УДК 519.25

МЕТОДОЛОГИЧЕСКИЕ АСПЕКТЫ ПРАКТИЧЕСКОГО РЕГРЕССИОННОГО ОЦЕНИВАНИЯ: УЛУЧШЕНИЕ ОЦЕНОЧНЫХ СВОЙСТВ МОДЕЛЕЙ ПУТЕМ КОДИРОВАНИЯ ЗНАЧЕНИЙ КОЛИЧЕСТВЕННЫХ ПРИЗНАКОВ

Е.К. Корноушенко, А.А. Лобко

Предложены эвристические процедуры решения двух важных проблем, возникающих при практическом регрессионном оценивании. Одна из них связана с выбором нелинейных преобразований для количественных данных. Для ее решения предложен метод нелинейного преобразования значений количественных данных с использованием понятия близости значений, что позволило существенно повысить качество моделей. Другая проблема связана с построением моделей с улучшенными прогнозными свойствами, проверяемыми на тестовой выборке. Предложено несколько эвристических процедур, с помощью которых производится последовательный отбор регрессоров на обучающей и контрольной выборках для построения модели с требуемыми свойствами. Эффективность этих процедур показана на практическом примере оценки стоимости земельных участков в Калужской области.

Ключевые слова: выборка, регрессионная модель, метки, прогнозные свойства модели.

ВВЕДЕНИЕ

В работе [1] рассматривался ряд этапов практической процедуры построения регрессионных моделей с количественной зависимой переменной, применяемых, в частности, для оценки объектов недвижимости. Подчеркивалась важность использования процедуры приписывания меток как для качественных, так и для количественных[1] признаков. В вычислительном плане приписывание меток представляет собой нелинейные преобразование исходных (числовых, вербальных) значений качественных и количественных признаков, улучшающее качество строящейся регрессионной модели.

В данной работе предложена процедура приписывания меток значениям количественных признаков, более эффективная по сравнению с аналогичной процедурой, приведенной в работе [1]. Важность предложенной процедуры обусловлена следующими причинами.

Из практики оценки объектов недвижимости известно, что местоположение оцениваемого объекта определяет около половины его рыночной стоимости. Зарубежные методики оценки объектов недвижимости принципиально ориентированы на применение электронных карт (см., например, работы [2, 3]), определяющих положение объекта парой координат (x, y). В России электронные карты территорий доступны лишь для нескольких центральных регионов. В поступающей рыночной информации об объектах недвижимости из остальных регионов[2] местоположение объектов указывается с привлечением большой совокупности расстояний каждого объекта от тех или иных «опорных точек» (от центра города (поселка), от железнодорожного вокзала, от автостанции, от мест рекреа-

[1] Количественный признак отличается от качественного тем, что он может принимать любое значение из выбранного диапазона непрерывных значений, в то время как для качественного признака задается лишь конечное множество значений.

[2] Авторы данной работы имели непосредственное отношение к кадастровой оценке земель.

ции и т. п.). Как правило, корреляции этих расстояний со стоимостью объектов весьма незначительны, поэтому использование исходных (или нормированных) значений этих расстояний практически безрезультатно. Выход из этой ситуации состоит в таком нелинейном преобразовании расстояний (и, быть может, других количественных признаков), которое увеличит корреляции преобразованных признаков с зависимой переменной, т. е. усилит их влияние на стоимость объектов[3]. Одним из таких преобразований является рассматриваемое в статье преобразование с использованием понятия близости значений.

Другая важная проблема состоит в построении по исходной выборке модели с улучшенным качеством оценки объектов, не входящих в исходную выборку, т. е. находящихся вне выборки (*out-of-sample*). Построению таких моделей уделяется большое внимание в последнее время. Помимо классического статистического подхода, когда элементы выборки рассматриваются как случайные переменные с тем или иным распределением, применяются различные методы формирования по исходной выборке обучающей и контрольной выборок (*cross-validation*) [4], различные методы бутстрэпа (*bootstrap*) для случаев фиксированной и нефиксированной обучающей выборки [5] и т. д. В данной работе для построения моделей с улучшенными прогнозными свойствами предлагаются несколько эвристических процедур, использующих понятие близости для количественных признаков и некоторые вспомогательные понятия.

1. НЕЛИНЕЙНЫЕ ПРЕОБРАЗОВАНИЯ ЭЛЕМЕНТОВ ИСХОДНОЙ ВЫБОРКИ

Будем считать, что исходными данными для анализа является некоторая выборка объектов, характеризующихся одним и тем же набором признаков, принимающих для разных объектов, возможно, разные количественные и качественные значения; предполагается, что зависимая переменная может быть только *количественной*. В практических задачах оценки нередко значения качественных признаков заданы либо словесно, либо в виде ведомственных числовых кодов, что препятствует их последующей компьютерной обработке. Процедура приписывания числовых меток значениям качественных признаков состоит в следующем [1].

1. Для каждого конкретного значения качественного признака из исходной выборки выделяется совокупность объектов с этим значением.

2. Формируется совокупность значений зависимой переменной, соответствующих выделенным объектам, и находится среднее значение зависимой переменной в этой совокупности.

3. Пункты 1 и 2 повторяются для всех значений рассматриваемого признака и формируется множество, элементами которого являются средние значения зависимой переменной соответствующих совокупностей.

4. Каждое из этих средних делится на медиану (или среднее[4]) множества средних из пункта 3. Результат такого деления для каждого среднего и является *меткой*, приписываемой соответствующему значению качественного признака.

Обобщим эту процедуру для количественных признаков. Пусть X — некоторый количественный признак, X_i — какое-либо его значение, относящееся к некоторому объекту из исходной выборки, $S(X)$ — множество значений признака X у объектов исходной выборки.

Определение. Пусть X_i, $X_j \in S(X)$ — два значения признака X, а $0 < d < 1$ — некоторое число. Будем говорить, что значение X_j является *d-близким* к значению X_i, если справедливо

$$|X_i - X_j| \leqslant dX_i. \blacklozenge$$

Обозначим через $S(d, X_i)$ совокупность d-близких значений для значения X_i. Очевидно, что $S(0, X_i) = X_i$. Функция $|S(d, X_i)|$ является неубывающей функцией параметра d.

Процедура приписывания меток количественным признакам отличается от указанной выше процедуры лишь пунктом 1: для каждого выбранного значения количественного признака формируется совокупность тех объектов исходной выборки, у которых значения этого признака d-близки к рассматриваемому значению. При выборе значений «параметра близости» d следует руководствоваться предметными неформальными соображениями о «приемлемых» значениях d для того или другого признака. Для каждого количественного признака в исходной выборке существует такое d, при котором коэффициент корреляции этого помеченного признака с зависимой переменной будет наибольшим (по модулю) из всех возможных значений. Этот факт используется далее в § 3 при оптимизации модели, т. е. при выборе

[3] Увеличение (по модулю) корреляции преобразованного признака с зависимой переменной приводит к увеличению соответствующего этому признаку коэффициента в регрессионной модели, т. е. к увеличению влияния этого признака на зависимую переменную.

[4] Кроме среднего и медианы, возможен выбор смешанных функций от этих величин.

«субоптимального» значения d для каждого из признаков, используемых в модели.

Приписывание меток как предварительное нелинейное преобразование исходных значений признаков рассматриваемой выборки приводит к следующим важным последствиям:

- увеличение коэффициентов корреляции «помеченных» количественных и качественных признаков с зависимой переменной;
- повышение независимости количественных признаков (уход от проблемы мультиколлинеарности): в силу того, что поскольку отношение близости не является транзитивным, корреляция исходных коллинеарных (пропорциональных) признаков существенно уменьшается;
- возможность улучшения качества регрессионной модели путем выбора того или иного значения «параметра близости» для каждого из количественных признаков.

Замечание 1. Имея в виду ограниченные возможности линейных и внутренне линейных регрессионных моделей в плане аппроксимации неизвестных регрессионных зависимостей и стремясь сохранить такое «удобное» свойство этих моделей как линейность по параметрам, аналитики вводят в рассмотрение «сложные» регрессоры как различные функции от исходных регрессоров (произведения и (или) степени регрессоров, отношения регрессоров и т. д.). Процедура подбора таких функций и соответствующее программное обеспечение используются, в частности, в методе RETINA [6]. Описанные выше процедуры приписывания меток значениям качественных и количественных признаков также сохраняют свойство линейности модели по параметрам. ♦

2. ПОСТРОЕНИЕ РЕГРЕССИОННОЙ МОДЕЛИ

2.1. Формирование обучающей, контрольной и проверочной выборок

Процедура построения регрессионной модели по исходной выборке начинается с предварительной обработки выборки (проверки корректности описаний элементов выборки, удаления выбросов и т. д.) и формирования по обработанной выборке обучающей, контрольной и проверочной выборок. Для этого проводится равновероятное случайное разбиение исходной выборки на три выборки. На обучающей выборке после соответствующих процедур строится модель оценки (*estimation*) объектов, контрольная выборка служит для выделения признаков, существенных для оценки объектов на тестовой выборке, и оптимизации результирующей модели, тестовая выборка служит для оценки качества прогнозирования (*validation* [7]) результирующей модели[5].

После формирования трех выборок обучающая выборка дополняется некоторыми объектами из контрольной и проверочной выборок для выполнения следующих условий:

— все значения качественных признаков, относящиеся к объектам контрольной и проверочной выборок, должны присутствовать у объектов обучающей выборки;

— диапазон значений всякого количественного признака у объектов контрольной и проверочной выборок может превышать не более чем на A процентов[6] диапазон значений соответствующего количественного признака у объектов обучающей выборки.

2.2. Приписывание меток значениям признаков на обучающей выборке

Следующий этап заключается в приписывании меток значениям качественных и количественных признаков на обучающей выборке согласно процедурам, приведенным в § 1. Значения «параметра близости» d для количественных признаков следует задавать, исходя из предметных соображений и не стремясь выбирать «оптимальные» значения (так, для количественных признаков земельных участков целесообразно выбирать d в интервале 0,15...0,2).

2.3. Отбор признаков и построение предварительной модели на обучающей выборке

Процедура 1. Отбор признаков производится путем рассмотрения абсолютных значений коэффициентов корреляции помеченных признаков-кандидатов с зависимой переменной (как, к примеру, в методе RETINA [6]). Согласно эмпирическому критерию отбора признаков отбираются признаки с коэффициентами корреляции, составляющими

[5] Подобное разделение функций, реализуемых на каждой из выборок, практикуется при построении моделей с улучшенными прогнозными свойствами, в частности, оно используется в известном методе RETINA [6].

[6] Если данное значение количественного признака у некоторого объекта контрольной выборки выходит из диапазона значений этого признака у объектов обучающей выборки, то в предположении нормальности распределения значений этого признака у объектов обучающей выборки считаем, что величина A не может выходить за пределы 3σ от среднего значения этого распределения; в противном случае данное значение считается выбросом, и объект с этим значением признака удаляется из контрольной выборки. Если же распределение значений не является нормальным, то выбор величины A определяется пользователем из тех или иных предметных соображений.

не менее 30—50 % от наибольшего коэффициента корреляции. ♦

Помимо известной процедуры 1 для отбора признаков предлагаются следующие эвристические процедуры.

Процедура 2. Отбор признаков производится путем вычисления вкладов помеченных признаков в коэффициент детерминации R^2 предварительной линейной модели M_o, построенной по исходно выбранной совокупности признаков. Вклад i-го признака в коэффициент R^2 определяется по формуле [8]:

$$w_{\text{o}}^i = \frac{\text{cov}(D_{\text{o}}(:, i) \cdot a_i, Y_{\text{o}})}{\text{var}(Y_{\text{o}})}, \quad i = 1, ..., N, \qquad (1)$$

где $D_{\text{o}}(:, i)$ — i-й столбец матрицы меток (без единичного столбца, соответствующего свободному члену) исходного набора признаков, соответствующий i-му признаку, a_i — i-й коэффициент модели M_{o}, Y_{o} — значения зависимой переменной на обучающей выборке, $\text{var}(Y_{\text{o}})$ — дисперсия переменной Y_{o}, $\text{cov}(\cdot)$ обозначает ковариационную матрицу, N — число признаков в исходном наборе. Сумма всех вкладов составляет R^2, но эта связь справедлива только в квазилинейной форме. Сами вклады могут быть положительными либо отрицательными — вклад становится отрицательным только в случае разнонаправленности влияний на зависимую переменную одиночного фактора и с учетом всех факторов модели в целом. Как и в процедуре 1, отбираются признаки, абсолютные значения вкладов которых в коэффициент R^2 модели составляют не менее 30—50 % от наибольшего вклада. ♦

Процедура 3. Отбор признаков производится путем вычисления и сравнения так называемых показателей «изменчивости» для коэффициентов a_o, a_1,..., a_N модели M_{o}. Пусть $Y_{\text{o}} = (Y_{\text{o}}^1, ..., Y_{\text{o}}^{n_{\text{o}}})$ — значения зависимой переменной на обучающей выборке длины n_{o}, а $\hat{Y}_{\text{o}}^j = a_0 + \sum_{i=1}^{N} a_i X_{ji}$ — модельная оценка зависимой переменной для j-го объекта, $j = 1, ..., n_{\text{o}}$. Для каждого коэффициента a_i, $i = 0,..., N$ определяется его оценка по j-му наблюдению (объекту)

$$\hat{a}_i^j = \frac{Y_{\text{o}}^j - \hat{Y}_{\text{o}}^j + a_i X_{ji}}{X_{ji}}, \quad j = 1, ..., n_{\text{o}}, \qquad (2)$$

а также среднее значение $\bar{a}_i = \frac{1}{n_{\text{o}}} \sum_{j=1}^{n_{\text{o}}} \hat{a}_i^j$ таких оценок на обучающей выборке[7]. Величину $V_i = |a_i - \bar{a}_i|/a_i$, назовем *показателем «изменчивости»* коэффициента a_i ♦.

Показатель «изменчивости» того или иного коэффициента может быть положительным или отрицательным в зависимости от знака этого коэффициента. Поясним смысл такого определения. Раскроем выражение (2):

$$\hat{a}_i^j X_{ji} = Y_{\text{o}}^j - \hat{Y}_{\text{o}}^j + a_i X_{ji},$$

$$\Rightarrow Y_{\text{o}}^j - a_{\text{o}} - \sum_{s=1, s\neq i}^{N} a_s X_{js} - a_i X_{ji} + a_i X_{ji} - \hat{a}_i X_{ji} = 0,$$

т. е. оценка коэффициента по объекту показывает, какое значение должен принять данный коэффициент, чтобы невязка рыночной и модельной цен обратилась в ноль для этого объекта. Если усреднить по наблюдениям и преобразовать выражение (2), то получим выражение

$$V_i = \frac{\left| \sum_{j=1}^{n_{\text{o}}} \left(\frac{Y_{\text{o}}^j - \hat{Y}_{\text{o}}^j}{X_{ji}} \right) \right|}{n_{\text{o}} a_i},$$

из которого ясно, что все показатели «изменчивости» V_i пропорциональны среднему значению отношения невязки стоимости к значению признака.

В результате применения процедуры 3 все коэффициенты модели M_{o} упорядочиваются по возрастанию абсолютных значений их показателей «изменчивости», и по аналогичному эвристическому правилу «отбирается» левая часть этого ряда.

2.4. Перенесение меток на значения признаков объектов контрольной выборки и дальнейший отбор регрессоров

Поскольку модель M_{o} определена на помеченных регрессорах, необходимо метки значений этих регрессоров перенести на соответствующие значения признаков объектов контрольной выборки. Пусть X — некоторый признак, а X_K — *исходное* значение этого признака у рассматриваемого объекта контрольной выборки. Из совокупности исходных значений признака X у объектов обучающей

[7] В выражении (2) не может быть деления на 0, поскольку при оценке объектов недвижимости (удельная) стоимость объектов всегда положительна, так что все метки будут также ненулевыми.

выборки формируем совокупность $X_K = \{X_{K1}, X_{K2}\}$ значений, находящихся на минимальном расстоянии от значения X_K. Если совокупность X_K содержит более одного элемента, то значению X_K приписывается метка[8] элемента X_{K1}. Подобным образом переносятся метки на все значения всех признаков объектов контрольной выборки. Описание каждого объекта контрольной выборки в терминах перенесенных меток является приближенным в том смысле, что каждая перенесенная метка для некоторого значения количественного признака может соответствовать сильно различающимся исходным значениям этого признака (не путать близость этих исходных значений с d-близостью) у объектов обучающей и контрольной выборок. Подобная ситуация возникает и при перенесении меток с объектов обучающей выборки на объекты проверочной выборки при оценке её объектов. Для анализа «качества» перенесенных меток на контрольной выборке предлагается следующая процедура.

Процедура 4. Обозначим через $M_к$ ($n_к \times N$)-матрицу меток, перенесенных на значения признаков объектов контрольной выборки, где $n_к$ — длина контрольной выборки. Пусть $\hat{Y}_к$ — вектор модельных значений зависимой переменной на контрольной выборке:

$$\hat{Y}_к = M_к \cdot a, \qquad (3)$$

где a — вектор-столбец коэффициентов модели $M_о$. Вычислим коэффициенты детерминации $R^2_о$ модели $M_о$ и $R^2_к$ «модели», задаваемой соотношением (3), и найдем вклады признаков в соответствующие коэффициенты детерминации каждой из этих моделей. Вклад i-го признака модели $M_о$ в коэффициент $R^2_о$ определяется согласно формуле (1). Аналогично, вклад i-го признака «модели» $M_к$ в коэффициент $R^2_к$ определяется как

$$w^i_к = \frac{\mathrm{cov}(D_к(:, i) \cdot a_i, Y_к)}{\mathrm{var}(Y_к)}, \quad i = 1, ..., N,$$

где $Y_к$ — вектор известных значений зависимой переменной на контрольной выборке. Заметим, что сумма вкладов $w^i_к$ в значения коэффициента детерминации $R^2_к$ не равна $R^2_к$, поскольку коэффициенты модели $M_о$ отличаются от коэффициентов модели, которую можно было бы построить по матрице $D_к$ перенесенных меток и известным значениям $Y_к$ (и для которой сумма вкладов равнялась бы её коэффициенту детерминации). ♦

В данном случае сумма вкладов $w^i_к$ нас не интересует, важны лишь оценки изменчивости вкладов каждого из признаков на обучающей и контрольной выборках.

Определим показатель изменчивости Δw_i вклада i-го признака как

$$\Delta w_i = \frac{1}{w^i_о} |w^i_о - w^i_к|, \quad i = 1, ..., m.$$

Показатель Δw_i может быть положительным или отрицательным в зависимости от знака вклада $w^i_о$. Упорядочим признаки модели $M_о$ по возрастанию абсолютных значений показателей изменчивости их вкладов, и в этом порядке будем последовательно выбирать признаки для построения результирующей модели на обучающей выборке, задавшись некоторым допустимым значением показателя изменчивости. Влияние каждого отбираемого признака (в дополнение к уже отобранным признакам) на качество модели, включающей в себя этот признак, проверяется на контрольной выборке.

Конечно, процедуры 1—4 не являются взаимно исключающими (как показано далее в § 4 на конкретном примере, совокупности признаков, отобранные согласно каждой из этих процедур, могут иметь непустое пересечение). Каждая из них несет определенную информацию о влиянии тех или иных признаков на зависимую переменную, и сравнительный анализ этой информации позволяет с большей уверенностью провести окончательный отбор признаков для построения линейной модели (и внутренне линейных моделей) на обучающей выборке. Процедуры 1—4 применялись при анализе практических выборок в ходе оценивании объектов недвижимости в ряде регионов Российской Федерации и показали свою практическую пригодность для массовой оценки объектов недвижимости.

3. СХЕМА ПОСТРОЕНИЯ ОПТИМАЛЬНОЙ МОДЕЛИ НА ОТОБРАННЫХ ПРИЗНАКАХ С ИСПОЛЬЗОВАНИЕМ МЕТОК ПО БЛИЗОСТИ

Отобранные с использованием процедур 1—4 признаки назовем регрессорами и построим на них линейную (или какую-либо внутренне линей-

[8] В принципе, возможны и другие, не рассматриваемые здесь варианты приписывания меток значениям признаков объектов контрольной выборки.

ную) результирующую модель модель $M_{\text{рез}}$. За отбором факторов стоимости следует важный момент выбора оптимального параметра близости для каждого регрессора. Коэффициенты модели являются неявными функциями от выбранных значений d_i для разных признаков. Справедливо предположить (и это подтверждается на практике), что зависимость вычисляемых коэффициентов корреляции от значений «параметров близости» d_i в общем случае немонотонна и отличается для разных признаков. Отсюда возникает проблема нахождения такого вектора значений этих параметров, при которых построенная модель будет обладать улучшенным качеством оценки, т. е. проблема «оптимизации» модели $M_{\text{рез}}$. В силу неявного задания и сложности минимизируемой функции приходится ограничиваться нахождением некоторого субоптимального вектора искомых параметров. Оптимизацию можно проводить с использованием имеющегося в пакете MATLAB алгоритма многомерной оптимизации с ограничениями, основанного на алгоритме последовательного линейно-квадратичного программирования (*sequential linear-quadratic programming* (*SLQP*) *algorithm*). Для определения точки начального приближения алгоритма оптимизации применяется следующий подход. Каждая координата начального вектора $\vec{d}$ параметров близости выбирается как соответствующая максимуму коэффициента корреляции помеченного признака и зависимой переменной на *контрольной* выборке. Таким образом, метки при этом приписываются на обучающей выборке и переносятся на контрольную, следовательно, оценивается одновременно и удачность выбора параметра d_i и эффективность переноса меток. Затем запускается процесс оптимизации с эмпирическими краевыми ограничениями на вектор $\vec{d}$. Отметим, что начальное приближение, построенное на основе коэффициентов корреляции, дает неплохой эффект, поскольку сдвиг вектора $\vec{d}$ в ходе оптимизации обычно небольшой и происходит не по всем координатам. Укажем, что в ходе оптимизации целевым функционалом служит комбинация нормированных критериев качества на контрольной выборке. К учитываемым критериям качества относятся коэффициент детерминации, стандартное отклонение и относительная погрешность. В литературе часто выбирается один из критериев (чаще всего коэффициент дисперсии или детерминации), и модели сравниваются только по нему, мы же постарались провести более комплексный, всесторонний анализ.

4. ПРИМЕР
Оценка стоимости земельных участков, находящихся под индивидуальными жилыми домами в сельских населенных пунктах Калужской области

Исходная выборка, содержащая рыночную информацию о сделках с земельными участками (ЗУ) указанного выше назначения включала описания ЗУ в разрезе 25-ти признаков и рыночные цены предложений о продаже. Из этой выборки как выбросы были удалены ЗУ с удельной стоимостью более 1100 руб., а также ЗУ с площадью более 9000 м^2. В результирующей выборке осталось 483 ЗУ, которые далее были переупорядочены так, как описано в п. 2.1. Затем из этой выборки были выделены обучающая, контрольная и проверочная выборки.

4.1. Отбор признаков для построения модели на обучающей выборке

Предварительно были отобраны 13 признаков (остальные признаки были отброшены по результатам процедуры 1:

качественные: 1 — направление от Калуги; 2 — район;

количественные: 3 — площадь ЗУ; а также расстояния от ЗУ: 4 — до границы с Московской обл.; 5 — до Москвы; 6 — до автовокзала; 7 — до центра района; 8 — до пристани; 9 —до региональной транспортной магистрали; 10 — до федеральной транспортной магистрали; 11 — до ж/д станции; 12 — до Калуги; 13 — до автобусной остановки.

Таблица 1

Результаты применения процедур 1—4 к исходному набору признаков

Номер процедуры	Признаки												
	1	2	3	4	5	6	7	8	9	10	11	12	13
1	**0,38**	**0,48**	**0,48**	0,27	**0,31**	0,18	**0,34**	0,27	**0,32**	**0,35**	0,27	**0,35**	0,14
2	0,02	**0,07**	**0,16**	**−0,03**	**0,06**	0,01	0,02	**−0,04**	**0,05**	0,03	−0,01	**0,03**	0,01
3	**0,07**	**0,06**	**0,02**	**−0,04**	**0,03**	**0,07**	0,12	**−0,05**	**0,05**	0,09	−1,02	**0,07**	**0,04**
4	0,27	**0,00**	**0,07**	−0,10	**0,08**	0,53	0,44	−0,46	0,53	0,37	−0,79	**0,09**	0,77

Таблица 2

Условия для выделения признаков в процедурах 1—4

Номер процедуры	Условие	Выделяемые признаки
1	Не менее 0,3	1—3, 5, 7, 9, 10, 12
2	Не менее 0,03	2—5, 8, 9, 12
3	Менее $0{,}075 \cdot 10^5$	1—6, 8, 9, 12, 13
4	Менее 0,1	2, 3, 5, 12

Значениям всех этих признаков были приписаны метки согласно описанным в § 1 процедурам. При приписывании меток количественным признакам предварительное значение «параметра близости» было выбрано одинаковым $d = 0{,}2$ для всех количественных признаков. Применим к этому набору помеченных признаков процедуры 1—4, результаты этих процедур сведем в табл. 1, где полужирным шрифтом выделены подходящие для отбора значения признаков.

В табл. 2 для процедур 1—4 указаны допустимые требования на абсолютные значения соответствующих показателей для выделения того или иного признака и совокупности выделяемых признаков в каждой из этих процедур.

Видим, что самой «жесткой» (и самой эффективной) процедурой в плане отбора признаков является процедура 4. Согласно ей, для построения моделей на обучающей выборке в качестве регрессоров были отобраны признаки: 2 — район; 3 — площадь ЗУ; 5 — расстояние до Москвы; 12 — расстояние до Калуги, дополненные признаком 4 — расстояние до границы с Московской областью, для которого значения показателей в каждой из процедур либо удовлетворяют указанным в табл. 2 пороговым значениям, либо довольно близки к ним.

4.2. Сравнительный анализ качества моделей на каждой из выборок в зависимости от способа введения меток для значений регрессоров

Сравнение проводилось на основании 250-ти итераций по случайному переформированию обучающей и контрольной выборок с последующим усреднением. Проверочная выборка была выделена случайным образом и в ходе исследования не менялась. Сопоставлялись следующие варианты:

1) для качественного регрессора *район* — метки по процедуре из § 1, для остальных четырех количественных регрессоров — метки по близости;

2) для качественного регрессора *район* — метки по процедуре из § 1, а значения каждого количественного регрессора делились на среднее этого регрессора на обучающей выборке (для большей схожести диапазонов значений меток для разных регрессоров);

3) для большей представительности сравнения была построена модель, основанная на общепринятом в эконометрике подходе, без меток, согласно которому количественные факторы входят в модель без изменений, а качественные преобразовываются в бинарные (так называемые «*dummy*-переменные»).

В табл. 3 приведены средние по 250-ти реализациям значения указанных параметров качества модели. По соображениям экономии места приводятся результаты только по мультипликативной модели. Аналогичные результаты были получены для линейной и экспоненциальной моделей.

Третий вариант (стандартный эконометрический) хуже, чем два первых, по всем критериям и по всем выборкам. Представляет интерес сравнение двух вариантов, отличающихся преобразованиями количественных регрессоров. Приписывание меток по близости безусловным образом улучшает описательную (аппроксимационную) способность модели, т. е. её качество на обучающей выборке, но предсказательная способность страдает из-за неточности переноса меток на другие выборки; второй же вариант лишен этого недостатка. Из приведенных результатов можно заключить, что положительные свойства меток по близости значительно перевешивают их недостатки, причем по всем принятым критериям.

В табл. 4 приведены значения доверительных (при уровне значимости 0,025) интервалов средних по 250 реализациям значений стоимости на каждой из выборок для всех трех вариантов. Аналогичные

Таблица 3

Значение критериев качества по трем альтернативным подходам для мультипликативной модели

Критерий	Выборка	Вариант 1	Вариант 2	Вариант 3
Относительная погрешность, %	Обучающая	52,10	67,85	84,41
	Контрольная	76,96	76,70	123,04
	Проверочная	59,30	73,31	88,28
Стандартное отклонение	Обучающая	218,48	281,51	274,37
	Контрольная	280,85	302,14	319,01
	Проверочная	550,23	560,63	571,01
Коэффициент детерминации	Обучающая	0,55	0,26	0,37
	Контрольная	0,28	0,17	0,14
	Проверочная	0,20	0,09	0,13

Таблица 4

Доверительные интервалы средних значений стоимости

Выборка	Вариант 1	Вариант 2	Вариант 3
Обучающая	295,03—300,23	269,12—273,22	334,21—339,13
Контрольная	276,54—280,66	270,24—274,30	322,99—328,87
Проверочная	287,28—291,08	275,45—279,06	338,38—342,78

доверительные интервалы можно построить и для всех приведенных в таблице 3 параметров качества. По этим результатам можно заключить, что:

— средние значения стоимостей на каждой из выборок для каждого из вариантов статистически значимо различны;

— вариант 1 обладает статистически значимо лучшим (с учетом результатов в табл. 3) качеством оценки на всех трех выборках по сравнению с двумя другими вариантами.

ЗАКЛЮЧЕНИЕ

Приписывание меток с помощью понятия близости значениям количественных признаков позволяет усиливать степень влияния того или иного признака на зависимую переменную, что дает возможность улучшать аппроксимационные и оценочные свойства модели. Предложенные процедуры отбора признаков для результирующей модели направлены также на улучшение её оценочных свойств на контрольной и проверочной выборках. Сравнение с известными процедурами регрессионного оценивания, проведенное на представительном практическом примере, дало статистически значимые улучшения показателей оценки.

ЛИТЕРАТУРА

1. *Корноушенко Е.К.* Методологические аспекты практического регрессионного оценивания // Проблемы управления. — 2008. — № 2. — С. 34—41.
2. *Ward R.D.* Developing Location Effects Using Cluster Analysis with Response Surface Analysis // Journal of Property Tax Assessment & Administration. — 2005. — Vol. 3, issue 2. — P. 5—17.
3. *Ward R.D.*, *Weaver J.R.*, *German J.C.* Improving CAMA Models Using Geographic Information Systems Response Surface Analysis Location Factors // Assessment Journal. — 1999. — January / February. — P. 30—38.
4. *Kohavi Ron* A study of cross-validation and bootstrap for accuracy estimation and model selection // Proc. of the Fourteenth Intern. Joint Conference on Artificial Intelligence. — Montreal, 1995. — Vol. 2. — P. 1137—1143.
5. *MacKinnon J.G.* Bootstrap Methods in Econometrics // Queen's University Economic Record. — 2006. — Feb. — Vol. 82. — P. S2—S18.
6. *Perez-Amaral T.*, *Gallo G.M.*, *White H. A* Flexible Tool for Model Building: the Relevant Transformation of the Inputs Network Approach (RETINA) // Oxford Bulletin of Economics and Statistics. — 2003. — Vol. 65(s1). — P. 821—838.
7. *Validating* the Regression Model. — URL: http://www.ltrr.arizona.edu/~dmeko/notes_12.pdf (дата обращения 10.07.09).
8. *Fields G.S.* Regression-Based Decompositions: A New Tool for Managerial Decision-Making // 2004 Cornell University, March 2004. — URL: http://www.ilr.cornell.edu/directory/downloads/fields/ Author_decomposingRegressions_mar04.pdf (дата обращения 20.07.2009).

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

Корноушенко Евгений Константинович — д-р техн. наук, гл. науч. сотрудник, Институт проблем управления им. В.А. Трапезникова РАН, г. Москва, ☎ (495) 334-90-00, ✉ ekorno@mail.ru,

Лобко Алексей Александрович — студент VI курса, Московский физико-технический институт, г. Долгопрудный, ✉ Alex.lobko@gmail.com.