ДОКЛАДЫ АН ВШ РФ
2014 январь–март № 1 (22)

ТЕХНИЧЕСКИЕ НАУКИ

УДК 519.237.5

ЛОКАЛЬНО ВЗВЕШЕННОЕ ВОССТАНОВЛЕНИЕ СТРУКТУРНЫХ ЗАВИСИМОСТЕЙ В ЗАДАЧЕ АНАЛИЗА УСПЕВАЕМОСТИ

А.Ю. Тимофеева, О.Е. Аврунев
Новосибирский государственный технический университет

Рассмотрена задача построения локально взвешенной регрессии в условиях, когда один из входных факторов наблюдается со случайными ошибками, а другие являются детерминированными. Наличие погрешностей в объясняющей переменной приводит к ухудшению качества оценивания на основе взвешенного метода наименьших квадратов, поэтому предлагается восстанавливать ортогональную регрессию. Получено аналитическое решение, учитывающее наличие детерминированных факторов в модели. Однако возникает проблема с тем, что веса, задающие локальную область, зависят от параметров регрессии. В этой связи наряду с известным адаптивным алгоритмом разработана итерационная процедура оценивания. Для определения оптимального числа ближайших соседей предложено использовать корень из среднего квадрата остатков модели. В ходе вычислительного эксперимента подтверждена правомерность использования такого критерия при малом и среднем уровне зашумления данных. Большая степень засорения выборки приводит к проблемам со сходимостью итерационного алгоритма и со стабильностью результатов оценивания адаптивным алгоритмом. Это влечет за собой искажение оценок отклика, и тем самым гладкость восстанавливаемой кривой обеспечивается только при значительном числе ближайших соседей. Дальнейшее развитие алгоритмов связывается с повышением их устойчивости к сильному засорению данных. Разработанный итерационный алгоритм применен для исследования успеваемости студентов. Произведено сглаживание средних результатов первой сессии в зависимости от суммарного балла единого государственного экзамена (ЕГЭ), направленности блока изучаемых дисциплин и вида факультета технического вуза. Это позволило сделать качественные выводы об особенностях процесса освоения образовательных программ в вузе и об истинном уровне знаний студентов.

Ключевые слова: локально взвешенная регрессия, ближайший сосед, ортогональная регрессия, метод общих наименьших квадратов, детерминированный фактор, качественный признак, вычислительный эксперимент, оценка успеваемости.

1. Постановка проблемы

Наиболее гибким средством анализа неизвестных зависимостей между признаками можно считать непараметрическое сглаживание [1]. Суть его состоит в локальной аппроксимации зависимости в заданной окрестности каждой точки (или некоторых узловых точек), координаты которой определяются значениями входных признаков. Корректность результатов такого подхода обеспечивается только при ограничении на схему проведения эксперимента: значения объясняющих переменных должны фиксироваться без погрешностей [2, с. 189–192]. В противном случае полученные результаты не всегда позволяют верно оценить влияние факторов, поскольку его требуется отделять от воздействия случайных погрешностей, с которыми наблюдаются признаки. В то же время наряду с переменными, содержащими погрешности измерения, набор входных признаков может включать факторы, значения которых носят детерминированный характер. Именно проблеме непараметрического оценивания значений отклика в таких условиях и посвящена эта работа. Перейдем к формальной постановке задачи.

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 14-07-31171 мол_а.

© 2014 А.Ю. Тимофеева, О.Е. Аврунев

Предполагается, что на выходной признак Y оказывают влияние некоторая переменная P, наблюдаемая с ошибкой, и ряд детерминированных факторов $x_1, x_2, \ldots, x_m$, которые могут носить и качественный характер. Функциональная форма зависимости априорно не постулируется, и необходимо получить некоторую аппроксимацию неизвестной функции $Y = g(P, x_1, x_2, \ldots, x_m)$.

Пусть по результатам наблюдений получена выборка значений введенных переменных объемом N. При этом истинные значения Y_i и P_i остаются ненаблюдаемыми ввиду наличия погрешностей, следовательно, фиксируются значения $y_i = Y_i + \varepsilon_i$, $p_i = P_i + \delta_i$, $i = \overline{1, N}$, где ε_i, δ_i – случайные ошибки, относительно которых предполагается

$$E(\varepsilon_i) = E(\delta_i) = 0,\ D(\varepsilon_i) = \sigma_\varepsilon^2,\ D(\delta_i) = \sigma_\delta^2,\ \forall i,$$

$$\operatorname{cov}(\varepsilon_i, \varepsilon_j) = \operatorname{cov}(\delta_i, \delta_j) = 0,\ \forall i \neq j,\ \operatorname{cov}(\varepsilon_i, \delta_j) = 0,\ \forall i, j.$$

Искомую зависимость согласно [3, с. 503] называют структурной:

$$y_i = g(p_i - \delta_i, x_{i1}, x_{i2}, \ldots, x_{im}) + \varepsilon_i.$$

Задача состоит в оценивании значений отклика. Используемый при такой постановке задачи непараметрический подход в настоящее время активно развивается [4, 5]. Это развитие направлено на его интеграцию с известными методами решения задачи восстановления структурных зависимостей. Особенность этой задачи, как уже подчеркивалось, состоит в наличии погрешностей в объясняющих переменных, что не позволяет использовать стандартные процедуры статистического анализа (в частности, регрессионного и дисперсионного). Сейчас преимущественно разрабатываются подходы, предполагающие привлечение обширной дополнительной информации (инструментальных переменных, повторных наблюдений [6]), что влечет за собой затраты по сбору таких данных. В этой работе предлагается ориентироваться на метод общих наименьших квадратов [7], в линейном случае приводящий к ортогональной регрессии [8]. Он требует лишь фиксации значения соотношения дисперсий ошибок входной переменной и отклика, которое может быть задано исходя из априорных представлений исследователя. Авторами ранее этот подход комбинировался со штрафными регрессионными сплайнами [9]. Здесь в качестве непараметрического метода оценивания выбрана локально взвешенная регрессия.

2. Локально взвешенная регрессия

Наиболее популярный алгоритм оценивания локально взвешенной регрессии предложен в [10]. Здесь остановимся на кратком его изложении.

Основная идея состоит в построении оценок отклика в выбранных точках пространства входных признаков $\tilde{z}_j = (\tilde{p}_j, \tilde{x}_{j1}, \ldots, \tilde{x}_{jm})$ путем восстановления линейной регрессии по k ближайшим к $\tilde{z}_j$ точкам, $1 < k \le N$. Исходя из расстояния h_j от j-й точки до ее k-го ближайшего соседа рассчитываются веса всех точек выборки, определяющие локальную область $\tilde{z}_j$:

$$w_i(\tilde{z}_j) = W\left(h_j^{-1} \rho_{\tilde{z}_j, z_i}\right), \qquad (1)$$

где $\rho_{\tilde{z}_j, z_i}$ – евклидово расстояние от $\tilde{z}_j$ до i-й выборочной точки $z_i = (p_i, x_{i1}, \ldots, x_{im})$. Обычно [10, 11] в качестве $W(x)$ используется функция

$$W(x) = (1 - x^3)^3 H(1 - x),$$

где $H(\cdot)$ – функция Хэвисайда. При вычислении расстояний разница в масштабе измерения входных признаков должна быть компенсирована за счет стандартизации их значений [11, с. 597]. Построенные таким образом веса используются для оценки линейной регрессии в каждой точке $\tilde{z}_j$ с помощью взвешенного метода наименьших квадратов (ВМНК) [12, с. 99–101].

3. Ортогональная регрессия

Применение при сглаживании ВМНК и определение весов на основе исходных значений входных факторов предполагает отсутствие ошибки при фиксации их значений. При отклонении от этого предположения предлагается в заданных точках пространства входных факторов восстанавливать ортогональную регрессию [13, с. 309–310]. При ее построении должно быть задано соотношение дисперсий ошибок $\gamma = \sigma_\delta^2 / \sigma_\varepsilon^2$. Обычно исходят из того, что $\gamma = 1$ и переменные имеют одинаковый масштаб. Здесь будем основываться на более общей постановке [8]: γ может принимать любые положительные значения в соответствии с представлениями исследователя относительно уровня дисперсий ошибок переменных.

Пока будем пренебрегать наличием детерминированных входных факторов и рассмотрим уравнение парной зависимости $y_i = g(p_i - \delta_i) + \varepsilon_i$. Пусть необходимо восстановить локально взвешенную регрессию в каждой из n заданных точек $\tilde{z}_j = \tilde{p}_j$. Задача построения взвешенной ортогональной регрессии в j-й точке сводится к поиску минимума выражения:

$$G_j = \sum_{i=1}^{N} w_i(\tilde{p}_j) \left[\frac{1}{\gamma} \left(p_i - P_{ij} \right)^2 + \left(y_i - \alpha_j - \beta_j P_{ij} \right)^2 \right]$$

по ненаблюдаемым значениям P_{ij} и неизвестным параметрам α_j и β_j. Известно [8], что рассмотренная функция приводится к виду

$$G_j = \frac{1}{1 + \gamma \beta_j^2} \sum_{i=1}^{N} w_i(\tilde{p}_j) \left(y_i - \alpha_j - \beta_j p_i \right)^2, \tag{2}$$

и при фиксированных значениях $w_i(\tilde{p}_j)$ существует аналитическое решение, которое можно представить как

$$\hat{\beta}_j = \frac{\left(s_{yj}^2 - \frac{1}{\gamma} s_{pj}^2 \right) + \sqrt{\left(s_{yj}^2 - \frac{1}{\gamma} s_{pj}^2 \right)^2 + \frac{4}{\gamma} s_{ypj}^2}}{2 s_{ypj}}, \tag{3}$$

$$\hat{\alpha}_j = \bar{y}_j - \hat{\beta}_j \bar{p}_j, \quad \hat{P}_{ij} = \frac{p_i + \gamma \hat{\beta}_j \left(y_i - \hat{\alpha}_j \right)}{1 + \gamma \hat{\beta}_j^2},$$

где s_{yj}^2, s_{pj}^2 – выборочные оценки дисперсий отклика и входной переменной для точки $\tilde{p}_j$ соответственно, s_{ypj} – выборочная оценка ковариации между ними, $\overline{y}_j$, $\overline{p}_j$ – выборочные оценки среднего объясняемой переменной и входного фактора для точки $\tilde{p}_j$ соответственно. Все эти выборочные оценки строятся с учетом весов $w_i(\tilde{p}_j)$.

Заметим, однако, что веса $w_i(\tilde{p}_j)$, определяемые на основе расстояния до ближайших соседей, не могут быть непосредственно вычислены из (1), поскольку значения p_i наблюдаются с ошибкой. Поэтому в качестве p_i при расчете расстояний необходимо использовать оценки истинных значений $\hat{P}_{ij}$, вследствие чего возникает проблема зависимости весов от параметров регрессии.

Задача построения локально взвешенной ортогональной регрессии для парной зависимости рассмотрена в [14], где для решения отмеченной выше проблемы разработан алгоритм адаптивного оценивания. При этом предполагается, что оценки параметров для ближайших точек пространства входных факторов существенно не отличаются и могут быть использованы для расчета оценок объясняющих переменных и весов. Однако при сильном засорении данных такое предположение нарушается и алгоритм адаптивного оценивания может давать нестабильные результаты. Поэтому предлагается итерационная процедура, включающая следующую последовательность шагов.

Шаг 1. Определяется начальное приближение оценок $\hat{\alpha}_j$ и $\hat{\beta}_j$ с весами, вычисленными по (1).

Шаг 2. На основе найденных оценок параметров рассчитываются значения $\hat{P}_{ij}$, которые подставляются в качестве p_i в соотношение (1), и определяются веса.

Шаг 3. По текущим значениям весов рассчитываются оценки $\hat{\alpha}_j$ и $\hat{\beta}_j$.

Шаги 2 и 3 повторяются, пока норма отклонений прогнозных значений отклика на соседних итерациях Δ превышает заданную малую величину.

Далее на модельном примере исследована работа предложенного алгоритма.

Кроме того, остается открытым вопрос выбора оптимального числа ближайших соседей для локально взвешенной ортогональной регрессии. Известно [1], что параметр k определяет степень сглаживания, и для аналогичных целей в непараметрических методах также используются ширина окна и параметр сглаживания. Выбор их значений может быть осуществлен по критерию кросс-валидации [1, с. 44]. Далее остановимся подробнее на этой проблеме.

4. Выбор числа ближайших соседей

При малом числе ближайших соседей восстановленная кривая сильно подвержена случайным колебаниям. В то же время при слишком большом числе k будет обнаруживаться эффект «пересглаживания» [15, с. 54]. При решении конкретных задач результаты определяются вариабельностью истинной функции регрессии и уровнем шума. Поэтому значение k должно подбираться в зависимости от результатов оценивания.

Ранее установлено [9], что применение суммы взвешенных расстояний G_j от точек корреляционного поля до линии ортогональной регрессии в качестве критерия выбора параметра сглаживания чревато рядом проблем. Так в случае, если корреляция между входным фактором и откликом близка к нулю, то линия обыч-

ной регрессии почти горизонтальна, тогда как коэффициент ортогональной регрессии в силу деления на ноль в (3) будет стремиться к бесконечности. Следовательно, при использовании линейных штрафных сплайнов получается ряд практически вертикальных участков [9]. При этом функция (2) имеет минимальное значение, а линия регрессии резко отклоняется от точек корреляционного поля, что противоречит поставленной задаче подгонки кривой.

Таким образом, с помощью сглаживания должны компенсироваться имеющиеся отрицательные эффекты использования целевой функции (2). Однако привлечение критерия кросс-валидации для выбора оптимального размера локальной области сглаживания при оценивании ортогональной регрессии представляется весьма трудоемкой задачей ввиду нелинейности по y оценок параметров. Поэтому можно предложить более простой критерий минимума показателя

$$RMSE_y = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(y(\tilde{z}_j)-\hat{y}(\tilde{z}_j)\right)^2}\ , \tag{4}$$

где $y(\tilde{z}_j)$, $\hat{y}(\tilde{z}_j)$ – наблюдаемое и прогнозное значение отклика в точке $\tilde{z}_j$. Этот показатель должен препятствовать сильному отклонению регрессионной кривой от точек корреляционного поля. Далее на модельном примере исследована пригодность показателя (4) для выбора оптимального числа ближайших соседей.

5. Модель с детерминированными факторами

Далее рассмотрим, как повлияет на процедуру оценивания наличие в модели детерминированных факторов. В этом случае функция G_j будет иметь вид

$$G_j = \sum_{i=1}^{N} w_i(\tilde{z}_j)\left[\frac{1}{\gamma}\left(p_i - P_{ij}\right)^2 + \left(y_i - \boldsymbol{\theta}_j \mathbf{X}_i - \beta_j P_{ij}\right)^2\right],$$

где $\boldsymbol{\theta}_j = \left(\theta_{j0}, \theta_{j1}, \ldots, \theta_{jm}\right)$ – вектор дополнительных параметров, подлежащих оцениванию; $\mathbf{X}_i = \left(x_{i0}, x_{i1}, \ldots, x_{im}\right)'$ – вектор значений детерминированных факторов в i-й точке наблюдений. Параметр θ_{j0} соответствует α_j ($x_{i0} = 1$).

Дифференцируя эту функцию по P_{ij} и приравнивая результат к нулю, получаем выражение для истинных значений входного признака:

$$P_{ij} = \frac{p_i + \gamma\beta_j\left(y_i - \boldsymbol{\theta}_j \mathbf{X}_i\right)}{1+\gamma\beta_j^2}.$$

Подставляя это выражение в исходную функцию, получим

$$G_j = \frac{1}{1+\gamma\beta_j^2}\sum_{i=1}^{N} w_i(\tilde{z}_j)\left(y_i - \boldsymbol{\theta}_j \mathbf{X}_i - \beta_j p_i\right)^2 . \tag{5}$$

Решение задачи минимизации функции (5) относительно неизвестных параметров $\theta_{j0}, \ldots, \theta_{jm}$ приводит к следующей системе линейных уравнений:

$$\sum_{l=0}^{m} s_{x_r l\, j}\theta_{jl} = s_{y x_r j} - s_{p x_r j}\beta_j\ , \quad r = \overline{0, m}\ ,$$

где $s_{x_r l j}$, $s_{y x_r j}$, $s_{p x_r j}$ – выборочные оценки ковариации между признаками x_r и x_l, y и x_r, p и x_r соответственно. В качестве весов при их расчете выступают значения $w_i(\tilde{z}_j)$.

Полученная система уравнений позволяет найти вектор неизвестных параметров θ_j следующим образом:

$$\boldsymbol{\theta}_j = \hat{\boldsymbol{\theta}}_{yxj} - \hat{\boldsymbol{\theta}}_{pxj}\beta_j, \tag{6}$$

где $\hat{\boldsymbol{\theta}}_{yxj}$, $\hat{\boldsymbol{\theta}}_{pxj}$ – векторы оценок параметров уравнений, описывающих влияние детерминированных признаков на y и p соответственно, построенные с учетом весов $w_i(\tilde{z}_j)$. При подстановке (6) в (5) получаем оптимизационную задачу относительно β_j с целевой функцией:

$$G_j = \frac{1}{1+\gamma\beta_j^2}\sum_{i=1}^{N} w_i(\tilde{z}_j)\left(y_i' - \beta_j p_{ij}'\right)^2,$$

где $y_i' = y_i - \hat{\boldsymbol{\theta}}_{yxj}\mathbf{X}_i$, $p_{ij}' = p_i - \hat{\boldsymbol{\theta}}_{pxj}\mathbf{X}_i$. В результате задача сведена к построению ортогональной регрессии, решение которой подробно описано выше.

Отметим, что если детерминированные факторы качественные и включаются в регрессию с помощью фиктивных переменных, в локальную область могут попадать объекты, ни один из которых не соответствует каким-либо из уровней фактора. Тогда возникает проблема с вырожденностью ковариационной матрицы детерминированных факторов. В таком случае имеет смысл провести редукцию [16, с. 229] и обнулить параметр, соответствующий отсутствующему уровню фактора.

6. Задача оценки успеваемости

Предложенный подход к оцениванию локально взвешенной регрессии со стохастическими и детерминированными входными факторами применен для анализа успеваемости студентов технического вуза в первую сессию. Ранее в ходе исследования процесса обучения специалистов [17] подтверждена гипотеза о том, что более высокий уровень подготовки абитуриентов определяет их возможность раннего и углубленного приобретения профессиональных знаний и навыков.

Процесс освоения образовательной программы и оценки знаний студента может быть представлен в самом простом виде как некая функция f, аргументами которой являются знания, полученные на предыдущем этапе, и факторы, характеризующие условия обучения на текущем этапе (семестре). Таким образом, запишем выражение для уровня знаний $U^{(t)}$ в семестре t:

$$U^{(t)} = f\left(U^{(t-1)}, V^{(t)}\right),$$

где $V^{(t)}$ – факторы, характеризующие условия обучения в семестре t.

Результаты сессии $y^{(t)}$ в семестре t можно рассматривать как действительный уровень знаний студента, наблюдаемый с ошибкой $\varepsilon^{(t)}$:

$$y^{(t)} = f\left(U^{(t-1)}, V^{(t)}\right) + \varepsilon^{(t)}.$$

Знания, которыми обладает студент на момент поступления, частично наблюдаемы посредством результатов ЕГЭ, на основании которых он поступил в вуз:

$$p^{(0)} = P^{(0)} + \delta^{(0)},$$

где $p^{(0)}$ – наблюдаемые результаты ЕГЭ; $P^{(0)}$ – истинный уровень знаний по предметам, включенным в ЕГЭ; $\delta^{(0)}$ – случайная ошибка. Случайные величины $\varepsilon^{(t)}$ и $\delta^{(0)}$ можно считать некоррелированными, так как субъективные факторы, вносящие погрешности в измерения уровня знаний, не взаимосвязаны: аттестации по ЕГЭ и результатам сессии достаточно разнесены по времени и отличаются условиями проведения, а также контролирующими материалами.

При этом общий уровень знаний студента включает также $B^{(0)}$ – знания по предметам, не оцениваемым в рамках ЕГЭ, т. е.

$$U^{(0)} = P^{(0)} + B^{(0)}.$$

Таким образом, при решении задачи анализа успеваемости наблюдаемыми будут результаты ЕГЭ и факторы $V^{(t)}$, характеризующие учебный процесс в семестре t, которые могут быть как количественными, так и качественными переменными. В то же время уровень знаний студентов $B^{(0)}$, не оцениваемый с помощью ЕГЭ, ненаблюдаем в первую сессию, поэтому далее мы им будем пренебрегать.

7. Результаты экспериментальных исследований

Для моделирования зависимости истинного уровня знаний Y в первую сессию от входного уровня знаний P студентов по предметам, включенным в ЕГЭ, использовалась степенная функция:

$$Y = P^{a_l},\ l = \overline{0,2},$$

где значения P заданы по равномерной сетке из интервала $[0,1]$, коэффициенты a_l отражали воздействие качественного фактора в ходе обучения. Рассмотрено три уровня фактора: отсутствие изменений в истинном уровне знаний при $a_0 = 1$; более низкий по сравнению со входным текущий уровень знаний по специальным предметам, не включенным в ЕГЭ, при $a_1 > 1$; наконец, приращение уровня знаний при $a_2 = 1/a_1$. Задаваемые величины a_1 представлены в таблице. При каждом a_l генерировалось по 300 наблюдений, $N = 900$. Истинные значения зашумлены независимыми нормально распределенными случайными ошибками с уровнем шума φ (см. таблицу). Понятие уровня шума введено в [18, с. 97; 19, с. 13]. Величина γ определялась как отношение дисперсии P к дисперсии Y.

Оценивание производилось с помощью адаптивного и итерационного алгоритмов в отдельных точках выборки, $m = 150$, по 50 на каждый уровень фактора. Число ближайших соседей варьировалось от 10 до 90 % от объема выборки с шагом 10. В ряде случаев итерационный алгоритм не сходился, поскольку наблюдалось чередование двух локальных оптимумов. Вследствие этого дополнительно учтена сходимость по модулю разности Δ на соседних итерациях. Кроме того, остановка работы алгоритма производилась при достижении максимального числа итераций $\nu_{\max} = 10$, которого при среднем засорении выборки было достаточ-

но для сходимости. Результаты усреднения по 100 выборкам представлены в таблице. В качестве среднего использовалась медиана. Величина показателя (4) с истинными значениями отклика вместо наблюдаемых $RMSE_Y$ умножалась на 100. В скобках приведена величина межквартильного размаха [13, с. 181].

Результаты вычислительных экспериментов

Схема		Алгоритм									
		Адаптивный				Итерационный					
a_1	φ	$RMSE_Y^*$	$RMSE_Y$	d^*	d	$RMSE_Y^*$	$RMSE_Y$	d^*	d	ν^*	ν
2	1	1,4 (0,20	1,4 (0,3)	0,2	0,2	0,8 (0,2)	0,8 (0,2)	0,1	0,1	3	3
	5	2,3 (0,8)	2,8 (1,0)	0,3	0,5	2,1 (0,9)	2,5 (1,2)	0,2	0,2	5	5
	10	3,5 (1,3)	4,2 (1,2)	0,5	0,6	3,3 (1,2)	3,9 (1,6)	0,2	0,5	8	4
4	1	2,8 (0,5)	3,1 (0,4)	0,3	0,4	1,1 (0,4)	1,1 (0,4)	0,1	0,1	4	4
	5	3,7 (0,9)	4,1 (1,2)	0,4	0,5	2,9 (1,3)	3,3 (2,3)	0,2	0,2	6	6
	10	5,6 (1,6)	6,6 (2,1)	0,5	0,6	4,7 (2,1)	6,8 (3,0)	0,2	0,6	9	6
8	1	2,6 (1,1)	2,9 (1,2)	0,3	0,3	1,5 (0,8)	1,6 (1,0)	0,1	0,1	5	5
	5	5,2 (2,8)	6,2 (3,8)	0,4	0,5	3,9 (2,3)	6,3 (5,9)	0,2	0,2	7	6
	10	8,9 (6,4)	11 (7,9)	0,5	0,6	6,5 (5,3)	11,4 (7,9)	0,2	0,7	10	5

В таблице представлены показатели, рассчитанные при числе соседей, соответствующем оптимальному по критерию (4) с $y(\tilde{z}_j)$ и с $Y(\tilde{z}_j)$ вместо $y(\tilde{z}_j)$ (отмечены *). Медианы оптимальных значений доли числа ближайших соседей в объеме выборки d и d^* по двум критериям при малом и среднем уровне шума практически совпадают. Следовательно, использование предложенного критерия для выбора числа ближайших соседей в этом случае вполне оправдано.

При сильно зашумленных данных возникают проблемы в работе обоих алгоритмов: итерационный алгоритм медленно сходится (требуется большое число итераций ν), адаптивный – дает нестабильные оценки. В таких условиях лучшие результаты достигаются при увеличении локальной области сглаживания.

Предложенный итерационный алгоритм применен для анализа зависимости успеваемости студентов очной формы обучения в первую сессию от суммарного балла ЕГЭ. Фактические данные получены из информационной системы НГТУ [20], где накапливается информация как обо всех результатах обучения студентов, так и о факторах, способных влиять на успешность освоения ими образовательной программы: результаты ЕГЭ, на основании которых студенты были зачислены в вуз, а также информация об учебных планах и организации учебного процесса.

В качестве детерминированных входных факторов рассматривались:

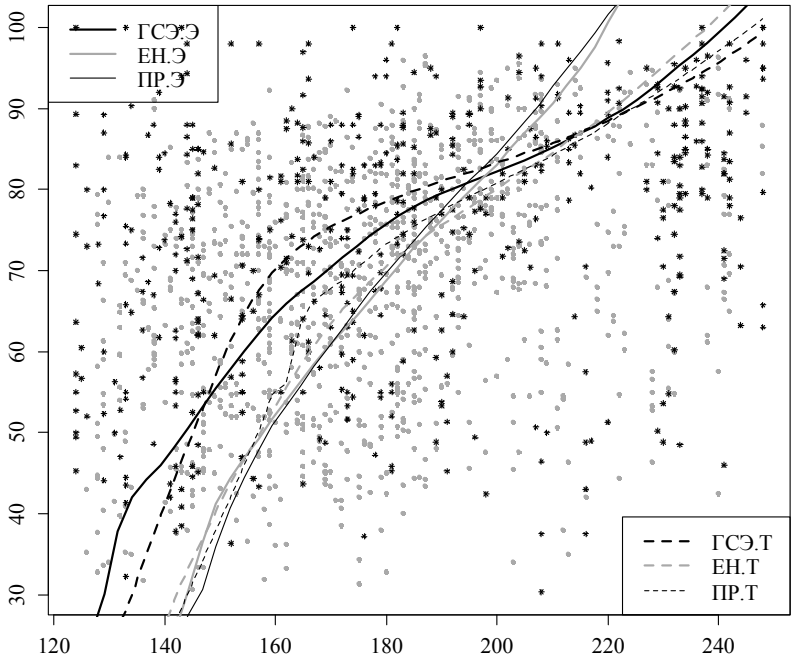
– вид факультета – ФБ и ФМА;

– направленность учебных дисциплин – профессиональные (ПР), естественно-научные (ЕН), гуманитарные и социально-экономические (ГСЭ) дисциплины.

Объем выборки составил 513 студентов. Успеваемость в осеннюю сессию 2012 года определялась по среднему баллу по всем дисциплинам соответствующего блока в 100-балльной шкале. Исходные данные изображены на рисунке светлыми маркерами по экономическому факультету, темными – по техническому.

При оценивании предполагалось, что соотношение дисперсий ошибок пропорционально дисперсиям наблюдаемых переменных. Для получения наилучшего результата в силу сильной зашумленности данных выбрано большое число ближайших соседей (97,5 % от объема выборки). Прогнозные значения отклика изображены линиями на рисунке. Видно, что прогнозные значения результативности освоения одних и тех же групп дисциплин близки для разных факультетов.

Наиболее существенные отличия проявляются в освоении различных блоков дисциплин. Так, при одном и том же входном уровне знаний результаты освоения студентами гуманитарных и социально-экономических дисциплин оцениваются более высоко, чем естественно-научных и профессиональных.



Зависимость успеваемости в первую сессию от суммарного балла ЕГЭ (ФБ и ФМА)

Таким образом, предложенный подход для задачи оценки успеваемости позволяет выявить группы дисциплин, где влияние входного уровня знаний студентов проявляется особенно сильно, что может помочь в корректировке методик обучения для снижения вероятности отчисления на начальных семестрах. В целом подход может стать хорошей альтернативой полупараметрическим методам [21], поскольку позволяет более гибко описывать влияние детерминированных (качественных) признаков. Дальнейшее развитие локально взвешенного сглаживания структурных зависимостей связывается с повышением устойчивости к сильному засорению данных на основе идей робастного оценивания [22], а также с исследованием возможности глобальной оптимизации целевой функции по неизвестным параметрам, не требующей применения адаптивных и итерационных процедур.

ЛИТЕРАТУРА

[1] **Анатольев С.** Непараметрическая регрессия // *Квантиль*. – 2009. – №7. – С. 37–52.

[2] **Катковник В.Я.** Непараметрическая идентификация и сглаживание данных: метод локальной аппроксимации. – М.: Наука, 1985. – 336 с.

[3] **Кендалл М.** Статистические выводы и связи/ М. Кендалл, А. Стьюарт. – М.: Наука, 1973. – 899 с.

[4] **Blundell R.** Semi-nonparametric IV estimation of shape-invariant engel curves / R. Blundell, X. Chen, D. Kristensen // *Econometrica*. – 2007. – Vol. 75. – № 6. – Pp. 1613–1669.

[5] **Khan S.** Weighted And Two-Stage Least Squares Estimation Of Semiparametric Truncated Regression Models / S. Khan, A. Lewbel // *Econometric Theory.* – 2007. – Vol. 23. – № 2. – Pp. 309–347.

[6] **Schennach S.M.** Estimation of nonlinear models with measurement error // *Econometrica.* – 2004. – Vol. 72. – № 1. – Pp. 33–75.
[7] **Van Huffel S.**, Vandewalle J. The Total Least Squares Problem: Computational Aspects and Analysis // *SIAM*. – 1991. – 288 p.
[8] **Тимофеев В.С.** Идентификация зависимостей признаков стохастической природы на основе регрессии Деминга / В.С. Тимофеев, В.Ю. Щеколдин, А.Ю. Тимофеева // *Информатика и ее применения*. – 2013. – Т. 7. – Вып. 2. – С. 60–68.
[9] **Тимофеева А.Ю.** Полупараметрическое оценивание зависимостей между стохастическими переменными / А.Ю. Тимофеева, О.И. Бузмакова // *Научный вестник НГТУ*. – 2012. – № 4 (49). – С. 29–37.
[10] **Cleveland W.S.** Robust Locally Weighted Regression and Smoothing Scatterplots // *Journal of the American Statistical Association.* – 1979. – Vol. 74. – № 368. – Pp. 829–836.
[11] **Cleveland W.S.** Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting / W.S. Cleveland, S.J. Devlin // *Journal of the American Statistical Association.* – 1988. – Vol. 83. – № 403. – Pp. 596–610.
[12] **Вучков И.** Прикладной линейный регрессионный анализ / И. Вучков, Л. Бояджиева, Е. Солаков. – М.: Финансы и статистика, 1987. – 239 с.
[13] **Cramér H.** Mathematical Methods of Statistics. – Bombay: Asia Publishing House, 1962. – 575 p.
[14] **Pinson P.** Local Linear Regression with Adaptive Orthogonal Fitting for the Wind Power Application / P. Pinson, H.A. Nielsen, H. Madsen, T.S. Nielsen // *Statistics and Computing.* – 2008. – Vol. 18. – № 1. – Pp. 59–71.
[15] **Härdle W.** Applied Nonparametric Regression. – New York: Cambridge University Press, 1992. – 352 p.
[16] **Тимофеев В. С.** Эконометрика / В.С. Тимофеев, А.В. Фаддеенков, В.Ю. Щеколдин. – М.: Юрайт, 2013. – 328 с.
[17] **Борисова А. А.** Готовность будущих специалистов состояться в профессии (индикаторы мониторинга) / А.А. Борисова, В.С. Тимофеев, А.Ю. Тимофеева // *Труд и социальные отношения*. – 2013. – № 2. – С. 62–80.
[18] **Ивахненко А.Г.** Помехоустойчивость моделирования / А.Г. Ивахненко, В.С. Степашко. – Киев: Наукова думка, 1985. – 216 с.
[19] **Денисов В. И.** Устойчивые распределения и оценивание параметров регрессионных зависимостей / В. И. Денисов, В. С. Тимофеев // *Известия Томского политехнического университета*. – 2011. – Т. 318. – № 2. – С. 10–15.
[20] **Стасышин М.В.** Информационная система университета: опыт создания и текущее состояние / В. М. Стасышин, О. Е. Аврунев, Е. В. Афонина, К. Н. Лях // *Открытое и дистанционное образование*. – 2012. – № 2(46). – С. 9–15.
[21] **Yatchev A.** Semiparametric Regression for the Applied Econometrician. – Cambridge University Press, 2003. – 213 p.
[22] **Денисов В.И.** Устойчивое оценивание нелинейных структурных зависимостей / В.И. Денисов, А.Ю. Тимофеева, Е.А. Хайленко, О.И. Бузмакова // *Сибирский журнал индустриальной математики*. – 2013. – № 4. – С. 47–60.

LOCALLY WEIGHTED SMOOTHING OF STRUCTURAL RELATIONSHIPS FOR THE STUDENT PROGRESS ANALYSIS

Timofeeva A.Yu., Avrunev O.E.
Novosibirsk State Technical University, Novosibirsk, Russia

The problem of estimating locally weighted regression is considered in conditions when one of the input factors is observed with random errors, while others are deterministic. The presence of errors in the input variable leads to a degradation in the estimation quality based on the weighted least squares method. For this purpose it is proposed to estimate the orthogonal regression. An analytical solution has been found. It takes into account the presence of deterministic factors in the model. However, the problem is that the weights specifying the local area depend on regression parameters. Therefore, an iterative estimation procedure has been developed along with the known adaptive algorithm. To determine an optimal number of nearest neighbors it is proposed to use a root mean square error. In the computing experiments the validity of this criterion has been proved for small and medium noise levels in the data. Heavy contamination of the

sample leads to problems with the convergence of the iterative algorithm and with the stability of estimation results of the adaptive algorithm. This entails a distortion of response estimates, and thus the curve smoothness is provided only if the number of nearest neighbors is great. Further development of the algorithms is related to an increase in their resistance to erroneous data. The developed iterative algorithm has been used to assess the progress of students. Average results of the first examination session were smoothed depending on the total score of the unified state exam, the subjects studied and the university department. This allowed making qualitative assessments of the efficiency of the process of learning at the University and of the actual level of student knowledge.

Keywords: locally weighted regression; nearest neighbor; orthogonal regression; total least squares; deterministic factor; qualitative attribute; computing experiment; progress evaluation.

REFERENCES

[1] **Anatol'ev S.** Neparametricheskaia regressiia [Nonparametric regression]. *Kvantil'*, 2009, no. 7, pp. 37–52.

[2] **Katkovnik V.Ia.** *Neparametricheskaia identifikatsiia i sglazhivanie dannykh: metod local'noi approksimatsii* [Nonparametric identification and data smoothing: local approximation method]. Moscow, Nauka Publ., 1985. 336 p.

[3] **Kendall M., St'iuart A.** *The Advanced Theory of Statistics: Inference and relationship*. London, Charles Griffin and Co., Ltd., 1961, 676 p. (Russ. ed.: M. Kendall, A. St'iuart *Statisticheskie vyvody i sviazi*. Moscow, Nauka Publ., 1973. 899 p.).

[4] **Blundell R., Chen X., Kristensen D.** Semi-nonparametric IV estimation of shape-invariant engel curves. *Econometrica*, 2007, no. 6, pp. 1613–1669. doi: 10.1111/j.1468-0262.2007.00808.x.

[5] **Khan S., Lewbel A.** Weighted And Two-Stage Least Squares Estimation Of Semiparametric Truncated Regression Models. *Econometric Theory*, 2007, no. 2, pp. 309–347. doi: 10.1017/S0266466607070132.

[6] **Schennach S.M.** Estimation of nonlinear models with measurement error. *Econometrica*, 2004, no. 1, pp. 33–75. doi:10.1111/j.1468-0262.2004.00477.x.

[7] **Van Huffel S., Vandewalle J.** *The Total Least Squares Problem: Computational Aspects and Analysis*. Philadelphia, SIAM, 1991, 288 p.

[8] **Timofeev V.S., Shchekoldin V.Iu., Timofeeva A.Iu.** Identifikatsiia zavisimostei priznakov stokhasticheskoi prirody na osnove regressii Deminga [The error-in-variables model identification on the basis of Deming's approach]. *Informatika i ee primenenija*, 2013, no. 2, pp. 60–68.

[9] **Timofeeva A.Iu., Buzmakova O.I.** Poluparametricheskoe otsenivanie zavisimostei mezhdu stokhasticheskimi peremennymi [Semiparametric estimation of regression with stochastic variables]. *Nauchnyi vestnik NGTU*, 2012, no. 4, pp. 29–37.

[10] **Cleveland W.S.** Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal of the American Statistical Association*, 1979, no. 368, pp. 829–836. doi: 10.1080/01621459.1979.10481038

[11] **Cleveland W.S., Devlin S.J.** Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting. *Journal of the American Statistical Association*, 1988, no. 403, pp. 596–610. doi: 10.1080/01621459.1988.10478639

[12] **Vuchkov I., Boiadzhieva L., Solakov E.** *Prikladnoi lineinyi regressionnyi analiz* [Applied Linear Regression Analysis]. Moscow, Finansy i statistika Publ., 1987, 239 p.

[13] **Cramér H.** *Mathematical Methods of Statistics*. Bombay, Asia Publishing House, 1962, 575 p.

[14] **Pinson P., Nielsen H.A., Madsen H., Nielsen T.S.** Local Linear Regression with Adaptive Orthogonal Fitting for the Wind Power Application. *Statistics and Computing*, 2008, no. 1, pp. 59–71. doi: 10.1007/s11222-007-9038-7.

[15] **Härdle W.** *Applied Nonparametric Regression*. New York, Cambridge University Press, 1992, 352 p.

[16] **Timofeev V.S., Faddeenkov A.V., Shchekoldin V.Iu.** *Ekonometrika* [Econometrics]. Moscow, Iurait Publ., 2013. 328 p.

[17] **Borisova A.A., Timofeev V.S., Timofeeva A.Iu.** Gotovnost' budushchikh spetsialistov sostoiat'sia v professii (indikatory monitoringa) [Future specialists' commitment to become

competent professionals (monitoring indicators)]. *Trud i sotsial'nye otnosheniia*, 2013, no. 2, pp. 62–80.

[18] **Ivakhnenko A.G., Stepashko V.S.** *Pomekhoustoichivost' modelirovanija* [Noise stability of modeling]. Kiev, Naukova dumka Publ., 1985. 216 p.

[19] **Denisov V.I., Timofeev V.S.** Ustoichivye raspredeleniia i otsenivanie parametrov regressionnykh zavisimostei [Stable distributions and estimating parameters of regression dependences]. *Izvestiia Tomskogo politekhnicheskogo universiteta*, 2011, no. 2, pp. 10–15.

[20] **Stasyshin V.M., Avrunev O.E., Afonina E.V., Liakh K.N.** Informatsionnaia sistema universiteta: opyt sozdaniia i tekushchee sostoianie [University information system: experience of creating and current state]. *Otkrytoe i distantsionnoe obrazovanie*, 2012, no. 2, pp. 9–15.

[21] **Yatchev A.** *Semiparametric Regression for the Applied Econometrician*. Cambridge, University Press, 2003, 213 p.

[22] **Denisov V.I., Timofeeva A.Yu., Khailenko E.A., Buzmakova O.I.** Robust estimation of nonlinear structural models. *Journal of Applied and Industrial Mathematics*, 2014, no. 1, pp. 28–39. doi: 10.1134/S1990478914010049.

СВЕДЕНИЯ ОБ АВТОРАХ

Тимофеева Анастасия Юрьевна – родилась в 1984 году, канд. экон. наук, старший преподаватель кафедры экономической информатики Новосибирского государственного технического университета. Область научных интересов: развитие методов статистического анализа объектов стохастической природы, в том числе социально-экономических явлений. Опубликовано 25 научных работ. (Адрес: 630073, Россия, Новосибирск, пр. Карла Маркса, 20. Email: a.timofeeva@corp.nstu.ru)

Timofeeva Anastasia Yur'evna (b. 1984) – PhD (Econ.), Senior Lecturer of Computer Science in Economics Department of the Novosibirsk State Technical University. Her research interests are currently focused on the methods development for the statistical analysis of stochastic objects nature, including socio-economic phenomena. She is author of 25 scientific papers. (Address: 20, Karl Marx Av., Novosibirsk, 630073, Russia. Email: a.timofeeva@corp.nstu.ru)

Аврунев Олег Евгеньевич – родился в 1981 году, окончил Новосибирский государственный технический университет (НГТУ), с 2013 года аспирант кафедры программных систем и баз данных НГТУ, директор Центра информатизации университета НГТУ. Область научных интересов: статистическое моделирование учебного процесса, разработка информационных систем. (Адрес: 630073, Россия, Новосибирск, пр. Карла Маркса, 20. Email: avrunev@ciu.nstu.ru)

Avrunev Oleg Evgen'evich (b. 1981) – graduated from the Novosibirsk State Technical University (NSTU), Post-graduated Student of Software Systems and Databases Department of the NSTU, Deputy Manager of Information Technologies Center in NSTU. His research interests are currently focused on statistical simulation of the educational process, information systems development. (Address: 20, Karl Marx Av., Novosibirsk, 630073, Russia. Email: avrunev@ciu.nstu.ru)

Статья поступила 17 февраля 2014 г.
Received 17 Feb. 2014

To Reference:
Timofeeva A.Yu., Avrunev O.E. Lokal'no vzveshennoe vosstanovlenie strukturnykh zavisimostei v zadache analiza uspevaemosti [Locally weighted smoothing of structural relationships for the student progress analysis]. *Doklady Akademii Nauk Vysshei Shkoly Rossiiskoi Federatsii* [*Reports of Russian Higher Education Academy of Sciences*], 2014, no. 1(22), pp. 135–146. (in Russ.).