МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО»

Кафедра Математической экономики

Непараметрическое оценивание экономических зависимостей

АВТОРЕФЕРАТ МАГИСТЕРСКОЙ РАБОТЫ

студентки 2 курса 247 группы

направление 09.04.03 – Прикладная информатика

механико-математического факультета

Худошиной Анастасии Олеговны

Научный руководитель
доцент, к.ф. м.н _____ В.В. Новиков

Зав. кафедрой
зав.каф., д.ф. м.н _____ С.И. Дудов

Саратов 2019

Введение. Под непараметрическими методами понимают статистические приемы, которые не требуют спецификации функциональных форм оцениваемых объектов. В таких методах данные сами определенным образом создают модель. Этот подход называется «непараметрическая регрессия».

Актуальность непараметрических методов обусловлена тем, что при анализе данных чисто параметрический подход к задачам оценивания не обладает необходимой гибкостью. Непараметрические методы имеют важную особенность, а именно, происходит ослабление параметрических предпосылок (накладываемых на процесс), и появляется возможность данным самим установить требуемую модель.

Предметом исследования являются непараметрические и полупараметрические ядерные методы. Цель работы - исследовать набор данных с применение пакета np, а именно, провести сравнительный анализ моделей. Для анализа данных мы использовали среду R. R - это мощный язык для статистических вычислений и графиков, который может справиться с любой задачей в области обработки данных Работа состоит из введения, двух глав, заключения, списка использованных источников и приложения.

Математические аспекты непараметрического оценивания регрессионных моделей. Рассмотрим произвольную случайную величину X, имеющую плотность $f(x)$. Предположим, что имеется IID-выборка из неизвестного распределения, и требуется смоделировать ее функцию плотности, $f(x)$.

Введем одномерную ядерную оценку плотности.

$$\widehat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) \qquad (2).$$

Оценку (2) называют оценкой Розенблатта-Парзена.

Свойства ядерной оценки одномерной плотности:

$$K(z) \geq 0, \quad \int K(z)dz = 1, \quad \int zK(z)dz = 0, \quad \int z^2K(z)dz = k_2 < \infty.$$

Обозначим через $f = f(x)$ непрерывную функцию плотности случайной величины X в точке x, и пусть $x_1, ..., x_n$- наблюдения из f. Оценка плотности

ядра $\widehat{f}(x)$ может быть переписана как

$$\widehat{f} = \widehat{f}(x) \sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right) = \frac{1}{n} \sum_{i=1}^{n} w_i,$$

$$w_i = w_{ni}(x) = \frac{1}{h} K\left(\frac{x_i - x}{h}\right),$$

где K - функция ядра, h - ширина окна, $w_{ni}(x)$ - вес функция, которая зависит от расстояния x_i от x и размера выборки h. Сделаем следующие предположения:

1) Наблюдения $x_1, ..., x_n$ независимы и одинаково распределенные.

2) Ядро K является симметричной функцией вокруг нуля, удовлетворяющей
   а) $\int K(z)\ dz = 1$,
   б) $\int z^2 K(z)\ dz = \mu_2 \neq 0$,
   в) $\int K^2(z) dz < \infty$.

3) Производные второго порядка f непрерывны и ограничены в некоторых окрестности x.

4) $h = h_n \to 0,\ n \to \infty$.

4) $nh_n \to \infty,\ n \to \infty$.

Главная задача ядра – обеспечить гладкость и дифференцируемость результирующей оценки. Впервые такое ядро было предложено Епанечниковым в 1969 году.

$$K_e(z) = \begin{cases} \frac{3}{4\sqrt{5}}(1 - \frac{1}{5}z^2), & -\sqrt{5} \leqslant z \leqslant \sqrt{5} \\ 0, & z > \sqrt{5} \cup z < -\sqrt{5} \end{cases}$$

Для проведения качественного непараметрического оценивания необходимо выбрать подходящую ширину окна. Хотя ядерная функция остается важной, ее главная роль состоит в обеспечении дифференцируемости и гладкости получающейся оценки. Ширина окна, определяет поведение оценки в конечных выборках, что ядерная функция сделать просто не в состоянии. Существуют четыре общих подхода к выбору ширины окна: 1) референтные эвристические правила, 2) методы подстановки, 3) методы кросс-валидации

и 4) бутстраповские методы. Заметим, что данными методы выбора ширины окна не всегда гарантируют хороший результат.

Кросс-валидация на основе наименьших квадратов - это полностью автоматический и диктуемый данными метод выбора сглаживающего параметра. Этот метод основан на принципе выбора ширины окна, минимизирующей интегральную среднеквадратическую ошибку получающейся оценки. Интеграл квадрата разности $\widehat{f}(x)$ и $f(x)$ имеет вид

$$\int [\widehat{f}(x) - f(x)]^2 dx = \int \widehat{f}(x)^2 dx - 2\int \widehat{f}(x) f(x) dx + \int f(x)^2 dx.$$

Кросс-валидация на основе правдоподобия дает оценку плотности, имеющую интерпретацию в терминах энтропии, а именно: оценка будет близка к истинной плотности в смысле Кулбака–Лайблера. Кросс-валидация на основе правдоподобия выбирает h, чтобы максимизировать логарифм функции правдоподобия (построенной по всей выборке за исключением одного наблюдения), имеющей вид

$$\Im = \log L = \sum_{i=1}^{n} \log \widehat{f}_{-i}(x),$$

где $\widehat{f}_{-i}(x)-$ ядерная оценка $f(X_i)$, построенная по всей выборке за исключением одного наблюдения X_i , то есть

$$\widehat{f}_{-i}(x) = \frac{1}{(n-1)h} \sum_{j=1, j\neq i}^{n} K\left(\frac{X_j - x}{h}\right).$$

Пусть имеется смесь дискретных и непрерывных данных, и необходимо смоделировать их совместную функцию плотности. В таком случаи ученные традиционно используют ядерные методы, прибегая к частотному подходу. В этот методе непрерывные данные разбиваются на подмножества в соответствии с реализациями дискретных данных. С ростом числа подмножеств количество данных в каждой клетке уменьшается, что приводит к проблеме редких "данных". Из-за недостаточного количества данных трудно оценить плотность, т.е. будет большая дисперсия. Поэтому лучше рассмотрим дру-

гой подход, который использует концепцию Обобщенных мультипликативных ядер. Для непрерывных переменных используем стандартные непрерывные ядра $W(.)$. В случае упорядоченной дискретной переменной можно использовать ядро вида:

$$\tilde{l}(\widetilde{X_i^d}, \widetilde{x^d}) = \begin{cases} 1-\lambda, & \widetilde{X_i^d} = \widetilde{x^d} \\ \frac{1-\lambda}{2}\lambda|\widetilde{X_i^d} - \widetilde{x^d}|, & \widetilde{X_i^d} \neq \widetilde{x^d} \end{cases}$$

Обощенное мультипликативное ядро для одной непрерывной, одной неупорядоченной дискретных переменных запишутся в следующем виде:

$$K(.) = W(.) \times \bar{l}(.) \times \widetilde{l}(.) \tag{8}$$

Применяя данный подход можно расширить сферу применения ядерных методов. Оценивание совместной функции плотности, на смешанных данных осуществляется при помощи обобщенных мультипликативных ядер.

В основе многих статистических объектов лежит функция условной плотности, но при этом их обычно не моделируют напрямую в рамках непараметрических моделей. Рассмотрим ядерное оценивание функции условной плотности. Пусть $f(.)$ - совместная плотность (X, Y), а $\mu(.)$ - маргинальная плотность X, где X и Y могут быть непрерывными, неупорядоченными и упорядоченными переменными. Предположим. что $X-$ независимая, $Y-$ зависимая переменная. Обозначим $\widehat{f}$ и $\widehat{\mu}$ для соответствующих ядерных оценок, и запишем условную плотность

$$g(y|x) = f(x, y)/f(x)$$

как

$$\widehat{g}(y|x) = \widehat{f}(x, y)/\widehat{f}(x). \tag{9}$$

Пусть $F(y|x)$ - функция распределения Y при $X = x$, а $f(x)$ - функция маргинальной плотности X. Оценим $F(y|x)$ как

$$\widehat{F}(y|x) = \frac{n^{-1}\sum_{i=1}^{n} G(\frac{y-Y_i}{h_0})K_h(X_i, x)}{\widehat{f}(x)}, \tag{10}$$

где $G(.)$ - ядерная функция распределения, например стандартного нормального распределения, h_0- параметр сглаживания, соотвествующий Y, а $K_h(X_i, x)$ - мультипликативное ядро определенное в (8) где каждое одномерное непрерывное ядро поделено на соответствующую ширину окна.

Самый известный метод непараметрической ядерной оценки регрессии был предложен Надараи и Уотсоном. По определению, условное среднее непрерыной случайной величины Y равно

$$g(x) = \int yg(y|x)dy = \int y\frac{f(y,x)}{f(x)}dy = \frac{m(x)}{f(x)},$$

где $g(y|x)$ - функция условной плотности, а $m(x) = \int yf(y,x)dy$.

Локально постоянная оценка условного среднего получается:

$$\widehat{g}(x) = \frac{\sum_{i=1}^{n} Y_i K(\frac{X_i - x}{h_x})}{\sum_{i=1}^{n} K(\frac{X_i - x}{h_x})}. \qquad (11).$$

К наиболее востребованным методам гибкого оценивания относится полупараметрический метод. Такие модели возникают при комбинировании непараметрических и параметрических моделей определенным способом. Полупараметрические модели применимы в случаях, когда полностью непараметрические модели срабатывают не очень хорошо, например, если проклятие размерности ведет к большой вариабельности оценок, или исследователь использует параметрическую модель регрессии, но не знает плотность распределения ошибок.

Существует достаточно много полупараметрических методов, в работе мы рассмотрели три популярных метода, а именно: одноиндексные модели и модели с переменными коэффициентами, и частично линейные.

Полупараметрическая частично линейная модель задается следующим образом:

$$Y_i = X_i'\beta + g(Z_i) + u_i, \quad i = 1, ..., n, \qquad (12)$$

где $X_i = p \times 1$ вектор, $\beta - p \times 1$ вектор неизвестных параметров, и $Z_i \in R^q$. Функциональная форма $g(.)$ не специфицирована. Конечномерный параметр β составляет параметрическую часть модели, а неизвестная функ-

ция $g(.)$ - непараметрическую часть. Предполагается, что данные являются IID с $E[u_i|X_i, Z_i] = 0$, а процесс для ошибок условно гетероскедастичен с $E[u_i^2|x, z] = \sigma^2(x, z)$ неизвестной формы.

Полупараметрическая индексная модель имеет вид

$$Y = g(X'\beta_0) + u, \tag{13}$$

где Y - зависимая переменная, $X \in R^q$ - вектор объясняющих переменных, $\beta_0 - q \times 1$ вектор неизвестных параметров, и u - ошибка, удовлетворяющая $E[u|X] = 0$. Член $X'\beta_0$ называется одиночным индексом, так как это одномерная величина, хотя X - вектор. Функциональная форма $g(.)$ неизвестна. Модель является полупараметрической по природе, так как функциональная форма линейного индекса специфицирована, а $g(.)$ - нет.

Модель с гладкими коэффициентами была предложена в работе и имеет следующий вид:

$$Y_i = \alpha(Z_i) + X_i'\beta(Z_i) + u_i = (1 \ X_i') \begin{pmatrix} \alpha(Z_i) \\ \beta(Z_i) \end{pmatrix} = W_i'\gamma(Z_i) + u_i. \tag{14}$$

где $X_i - k \times 1$ вектор, а $\beta(z)-$ вектор неспецифицированных гладких функций от z.

Подогнанная модель имеет вид:

$$Y_i = \widehat{Y}_i + \widehat{u}_i = W_i'\gamma(Z_i) + \widehat{u}_i.$$

Рассмотрим непараметричесую регрессионную модель

$$Y_i = m(X_i) + \epsilon_i, i = 1, ..., n,$$

где $m(x) = E(Y|X = x)$ - неизвестная функция регрессии, которая оценивается на основе эмпирических данных $\{(Xi, Yi)\}_{i=1}^n$, $\{\epsilon_i\}_{i=1}^n$ - случайные ошибки.

Пусть, функция регрессии может быть представлена в виде ряда Фурье

$$m(x) = \sum_{j=0}^{\infty} \beta_j \varphi_j(x), \tag{15}$$

где $\{\varphi_j(x)\}_{j=0}^{\infty}$ - известная система базисных функций, а $\{\beta_j\}_{j=0}^{\infty}$ - неизвестные коэффициенты Фурье.

Пусть $P_j(x),\ j = 0, 1, ..., n$ - многочлены Якоби, ортонормированные на отрезке $[-1, 1]$ с весом

$$\rho(x) = (1-x)^{\alpha}(1+x)^{\beta}, \alpha, \beta > -1$$

и пусть $p := min\{\alpha; \beta\},\ q := max\{\alpha; \beta\}$. Через $Lip\ M\alpha,\ 0 < \alpha \leqslant 1$, обозначается класс функций f , удовлетворяющих на отрезке $[-1, 1]$ условию Липшица порядка α с константой $M > 0$. Через $C(p),\ C(p, q)$ будем обозначать постоянные, зависящие от одного или нескольких параметров.

Лемма 1. При условии $p \geqslant -\frac{1}{2}$ имеет место следующее весовое неравенство для ортонормированных многочленов Якоби

$$(1-x)^{\frac{\alpha}{2}+\frac{1}{4}}\ (1+x)^{\frac{\beta}{2}+\frac{1}{4}}\ \left|P_j^{(\alpha,\beta)}(x)\right| \leqslant C(\alpha, \beta), x \in [-1; 1]$$

или

$$\rho(x)\ \left|P_j^{(\alpha,\beta)}(x)\right| \leqslant C(\alpha, \beta)(1-x)^{\frac{\alpha}{2}-\frac{1}{4}}\ (1+x)^{\frac{\beta}{2}-\frac{1}{4}}.$$

Лемма 2. В каждой точке $x \in (-1, 1)$ для ортонормированных многочленов Якоби справедливо соотношение

$$P_j^{(\alpha,\beta)} = o_x(1), \tag{16}$$

которое выполняется равномерно на каждом интервале $[a, b] \subset (-1, 1)$.

Лемма 3. Положим для $n \geqslant 2$

$$A_n(\alpha, \beta) := \begin{cases} n, & p > 1/2 \\ n/\log n, & p = 1/2 \end{cases}$$

Тогда при $\alpha, \beta \geqslant -1/2$ справедливо

$$\sum_{i=1}^{n} \left(\int_{A_i} P_j^{(\alpha,\beta)}(x)\ \rho(x)dx \right)^2 \leqslant \frac{C(\alpha,\beta)}{A_n(\alpha,\beta)}.$$

Теорема 1. Пусть выполнены условия:

$$1)\ E\varepsilon_i = 0,\ E(\varepsilon_i\varepsilon_j) = 0,\ i \neq j,\ E\varepsilon_i^2 < C;$$

$$2)\ m(x) \in Lip_{M^1};$$

$$3)\ p = min\{\alpha; \beta\} \geqslant -\frac{1}{2};$$

$$4)\ N^2(n) = o\{A_n(\alpha; \beta)\}. \tag{17}$$

Тогда при $N(n) \to \infty$ имеем

$$\widehat{m}_{N(n)}(x) \to m(x),\ x \in (-1; 1).$$

Теорема 2. Пусть выполнены условия 1)-3) предыдущей теоремы, $q < \frac{1}{2}$, и, кроме того,

$$(N(n))^{2q+3} = o\{A_n(\alpha; \beta)\}, n \to \infty. \tag{18}$$

$$\widehat{m}_{N(n)}(x) \to m(x),\ x \in [-1; 1].$$

Оценивание непараметрических регрессий в среде R. Рассмотрим применение пакета np для непараметрических моделей. В выпускной квалификационной работе мы исследовали данные о заработной плате в Канаде. Данные состоят из случайной выборки, взятой из документов о переписи населения Канады 1971 года. Имеем 205 наблюдений и 2 переменные: логарифм заработной платы человека (logwage) и его возраст (age). С использованием встроенных функций в R, мы получили подходящую ширину окна и оценили непараметрическую модель, а так же рассчитали прогнозы, и для сравнения построили параметрическую модель.

Далее мы анализировали данные о почасовой оплате труда. Мы имеем 526 наблюдений. В нашем примере зависимой переменной является заработная плата (lwage), в то время как объясняющие переменные включают

три непрерывные переменные, а именно число лет образования (educ), количество лет потенциального опыта (exper) и количество лет с их нынешним работодателем (tenure) вместе с двумя качественными переменными, female («Женщина» / «Мужчина») и married («Женат» / «Не замужем»). Результат получился следующим: мужчины имеют ожидаемую заработная плата выше, чем у женщин. Существенной разницы нет между ожидаемой заработной платой состоящих в браке и не состоящих в браке лиц, а образование и стаж сначала растут, а затем падают по мере накопления опыта, при прочих равных условиях. Для этих данных были построены 4 модели: 1) полупараметрическая частичная линейная модель; 2) параметрическая модель; 3) непараметрическая модель; 4) полупараметрической модель с переменным коэффициентом. Итог следующий: полупараметрическая частичная линейная модель R^2=44,9%, немного лучше параметрической модели R^2=40,4%, и полупараметрической модель с переменным коэффициентами R^2=42,6%, но полностью непараметрическая модель R^2=51,5% имеет большее преимущество.

Следующим объетом исследовния стали данные динамики ВВП Италии для 21 региона за период 1951 по 1998г. Всего 1008 наблюдений и две переменные: ВВП и год. Исходя из природы данных, будем считать ВВП непрерывной переменной, а год - упорядоченная дискретная переменная. Ширина окна вычислялась с помощью кросс-валидации на основе правдоподобия. Итог: распределение доходов изменилось с унимодального в начале 1950-х годов до бимодального в 1990-х годах. Этот результат является устойчивым к выбору ширины окна и наблюдается независимо от того применяется ли кросс-валидация на основе наименьших квадратов или правдоподобия. Метод ядра раскрывает тенденцию, которую можно легко упустить, при использовании параметрической модели распределения доходов.

Заключение. Методы ядерного сглаживания применяются исследователями-практиками в целом ряде дисциплин. Непараметрические ядерные подходы используют при неправильной спецификации параметрической модели. Привлекательность непараметрических методов состоит, главным образом, в их устойчивости к неправильной спецификации функциональных форм, в отличие от параметрических моделей.

Мы построили простую параметрическую линейную регрессию, локально-линейную непараметрическую регрессию и полупараметрическую частичную линейную регрессию, и провели сравнение по критерию R^2. Результат оказался следующим: полупараметрическая частичная линейная модель R^2=44,9% , немного лучше параметрической модели R^2=40,4%, тогда как полностью непараметрическая модель R^2=51,5% превосходит как параметрическую модель так и частичную линейную модель. Кроме того, рассмотрели модель с изменяющимся коэффициентом и по критерию R^2 результат составил R^2=42,6%. Для панельных данных ВВП Италии были рассчитаны непараметрические оценки условной плотности и условного распределения, а так ж непараметрические оценки условных квантилей.

Заметим, что непараметрические ядерные методы могут быть вычислительно трудными, особенно при обработке большого объема данных. Это происходит из-за того, что на практике необходимо применение диктуемых данными методов выбора ширины окна, а скорость выполнения таких алгоритмов экспоненциально растет с объемом доступных данных. Несмотря на то, существуют методы приближенных вычислений, которые имеют возможность значительно снизить объем необходимых вычислений.