Уфа : УГАТУ, 2007 Вестник УГАТУ • Математика Т. 9, №3 (21). С. 46–50

УДК 519.237.5

А. А. ВАРИКОВ

МОДИФИЦИРОВАННАЯ НЕПАРАМЕТРИЧЕСКАЯ МОДЕЛЬ ОЦЕНКИ РЕГРЕССИИ И ЕЕ ПРИМЕНЕНИЕ ДЛЯ РАСЧЕТА НАЛОГОВЫХ ОБЯЗАТЕЛЬСТВ

Рассматривается модифицированная непараметрическая модель регрессии для расчета налоговых обязательств. Приведены результаты тестирования этой модели. *Непараметрическая модель регрессии, расчет налоговых обязательств*

ВВЕДЕНИЕ

Будем использовать следующую терминологию и обозначения при изучении статистической модели. Пусть $(X_i, Y_i)(i = 1, .., N)$ – множество статистических данных.

При каждом значении x Y является случайной величиной, поэтому желательно найти его математическое ожидание $E(Y)$ при условии, что X приняло конкретное значение x , то есть находим $f(x) = E(Y|x)$.

Функцию $f(x)$ будем называть истинной функцией регрессии Y на X.

Основная гипотеза классического регрессионного анализа состоит в том, что считается известным вид функциональной зависимости, а неизвестны лишь некоторые параметры этой зависимости. То есть предполагается, что $f(x) = E(Y|x) = F(x, Q)$, где $Q = (Q_1, Q_2, ..., Q_m)$– параметры модели, которые подлежат определению. Как правило, для их определения применяется метод наименьших квадратов, т.е. параметры Q выбираются так, что функционал $S(Q) = \sum_{j=1}^{N}(F(x_j, Q) - Y_j)^2$ достигает минимума.

Классический регрессионный анализ используется, например, при построении производственной функции предприятия.

В тех случаях, когда вид функциональной зависимости неизвестен, что, как правило, имеет место при моделировании налоговых обязательств, встает задача приближения регрессионной зависимости $f(x)$. Здесь возможны различные подходы (см. например [1],[2]). При моделировании налоговых обязательств в основном развиваются два подхода([3],[4]): метод нейросетевого моделирования и метод, сочетающий нейросетевой подход с непараметрическим оцениванием регрессии. Оба метода позволяют со сколь угодно высокой точностью приблизить любую функцию. Например, можно построить такую функцию $\hat{f}(x)$, что величина $\sum_{j=1}^{N}(\hat{f}(x_j) - Y_j)^2$ будет практически равна нулю, если $x_i \neq x_j$ $(i, j = 1...N)$. Однако в этом случае функция $\hat{f}(x)$ аппроксимирует траекторию случайного процесса, проходящую через точки (x_j, y_j), а не регрессионную зависимость $f(x)$.

Условимся далее называть аппроксимативными (или неклассическими) методами моделирования статистической зависимости такие, в которых не предполагается известным вид функциональной зависимости $f(x) = E(Y|x)$.

Допустим, что мы нашли некоторое приближение $\hat{f}(x)$ функции регрессии $f(x)$. Тогда определена модель:

$$\hat{Y}(x) = \hat{f}(x).$$

В этом случае отклонение фактического значения $Y(x)$ от расчетного $\hat{Y}(x)$ состоит из двух частей:

Δ_1 – отклонение от истинной функции регрессии, Δ_2 – ошибка модели.

Возможная величина отклонения Δ_1 характеризуется собственной стохастичностью изучаемого явления.

Оценка качества модели сводится к оценке ее адекватности, точности, устойчивости.

В ситуации, когда функция регрессии неизвестна, мерой адекватности модели наиболее разумно принять оценку истинной функции регрессии от ее приближения: $|f(x) - \hat{f}(x)| \leq \Delta$.

При использовании методов непараметрического оценивания регрессии такую оценку можно получить из [1], используя неравенство $\left|f(x)-\hat{f}(x)\right| \le 2\hat{\sigma}/\sqrt{N}$, $\hat{\sigma}=\min_b d(b)$, где функция $d(b)$ определяется формулой (3). Отсюда следует, что при увеличении числа наблюдений разница между приближенной $\hat{f}(x)$ и точной $f(x)$ моделью исследуемого показателя Y стремится к нулю, т. е. $\Delta_2 \to 0$. Величина $\hat{\sigma}$ задает асимптотическую несмещенную оценку среднеквадратичного отклонения исследуемого показателя от своего математического ожидания, т. е., получена оценка собственной стохастичности Δ_1 показателя Y. Эта оценка дает возможность определить количество наблюдений, достаточных для аппроксимации математического ожидания $E(Y|x)$ (т. е. функции $f(x)$) с заданной точностью.

Приведенные выше соображения могут быть использованы для оценки суммы ошибок вида Δ_1 и Δ_2.

Если кратко охарактеризовать суть нейросетевого моделирования, то это построение регрессионной зависимости в виде смешанной кусочно-линейной функции. При этом активационная функция служит для непрерывного сращивания функции на стыках подобластей, а число нейронов в скрытых слоях определяет количество выделяемых подобластей. Ясно, что таким образом можно аппроксимировать любую функцию. Поэтому нейросетевой метод моделирования можно отнести к классу аппроксимативных.

В данной работе мы кратко остановимся на другом аппроксимированном методе выявления регрессионной зависимости, сочетающим методы нейросетевого моделирования и непараметрического оценивания регрессии [1].

В основе лежит непараметрическое оценивание регрессии, а с нейросетевым методом предлагаемый ниже подход объединяет использование линейных подмоделей и кластеризации.

1. ОБЩАЯ ХАРАКТЕРИСТИКА НЕПАРАМЕТРИЧЕСКИХ МОДЕЛЕЙ

Непараметрический подход к моделированию позволяет ослабить два основных требования классической постановки регрессионной зависимости. Первое – предположение о том, что $\hat{Y}=E(Y/X)$ как функция X представима в виде $f(X;Q)$, где $f(...,...)$ – известная функция своих аргументов, а Q – вектор неизвестных параметров, оцениваемый по выборочным данным, заменяется на более слабое предположение, что $f(X)$ – непрерывная и гладкая функция X. Второе – требование постоянства дисперсии $\sigma^2(X)$ заменяется на предположение ее непрерывности.

Возможны различные варианты построения непараметрических моделей [1]. Например, можно построить модель следующего вида:

$$\hat{Y}(x,\vec{X}_b)=\frac{\sum_i \omega(X,X_i)\cdot y_i}{\sum_i \omega(X,X_i)}, \qquad (1)$$

где

$$\omega(X,X_i)=(1+b\cdot\sum_k (x_k-x_{ik})^2)^{-\frac{1}{2}}. \qquad (2)$$

Здесь X – точка в факторном пространстве $X=(x_1,x_2,...,x_n)$, $\vec{X}_b$ - выборка из точек (X_i,y_i) , b – параметр метода.

Модель (1) можно интерпретировать как многомерный аналог метода скользящего осреднения, где $\omega(x,x_i)$ – веса, которые достаточно быстро убывают по мере удаления точки x_i от расчетной точки x .

Выбор параметра b является ключевым моментом в построении модели. Здесь должен быть установлен баланс между точностью и устойчивостью модели. Нетрудно показать, что при увеличении параметра b можно добиться того, что модель будет проходить по точкам очень близким к точкам выборки. Однако на другой выборке погрешность модели может оказаться очень высокой.

Наоборот, если b взять очень большим, то $\hat{Y}(x)$ будет близка к среднему значению Y.

Выбор параметра b осуществляется с помощью минимизации функции $d(b)$:

$$d^2(b)=n^{-1}\sum_{i=1}^{n}(y_i- -\sum_{j\neq i}\omega(x_j,x_i)\cdot y_i/\sum_{j\neq i}\omega(x_j,x_i))^2, \qquad (3)$$

которая задает асимптотически несмещенную оценку величины среднеквадратической погрешности непараметрической аппроксимации.

2. МОДИФИКАЦИЯ НЕПАРАМЕТРИЧЕСКОЙ МОДЕЛИ

Применение модели (1), (2) к расчету налога на добавленную стоимость для группы разномасштабных строительных предприятий дало хорошие результаты для средних предприятий и неудовлетворительные для мелких и крупных. Простое разделение группы на подгруппы может несколько уточнить модель, но при этом теряется свойство непрерывности на стыках кластеров, кроме того, в приграничных точках кластеров качество модели все равно ухудшается.

В предлагаемом модифицированном методе для уточнения модели используется оптимальная кластеризация, учитывается сила влияния отдельных факторов на выходной показатель с помощью предварительного построения линейных подмоделей, автоматическое удаление аномальных наблюдений.

Отметим основные этапы построения модели. При этом мы считаем, что экономическая модель построена.

Сбор данных

Первый этап состоит как обычно в сборе данных и их первичной обработке, которая состоит в экспертном анализе данных, удалении аномальных наблюдений, нормировки данных, т. е. переводе в безразмерные величины, по формуле $x'_k = x_k/\bar{x}_k$, где $\bar{x}_k = \frac{1}{N}\sum x_{ik}$, или $x'_k = x_k/\sigma(x_k)$, где $\sigma(x_k)$ – среднеквадратическое отклонение x_k. Далее в обозначениях штрих будем опускать.

Предварительное построение линейной модели. Кластеризация

Здесь возможны два варианта. В первом варианте число кластеров задается пользователем (экспертом), во втором случае выбирается оптимальное число кластеров, минимизирующее функционал (1.4). Кластеризация ведется по масштабным факторам. Пусть, например, $x_1, x_2, ..., x_e$ – масштабные факторы. Тогда масштаб субъекта налогообложения (СН) определяем по формуле $r = r(x) = (\sum_{k=1}^{e} a_k x_k^2)^{\frac{1}{2}}$, где a_k – коэффициенты линейной модели. Далее СН ранжируются по масштабу и одномерная совокупность чисел $r_i = (\sum_{k=1}^{n} a_k (x_{ik})^2)^{\frac{1}{2}}$ по методу k-средних разбивается на заданное число кластеров. Тогда каждый кластер $B_j (j = \overline{i.m})$ имеет граничные значения масштабов $z_{j1} = \min r_i$ при $X_i \in B_j$ и $z_{j2} = \max r_i$ при $X_i \in B_j$. Далее строим расширение $\tilde{B}_j$ кластеров B_j путем присоединения заданного числа, например, 10% ближайших точек из кластеров B_{j-1} и B_{j+1}, через $\tilde{r}_{j1}$ и $\tilde{r}_{j2}$ обозначим нижнюю и верхнюю границы расширенного кластера $\tilde{B}_j$. Расширение кластеров делается для того, чтобы построить непрерывную модель за счет сращивания подмоделей на соседних кластерах. В случае экспертного выбора числа кластеров, как показали эксперименты, наиболее рациональной является следующая стратегия. Для однородной группы СН полагаем $m = 1$, для группы СН с небольшой неоднородностью полагаем $m = 2$, а для группы СН, сильно отличающихся по масштабам, полагаем $m = 3$. То есть в последнем случае мы разбиваем их на мелкие, средние и крупные.

Основной этап

Основной этап построения модели реализуется в несколько подэтапов:

a) *Построение линейной модели на каждом расширенном кластере* $\tilde{B}_j (j = \overline{i,m})$.

b) *Построение предварительной модели.*

На каждом расширенном кластере по формуле (1) строится предварительная модель, где $\vec{X}_b = \tilde{B}_j$, а $\omega(X, X_i)$ определяется по формуле:

$$\omega(X, X_i) = (\sum_{k=1}^{n} \frac{a_{jk}^2}{1 + b \cdot \sum (x_k - x_{ik})^2})^{\frac{1}{2}}.$$

Таким образом построенная функция осреднения учитывает силу влияния каждого фактора $x_k (k = \overline{1,n})$.

Выбор параметра b осуществляется путем минимизации функции $d^2(b)$, определяемой по формуле (3).

c) *Стыковка моделей на пересечении кластеров.*

Достаточно описать стыковку кластеров с номерами j и $j+1$. Если $\tilde{r}^2_{(j-1)B} \le r(x) \le \tilde{r}^2_{(j+1)H}$, полагаем

$$\hat{Y}(x) = \hat{Y}(x, \tilde{B}_j).$$

Если $\tilde{r}_{jB} \le r(x) \le \tilde{r}^2_{(j+2)H}$, полагаем

$$\hat{Y}(x) = \hat{Y}(x, \tilde{B}_{j+1}).$$

В случае, если $\tilde{r}_{(j+1)H} \le r(x) \le \tilde{r}_{jB}$, считаем, что

$$\hat{Y}(x) = z_j + (z_{j+1} - z_j) \cdot \frac{r - r_{(j+1)H}}{r_{jB} - r_{(j+1)H}}.$$

Здесь $j = \overline{2, m-1}$, $z_j = \hat{Y}(\tilde{r}_{(j+1)H} \cdot \frac{x}{r(x)}, \tilde{B}_j)$, а $z_{j+1} = \hat{Y}(\tilde{r}_{jB} \cdot \frac{x}{r(x)}, \tilde{B}_{jH})$.

d) *Формирование обучающей выборки.*

Кроме экспертного способа отделения аномальных наблюдений метод предусматривает аналитический способ выделения аномальных наблюдений. В упрощенной форме он состоит в том, что из генеральной совокупности исключается либо определенный процент наблюдений, имеющих наибольшее относительное отклонение

$$\delta_i = \frac{Y_i - \hat{Y}_i}{\hat{Y}_i}$$

расчетных значений от фактических, либо отбрасываются все наблюдения, относительная ошибка δ_j, для которых больше определенной величины. Оставшиеся наблюдения образуют выборку. В случае необходимости список наблюдений, вошедших в выборку, может быть откорректирован экспертно.

e) *Построение модели.*

Построение модели осуществляется повторением этапов a)-c), но уже не на генеральной совокупности, а на выборке. В случае, когда число кластеров задается экспертно, построение модели заканчивается.

Для выбора оптимального числа кластеров требуется минимизировать функционал:

$$\Phi(k) = \sum_{j=1}^{k} \frac{d_j^2(\hat{B}_j)}{\hat{n}_j}, \qquad (4)$$

где $\hat{B}_j$ – совокупность наблюдений из кластера $\tilde{B}_j$, оставшихся в выборке, $\hat{n}_j$ – число наблюдений из $\hat{B}_j$, а $d_j(\hat{B}_j)$ – функция вида (3), рассчитанная по множеству $\hat{B}_j$ при оптимальном (минимизирующем) параметре b_j.

3. ТЕСТИРОВАНИЕ МОДЕЛИ

Тестирование модели проводилось двумя способами. Первый способ состоит в тестировании на модельных примерах. Суть данного подхода состоит в следующем. Выберем некоторую функцию $y = f(x)$ от одной или нескольких переменных $x = (x_1, x_2, \ldots, x_n)$ в определенной области, например, в квадрате. В заданной области выбираем случайным образом точки X_i $(i = 1, \ldots, N)$, далее множество значений $Y_i = f(X_i)$ возмущаем с помощью случайной величины ξ с математическим ожиданием равным нулю, т. е. от y_i перейдем к $Y_i = y_i + \xi_i$. По кортежу данных (X_i, Y_i) строится модель, целью которой является установить регрессионную зависимость Y от X. Ясно, что в данном случае известна точная регрессионная зависимость $E(Y(x)) = f(x)$, поэтому отклонение приближенно найденной по предлагаемой модели зависимости от точной регрессионной зависимости можно явно оценить.

Другой подход к тестированию модели состоит в сравнении расчетов по одному статистическому материалу, полученных с помощью нейросетевой модели и модифицированной непараметрической модели.

Вычислительные эксперименты проводились на линейных и нелинейных функциях $f(x)$ при различном числе переменных x. Выявлена экспериментальная зависимость точности модели от числа наблюдений N и стохастичности выходного показателя Y (т. е. от величины $\sigma(\xi)$).

Расчеты показали, что при $n = 5$, $\sigma(\xi) = 1$ и $N = 100$ для всех исследуемых функций погрешность модели $\max_x |f(x) - \hat{f}(x)|$ оказалась не более 10% при стохастичности показателя Y в 100%. Таким образом, можно сказать, что при 100 наблюдениях ошибка модели дает не более 10% суммарной ошибки.

Исследовалась также устойчивость модели к ошибкам в задании объясняющих переменных (факторов x) и выходного показателя Y.

ВЫВОДЫ

Сравнение с нейросетевой моделью проводилось на примере расчета налога на добавленную стоимость для группы строительных организаций. Расчеты для предприятий среднего масштаба оказались примерно одинаковые, а для мелких и крупных модифицированная непараметрическая модель дает более точные значения. Кроме того, эта модель более устойчивая к изменению выборки и генеральной совокупности.

Предлагаемый подход к построению модели реализован в рабочей программе [6]. Применение этой программы к отбору налогоплательщиков для выездных налоговых проверок успешно апробировано.

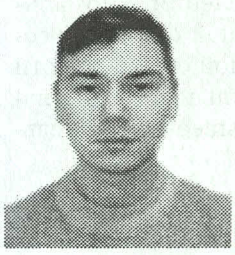
СПИСОК ЛИТЕРАТУРЫ

1. **Айвазян, С. А.** Прикладная статистика: Исследование зависимостей / С. А. Айвазян, И. С. Енюков, Л. Д. Мешалкин. – М.: Финансы и статистика. 1985. 487с.
2. **Дукарский, О. М.** Некоторые применения непараметрических оценок регрессии /

О. М. Дукарский, Б. Я. Левитан // Многомерный статистический анализ в социально-экономических исследованиях. М.: 1974, С.30-37.

3. **Букаев, Г. И.** Моделирование системы налогового контроля на основе нейросетевых информационных технологий / Г. И. Букаев, Н. Д. Бублик, С. А. Горбатков, Р. Ф. Саттаров. – М.: Наука, 2001. 344с.
4. **Бублик, Н. Д.** Теоретические основы разработки технологии налогового контроля и управления / Н. Д. Бублик, И. И. Голичев, С. А. Горбатков, А. В. Смирнов. – Уфа: РИО БашГу. 2004. 336с.
5. **Голичев, И. И., Вариков, А. А.** Аппроксимация регрессионной зависимости (программа для ЭВМ). Зарегистрирована в реестре программ для ЭВМ 10.01.2006 № 2006610133

ОБ АВТОРЕ



Вариков Андрей Александрович, дипл. математик (УГАТУ, 2003). Оконч. аспирантуру каф. ВВТиС УГАТУ (2006). Иссл. в области регрессионного анализа.