного ряда для объектов обучающей выборки*. Для объектов обучающей выборки был построен вариационный ряд по стоимости объектов и определены параметры пологого и крутого участков (рис. 1): пологий участок *1* включал в себя объекты с номерами от 1 по 50, крутой учас-

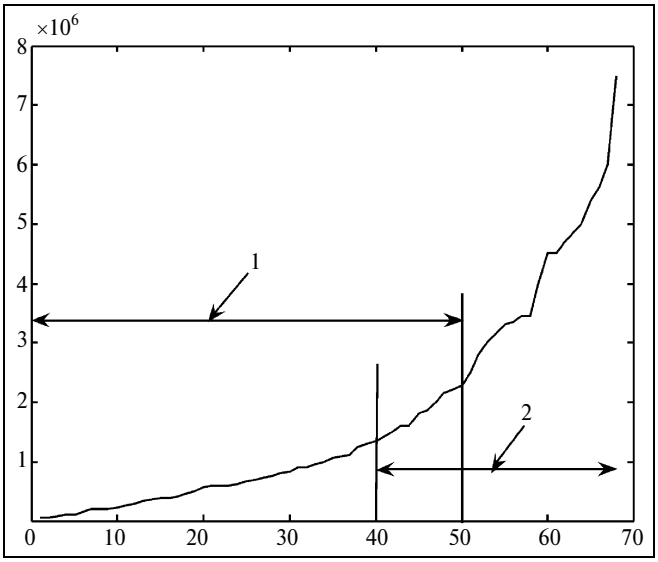

Рис. 1. Выделение в вариационном ряду для объектов обучающей выборки пологого и крутого участков
(по оси абсцисс указаны номера объектов в упорядоченной совокупности, по оси ординат — стоимость объектов, долл.)

ток *2* — с номерами от 40 по 68, таким образом, зона перекрытия содержала 11 объектов.

3. *Приписывание меток значениям факторов стоимости на пологом и крутом участках.* На пологом и крутом участках обучающей выборки *независимо* приписывались метки значениям факторов стоимости объектов согласно приведенным выше правилам. Значениям качественных факторов *Район* и *Зона* метки приписывались согласно процедуре приписывания меток значениям качественных факторов, для фактора *МКАД*, отрицательно коррелирующего со стоимостью объектов, проводилось переопределение его значений согласно формуле (1), а на двоичных факторах *НП*, *Комм* и *Подъезд* был построен разрядный фактор 1 + *Комм* + 2 **НП* + 4 **Подъезд*, для чего исходные значения (0, 1) факторов *Комм* и *Подъезд* переопределялись для обеспечения положительности корреляции этих факторов со стоимостью объектов.

4. *Построение регрессионных моделей на пологом и крутом участках.* На пологом и крутом участках в качестве моделей оценки были выбраны и построены линейная и мультипликативная модели. Кроме того, были построены линейная и мультипликативная модели по всей обучающей выборке для последующего сравнения качества оценки строящейся составной модели с этими моделями.

5. *Сравнение качества оценки составной модели с традиционными моделями на обучающей выборке.* Для простоты полагаем, что в составной модели «сочленяются» модели одного и того же класса, результирующую составную модель назовем либо линейной, либо мультипликативной составной моделью. Значения параметров качества составных моделей на обучающей выборке приведены в табл. 1. Для сравнения в ней приведены также значения параметров качества для линейной и мультипликативной моделей, построенных по всей обучающей выборке.

Здесь $\delta_{ср}$ — средняя относительная погрешность оценки, R^2 — множественный коэффициент корреляции (коэффициент детерминации), F — значение критерия Фишера, а σ — стандартное отклонение. Поскольку при уровне значимости 0,05, числе факторов в моделях, равном 5, и длине обучающей выборки, равной 68, пороговое значение критерия Фишера $F_{kp0,05;5,68-5-1} \approx 2$, все рассмотренные модели статистически значимы. Из табл. 1 следует, что составные модели на обучающей выборке обладают лучшим качеством оценки по сравнению с «одиночными» рассмотренными моделями.

4.3. Процедура оценки объектов контрольной выборки с использованием составной модели

Оценка объектов контрольной выборки с использованием составной модели производится согласно следующим этапам.

А. *Приписывание пар меток значениям факторов стоимости.* Прежде всего на значения факторов стоимости объектов контрольной выборки надо перенести метки совпадающих с ними (или ближайших к ним) значений факторов стоимости объектов пологого и крутого участков обучающей выборки. В результате каждое значение факторов стоимости объектов контрольной выборки будет иметь пару меток.

В. *Соотнесение объектам контрольной выборки соответствующих объектов обучающей выборки.* Такое соотнесение производится согласно разработанному алгоритму классификации, суть которого сформулирована в первой части работы. Результат действия алгоритма заключается в совокупности номеров объектов обучающей выборки в исходном вариационном ряду, однозначно определенных для объектов контрольной выборки. Эта совокупность в данном случае имеет следующий

Таблица 1

Значения параметров качества составных и традиционныхмоделей на обучающей выборке

Составная модель		«Одиночная» модель	
линейная	мультипликативная	линейная	мультипликативная
$\delta_{ср}$ = 50,88 % R^2 = 0,9570 F = 275,9801 σ = 402 080	$\delta_{ср}$ = 37,48 % R^2 = 0,9270 F = 157,4723 σ = 454 920	$\delta_{ср}$ = 96,08 % R^2 = 0,7505 F = 36,1016 σ = 796 850	$\delta_{ср}$ = 33,70 % R^2 = 0,7751 F = 41,3488 σ = 756 650

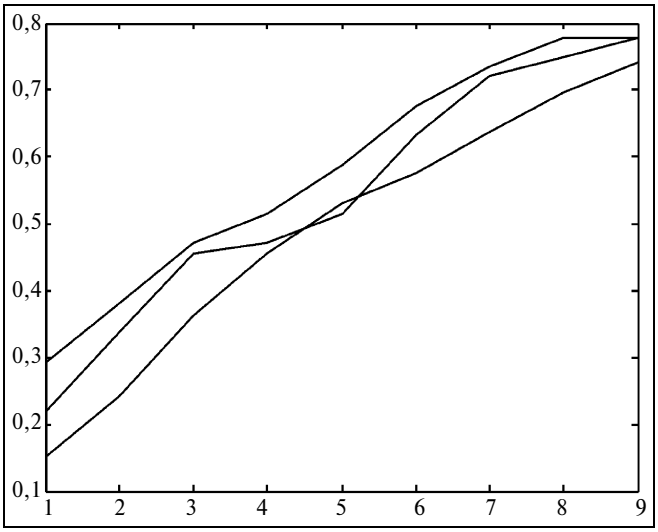

Рис. 2. Графики доли объектов, для которых относительная погрешность оценки не превосходит заданного уровня: «одиночная» мультипликативная модель (нижняя кривая); составная мультипликативная модель с использованием логистической модели (промежуточная кривая); составная мультипликативная модель с использованием предложенного алгоритма классификации (верхняя кривая)

вид: Ф = (16 35 16 34 1 40 11 40 1 8 12 14 13 16 29 8 24 17 63 40 21 23 23 13 35 17 23 9 30 3 38 28 40 9 34 67 30 44 24 42 42 52 44 35 24 64 20 18 20 54 66 57 55 45 31 45 57 66 8 49 66 62 54 59 65 65 26 30).

С. *Выбор моделей для оценки объектов контрольной выборки* производится по правилу:

— если номер $\Phi(i) < 41$, то оценкой стоимости i-го объекта является МП(i);

— если $\Phi(i) \in [41, 50]$, то оценка стоимости i-го объекта есть (МП(i) + МК(i))/2;

— если номер $\Phi(i) > 50$, то оценкой стоимости i-го объекта является МК(i).

Значения параметров качества составных моделей на контрольной выборке приведены в табл. 2. Для сравнения в ней приведены также значения параметров качества линейной и мультипликативной моделей (построенных по всей обучающей выборке) на контрольной выборке.

Таблица 2

Значения параметров качества составных моделей и традиционных моделей на контрольной выборке

Составная модель		«Одиночная» модель	
линейная	мультипликативная	линейная	мультипликативная
$\delta_{ср} = 83{,}88$ %	$\delta_{ср} = 69{,}12$ %	$\delta_{ср} = 136{,}26$ %	$\delta_{ср} = 82{,}86$ %
$R^2 = 0{,}6153$	$R^2 = 0{,}6731$	$R^2 = 0{,}3705$	$R^2 = 0{,}2831$
$F = 19{,}1959$	$F = 24{,}7102$	$F = 7{,}0613$	$F = 4{,}7387$
σ=1 005 600	σ = 926 990	σ=1 292 200	σ=1 378 900

Из табл. 2 следует, что на контрольной выборке качество линейной и мультипликативной составных моделей существенно лучше качества линейной и мультипликативной моделей, построенных по всей обучающей выборке. Об этом же говорит и рис. 2, на котором приведены графики (аналоги линии Лоренца) для различных моделей.

На рис. 2 по оси абсцисс отложены заданные уровни относительной погрешности оценки (в десятках процентов), а по оси ординат — доли объектов контрольной выборки, относительная погрешность оценки которых не превосходит заданного уровня. График для мультипликативной составной модели с применением предложенного алгоритма доминирует над остальными графиками, что говорит о лучшем качестве оценки объектов с помощью этой модели, а площадь зоны между графиками является количественной оценкой такого улучшения.

ЗАКЛЮЧЕНИЕ

Мотивация данной работы обусловлена желанием дополнить традиционные методы построения регрессионных моделей некоторыми практическими приемами, улучшающими качество этих моделей. К таким приемам относятся рассмотренные в работе процедуры приписывания меток и объединения традиционных моделей в составные модели в случае неоднородных выборок. Центральным в процедуре оценивания с использованием составных моделей является алгоритм классификации, от качества которого зависит результирующее качество оценивания. На конкретном примере продемонстрирована эффективность перехода к составным моделям в плане повышения качества оценки.

ЛИТЕРАТУРА

1. *Эккерт Дж.* Организация оценки и налогообложения недвижимости. — М.: Стар Интер, 1997. — Т. 2. — 442 с.
2. *Standard* on Automated Valuation Models (AVMs). International Association of Assessing Officers, 2003.
3. *Воронцов К.В.* Лекции по алгоритмическим композициям. — М.: МГУ, 2007. — 44 с. ⟨http://lib.mexmat.ru/books/13920⟩.
4. *Воронцов К.В.* Лекции по метрическим алгоритмам классификации. — М.: МФТИ, 2007. — 14 с. ⟨http://www.ccas.ru/voron/download/MetricAlgs⟩.
5. *Логистическая* регрессия и ROC-анализ — математический аппарат. Лаборатория Base Group, ⟨http://www.basegroup.ru/regressi on/logistic.htm⟩.

☎ *(495) 334-90-00, e-mail: ekorno@mail.ru*

Статья представлена к публикации членом редколлегии А.С. Рыковым. ❑