Непараметрическая эконометрика: вводный курс*

Джеффри Расин†

Университет МакМастер, Гамильтон, Канада

Настоящее эссе является вводным курсом для тех, кто хотел бы познакомиться с непараметрической эконометрикой. Хотя теория, лежащая в основе многих из рассматриваемых методов, может показаться устрашающей для практика, в эссе показано, как применять ряд непараметрических методов весьма непосредственным образом. Вместо энциклопедического покрытия всей области, внимание в эссе ограничено набором основополагающих разделов, а для иллюстративных целей широко используются примеры. Описаны ситуации, когда моделируются данные, состоящие из непрерывных, дискретных или категориальных (номинальных или порядковых) переменных, или любой их комбинации. Также рассматриваются недавние разработки для случаев, когда некоторые из включенных в модель переменных на самом деле могут быть несущественными, что значительно меняет поведение оценок и оптимальной ширины окон по сравнению с общепринятыми подходами.

1 Введение

Непараметрические методы представляют собой статистические приемы, которые не требуют спецификации функциональных форм оцениваемых объектов. Вместо этого данные сами определенным способом формируют модель. В рамках модели регрессии этот подход известен как «непараметрическая регрессия» или «непараметрическое сглаживание». Методы, которые рассматриваются в настоящем эссе, известны как ядерные[1] методы. Подобные методы становятся все более популярными в прикладном анализе данных; они лучше всего подходят для ситуаций, когда имеется большой объем данных, а число участвующих переменных довольно мало. Эти методы часто применяют, если обычные параметрические спецификации оказываются неподходящими для решения поставленной задачи, особенно когда формальное отвержение параметрической модели тестом на правильность спецификации не указывает направление поиска более приемлемой параметрической модели. Привлекательность непараметрических методов в том, что они ослабляют параметрические предпосылки, накладываемые на процесс, порождающий данные, и позволяют данным самим определить подходящую модель.

Непараметрические и полупараметрические методы привлекли большое внимание статистиков за несколько прошедших десятилетий, что подтверждается обширной статистической литературой на эту тему, в частности, работами Prakasa Rao (1983), Devroye & Györfi (1985), Silverman (1986), Scott (1992), Bickel, Klaassen, Ritov & Wellner (1993), Wand & Jones (1995), Fan & Gijbels (1996), Simonoff (1996), Azzalini & Bowman (1997), Hart (1997), Efromovich (1999), Eubank (1999), Ruppert, Carroll & Wand (2003), Härdle, Müller, Sperlich & Werwatz (2004) и Fan & Yao (2005). Тем не менее, относительно мало работ, нацеленных на потребности прикладных эконометристов; среди таковых на текущий момент нам известны работы Härdle (1990), Horowitz (1998), Pagan & Ullah (1999), Yatchew (2003) и Li & Racine (2007a).

*Перевод Б. Гершмана и С. Анатольева. Цитировать как: Расин, Джеффри (2008) «Непараметрическая эконометрика: вводный курс», Квантиль, №4, стр. 7–56. Citation: Racine, Jeffrey S. (2008) “Nonparametric econometrics: a primer,” Quantile, No.4, pp. 7–56.

†Адрес: McMaster University, 1280 Main Street West, L8S 4M4, Hamilton, Canada. Электронная почта: *racinej@mcmaster.ca*

[1] Ядро – это просто взвешивающая функция.

Первая опубликованная статья о ядерном оценивании вышла в 1956 г. (Rosenblatt, 1956), а сама идея была предложена в техническом докладе военно-воздушных сил США (USAF) как средство освобождения дискриминантного анализа от необходимости строгой параметрической спецификации (Fix & Hodges, 1951). С того времени данная область испытала экспоненциальный рост и стала неотъемлемой частью учебников бакалаврского уровня (см., например, главу 11 в Johnston & DiNardo, 1997), что свидетельствует о популярности этих методов как среди студентов, так и среди исследователей.

Хотя ядерные методы популярны, они являются лишь одним из многих подходов к построению гибких моделей. Такие подходы включают, например, сглаживание сплайнами, методы ближайших соседей, нейронные сети и гибкие методы сглаживания с помощью рядов. Но в настоящем эссе мы ограничимся классом непараметрических ядерных методов, а также затронем полупараметрические ядерные методы. Мы также сконцентрируемся на более прикладных аспектах этих методов, чтобы удержать объем эссе в разумных пределах; заинтересованный читатель может обратиться к работе Li & Racine (2007a) и к упомянутым выше источникам за подробностями теоретических оснований рассматриваемых методов.

Стоит отметить, что часто приходится слышать две жалобы по поводу непараметрических ядерных методов, а именно: 1) отсутствие соответствующего программного обеспечения и 2) вычислительные трудности, связанные с этими методами. Мы, конечно же, согласны с обеими. Последняя из них неизбежна и лежит, так сказать, в самой природе методов, хотя в разделе *Соображения по поводу вычислений* обсуждаются некоторые возможности. Однако первая проблема постепенно решается, и последние разработки дают надежду на прорыв в вычислительных возможностях. Многие статистические программные пакеты уже содержат некоторые элементарные непараметрические методы (оценивание одномерной плотности, парной регрессии), хотя они часто используют эвристические правила для выбора ширины окна, которые, будучи вычислительно простыми, могут не быть робастными во всех приложениях. Недавно для программной среды `R` был разработан пакет `np` (R Development Core Team, 2007), который содержит простой в использовании и открытый код для ядерного оценивания; заинтересованный читатель может обратиться за подробностями к работе Hayfield & Racine (2007). Все примеры в данном эссе были реализованы с помощью пакета `np` (код для повторения результатов доступен по запросу).

2 Оценивание плотности и функции вероятности

Обозначения и основные подходы, рассматриваемые в этой главе, представляют основу для остальных глав, и будут использоваться на протяжении всего эссе. В данном разделе дается больше деталей, чем в других, поскольку изрядная доля ключевых понятий, таких как «обобщенное мультипликативное ядро», ядра для категориальных данных, диктуемый данными выбор ширины окна и другие, полезна для понимания дальнейшего материала.

Читатель, несомненно, близко знаком с двумя популярными непараметрическими оценками, а именно: гистограммой и частотной оценкой. Гистограмма является негладким непараметрическим методом, который можно использовать для оценивания функции плотности распределения (ФПР) непрерывной случайной величины. Частотная оценка вероятности – это негладкий непараметрический метод для оценивания вероятности дискретных событий. Хотя негладкие методы могут на самом деле быть мощными, у них есть свои недостатки. Для глубокого изучения ядерного оценивания плотности заинтересованный читатель может обратиться к великолепным монографиям Silverman (1986) и Scott (1992), а для оценивания плотности в случае данных смешанного типа – к работе Li & Racine (2007a) и содержащимся в ней ссылкам на литературу. Начнем с иллюстративного *параметрического* примера.

2.1 Параметрические оценки плотности

Рассмотрим произвольную случайную величину X, имеющую плотность $f(x)$, где $f(\cdot)$ – объект интереса. Предположим, что имеется IID-выборка из неизвестного распределения, и требуется смоделировать его функцию плотности, $f(x)$. Это типичная ситуация, в которой оказывается исследователь-практик.

Для данного примера сгенерируем $n = 500$ реализаций, но тут же «забудем» вид истинного процесса, порождающего данные (DGP), делая вид, будто нам неизвестно, что данные получены из смеси нормальных распределений ($N(-2; 0{,}25)$ и $N(3; 2{,}25)$ с одинаковой вероятностью). Затем (наивно) предположим, что данные получены из, скажем, параметрического семейства нормальных распределений, а именно:

$$\hat{f}(x) = \frac{1}{\sqrt{2\pi\hat{\sigma}^2}} \exp\left\{-\frac{1}{2}\left(\frac{x-\hat{\mu}}{\hat{\sigma}}\right)^2\right\}.$$

Теперь оценим эту модель и получим $\hat{\mu} = 0{,}56$ и $\hat{\sigma} = 2{,}71$. Далее, как всегда рекомендуется, проведем тест на правильность спецификации, используя, скажем, тест Шапиро–Уилкса, и получим $W = 0{,}88$ с p-значением $< 2{,}2\times10^{-16}$, тут же отвергая эту параметрическую модель. Оцененная модель и истинный DGP изображены на Рис. 1.

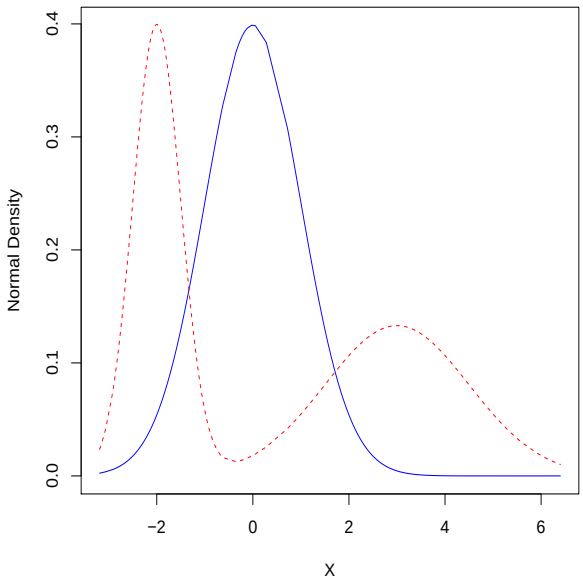

Рис. 1: $N(0{,}56; 2{,}71^2)$-оценка плотности (унимодальная, сплошная линия) и истинный процесс, порождающий данные (бимодальный, пунктирная линия).

Поскольку данная популярная параметрическая модель уверенно отвергается данными, есть две альтернативы, а именно: 1) искать более подходящую параметрическую модель или 2) использовать более гибкие оценки.

В дальнейшем будем предполагать, что читатель оказался именно в такой ситуации. То есть он честно применил параметрический метод и провел ряд тестов на адекватность модели, которые показали, что параметрическая модель не согласуется с DGP. Далее он обращается к более гибким методам оценивания плотности. Заметим, что, хотя в данный момент речь идет об оценивании плотности, точно так же можно было бы обсуждать почти любой параметрический подход, например, регрессионный анализ.

2.2 Гистограммы и ядерные оценки плотности

Построить гистограмму просто. Сначала строится ряд ячеек (возьмем x_0 за начало координат и h за ширину ячейки). Ячейки – это промежутки $[x_0 + mh, x_0 + (m+1)h)$ для положительных и отрицательных целых чисел m. Гистограмма определяется как

$$\hat{f}(x) = \frac{1}{n} \frac{\text{количество } X_i \text{ в одной ячейке с } x}{\text{ширина ячейки, содержащей } x} = \frac{1}{nh} \sum_{i=1}^{n} \mathbb{I}[X_i \text{ в одной ячейке с } x], \tag{1}$$

где $\mathbb{I}[A]$ – индикаторная функция, принимающее значение 1, если A верно, и 0 в противном случае. Исследователь должен выбрать начало координат и ширину ячейки, и получаемая оценка чувствительна к обоим параметрам выбора. Обычно для выбора обоих параметров используются эвристические правила. Хотя данная конструкция является крайне мощной, ее можно значительно улучшить. Гистограмма не является особо эффективной оценкой, говоря статистическим языком. Она разрывна, а значит, любой основанный на ней метод, требующий дифференцирования, будет недоступен в силу этого свойства. К тому же она не центрирована вокруг точки, в которой требуется оценить плотность. Хотя гистограмма является замечательным инструментом, ядерные методы представляют альтернативу, к изучению которой мы приступаем.

Одномерная ядерная оценка плотности была предложена для преодоления многих ограничений, связанных с гистограммой. Она всего-навсего заменяет индикаторную функцию в выражении (1) на ядро, симметричную взвешивающую функцию $K(z)$, обладающую рядом полезных свойств. Замена индикаторной функции в (1) на ядерную функцию дает

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

Эту оценку часто называют оценкой Розенблатта–Парзена (Rosenblatt, 1956; Parzen, 1962). На Рис. 2 изображены гистограмма и оценка Розенблатта–Парзена для сгенерированных данных из главы 2.1 с шириной окна, полученной по методу подстановки из Sheater & Jones (1991) (см. п. 2.3.2).

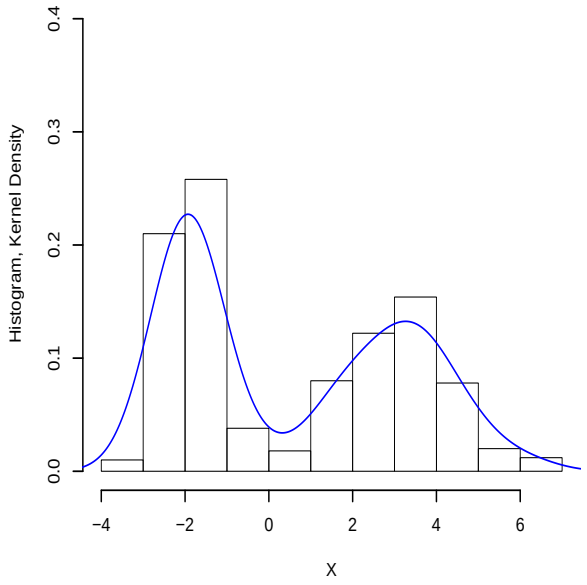

Рис. 2: Гистограмма и ядерная оценка одномерной функции плотности.

Как видно из Рис. 2, и гистограмма, и оценка Розенблатта–Парзена легко раскрывают бимодальную природу данных, в отличие от неправильно специфицированной унимодальной параметрической модели, представленной на Рис. 1. Читатель, сравнив рисунки 1 и 2, сразу заметит, что как гистограмма, так и ядерная оценка, являются *смещенными*, то есть они как бы недооценивают левый пик в конечных выборках, что на самом деле закономерно, как будет видно далее при изучении свойств оценки Розенблатта–Парзена. Но по мере увеличения n и уменьшения h определенным образом, который вскоре будет описан, ядерная оценка сойдется к истинному DGP с единичной вероятностью. Неправильно специфицированная модель никогда не сойдется к истинному DGP. Какой метод дает более адекватное описание DGP – унимодальная параметрическая модель или бимодальная непараметрическая?[2] Этот вопрос обсуждается ниже в главе 2.7.

Ядерное оценивание кумулятивной функции распределения (КФР) получило гораздо меньше внимания, чем оценивание функции плотности. Заинтересованный читатель может обратиться к основополагающей статье Bowman, Hall & Prvan (1998) и главе 1 в Li & Racine (2007a).

2.2.1 Свойства ядерной оценки одномерной плотности

Предположим, что ядерная функция $K(z)$ неотрицательна и обладает следующими свойствами:

$$\int K(z)\,dz = 1, \quad \int zK(z)\,dz = 0, \quad \int z^2K(z)\,dz = \kappa_2 < \infty.$$

Если не указано иное, нижним и верхним пределами интегрирования будут $-\infty$ и ∞ соответственно. Такое ядро часто называют «ядром второго порядка». Parzen (1962) показал, что можно выбирать ядра, позволяющие снизить поточечное смещение $\hat{f}(x)$, однако для этого необходимо отказаться от неотрицательности $K(z)$. Один из недостатков использования подобных ядер «более высокого порядка»[3] при оценивании функции плотности заключается в том, что оценки могут оказаться отрицательными, что, конечно, является нежелательным побочным эффектом. Ядра более высокого порядка иногда встречаются в многомерных случаях для обеспечения скоростей сходимости, необходимых для установления предельных распределений. В дальнейшем по умолчанию используется ядро второго порядка, если не указано иное.

Для оценки свойств многих ядерных методов используется критерий точечной среднеквадратической ошибки (MSE). Чтобы получить выражение для MSE, посчитаем смещение и дисперсию $\hat{f}(x)$. Учитывая, что

$$\mathbb{MSE}[\hat{f}(x)] = \mathbb{E}[\hat{f}(x) - f(x)]^2 = \mathbb{V}[\hat{f}(x)] + (\mathbb{B}[\hat{f}(x)])^2,$$

и используя разложение в ряд Тейлора и замену переменных, получаем асимптотическое смещение

$$\mathbb{B}[\hat{f}(x)] \approx \frac{h^2}{2} f''(x)\kappa_2, \tag{3}$$

и асимптотическую дисперсию

$$\mathbb{V}[\hat{f}(x)] \approx \frac{f(x)}{nh} \int K^2(z)dz. \tag{4}$$

[2] Возможно, уместным здесь будет высказывание Дж. Бокса о том, что «все модели неправильны, но некоторые полезны» (с. 424 в Draper, 1987).

[3] Вообще говоря, ядро порядка ν (где $\nu \geq 2$ – целое число) должно удовлетворять следующим свойствам: $\int K(z)\,dz = 1$, $\int z^l K(z)\,dz = 0$, $(l = 1, \ldots, \nu - 1)$ и $\int z^\nu K(z)dz = \kappa_\nu \neq 0$.

См. подробный вывод этих результатов на с. 23–24 в Pagan & Ullah (1999) или с. 11–12 в Li & Racine (2007a).

Заметим, что как смещение, так и дисперсия, зависят от ширины окна (смещение падает, а дисперсия растет по мере уменьшения h). Смещение также возрастает по $f''(x)$, то есть является наибольшим в пиках распределений. Но, если выполнены условия состоятельности, а именно $h \to 0$ при $n \to \infty$ (смещение $\to 0$) и $nh \to \infty$ при $n \to \infty$ (дисперсия $\to 0$), то смещение, связанное с $f''(x)$, уменьшается при увеличении выборки и в пределе исчезает. Отметим, что величину nh иногда называют «эффективным размером выборки», а требование $nh \to \infty$ при $n \to \infty$ просто означает, что по мере получения большего количества информации ($n \to \infty$) усреднение происходит по более узкой области ($h \to 0$), но в то же время количество «локальной информации» (nh) должно увеличиваться.

Приведенные выше формулы для смещения, дисперсии и среднеквадратической ошибки – это *точечные* свойства, то есть они выполняются в любой точке x. Интегральная среднеквадратическая ошибка (IMSE), с другой стороны, агрегирует среднеквадратические ошибки по всей области определения функции плотности, являясь глобальной мерой ошибки, и при подстановке выражений для асимптотического смещения и дисперсии принимает вид

$$\begin{aligned}
\mathbb{IMSE}[\hat{f}(x)] &= \int \mathbb{MSE}[\hat{f}(x)]dx = \int \mathbb{V}[\hat{f}(x)]\,dx + \int (\mathbb{B}[\hat{f}(x)])^2 dx \\
&\approx \int \left[\frac{f(x)}{nh} \int K^2(z)dz + \left(\frac{h^2}{2} f''(x)\kappa_2 \right)^2 \right] dx \\
&= \frac{1}{nh} \int K^2(z)dz \int f(x)\,dx + \left(\frac{h^2}{2}\kappa_2 \right)^2 \int \left[f''(x) \right]^2 dx = \frac{\Phi_0}{nh} + \frac{h^4}{4}\kappa_2^2 \Phi_1,
\end{aligned} \tag{5}$$

где $\Phi_0 = \int K^2(z)dz$, а $\Phi_1 = \int [f''(x)]^2 dx$. См. подробный вывод этого результата на с. 24 в Pagan & Ullah (1999) или с. 13 в Li & Racine (2007a).

Теперь можно минимизировать это выражение по ширине окна и ядерной функции для получения «оптимальной ширины окна» и «оптимального ядра». Это выражение также дает основу для диктуемого данными выбора ширины окна. Отметим, что, используя IMSE вместо MSE, мы выбираем ширину окна, обеспечивающую хорошую оценку «в целом», а не ту, которая хороша лишь для одной точки.

Минимизируя IMSE по h, получаем ширину окна, которая глобально балансирует смещение и дисперсию:

$$h_{opt} = \Phi_0^{1/5} \kappa_2^{-2/5} \Phi_1^{-1/5} n^{-1/5} = \left[\frac{\int K^2(z)dz}{\left(\int z^2 K(z)dz \right)^2 \int [f''(x)]^2 dx} \right]^{1/5} n^{-1/5} = cn^{-1/5}. \tag{6}$$

Заметим, что константа c зависит от $f''(x)$ и $K(\cdot)$, и если $h \propto n^{-1/5}$, то

$$o\left(\frac{1}{nh} \right) = o\left(\frac{1}{n^{4/5}} \right),$$

то есть использование оптимальной ширины окна дает оценку $\hat{f}(x)$, имеющую IMSE порядка $n^{-4/5}$:

$$\hat{f}(x) - f(x) = O_p\left(n^{-2/5} \right).$$

Отметим, что для *правильно специфицированной* параметрической оценки, скажем, $\hat{f}(x, \theta)$, мы бы получили

$$\hat{f}(x, \theta) - f(x) = O_p\left(n^{-1/2} \right),$$

то есть более высокую скорость сходимости, чем в непараметрическом случае, из-за которой подобные модели называют $\sqrt{n}$-состоятельными. Конечно, если параметрическая модель неверно специфицирована, она не является состоятельной, поэтому Robinson (1988) на с. 933 называет такие модели «$\sqrt{n}$-несостоятельными».

Получив оптимальную ширину окна, займемся получением оптимальной ядерной функции. Главная задача ядра – обеспечить гладкость и дифференцируемость результирующей оценки. В ином контексте Hodges & Lehmann (1956) впервые показали, что оптимальная с точки зрения IMSE взвешивающая функция имеет вид

$$K_e(z) = \begin{cases} \frac{3}{4\sqrt{5}}\left(1 - \frac{1}{5}z^2\right), & -\sqrt{5} \leq z \leq \sqrt{5}, \\ 0, & \text{иначе.} \end{cases}$$

Этот результат получен с помощью техники вариационного исчисления, его вывод можно найти на с. 27–28 в Pagan & Ullah (1999). В рамках оценивания функции плотности такое ядро впервые предложил Епанечников (Epanechnikov, 1969), поэтому его часто называют ядром Епанечникова. Оказывается, целый ряд ядерных функций приводит к оценками с той же относительной эффективностью[4], так что можно выбирать ядро на основе вычислительной сложности; популярным выбором является гауссовское ядро.

В отличие от выбора ядерной функции, выбор подходящей ширины окна является ключевым аспектом корректного непараметрического анализа.

2.3 Выбор ширины окна

Ключом к проведению качественного непараметрического оценивания является выбор подходящей ширины окна для имеющейся задачи. Хотя ядерная функция остается важной, ее главная роль состоит в обеспечении дифференцируемости и гладкости получающейся оценки. Ширина окна, с другой стороны, 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. Ради аккуратности подчеркнем, что диктуемые данными методы выбора ширины окна не всегда гарантируют хороший результат. Для простоты изложения рассмотрим далее оценку одномерной плотности для непрерывных данных. Модификация для многомерного случая и данных смешанного типа требует незначительных изменений, заинтересованный читатель может обратиться к работе Li & Racine (2003) за подробностями об оценке плотности для данных смешанного типа.

2.3.1 Референтные эвристические правила

Рассмотрим оценку одномерной функции плотности, определенную в (2), с оптимальной шириной окна из (6). Быстрый взгляд на (6) дает понять, что оптимальная ширина окна зависит от соответствующей функции плотности, которая неизвестна. Референтные эвристические правила выбора ширины окна используют стандартное семейство распределений для присвоения значения неизвестной константе $\int f''(z)^2\,dz$. Например, в случае семейства нормальных распределений можно показать, что $\int f''(z)^2\,dz = \frac{3}{8\sqrt{\pi}\sigma^5}$. Если к тому же использовать гауссовское ядро, то

$$\int K^2(z)dz = \frac{1}{\sqrt{4\pi}}, \qquad \int z^2 K(z)dz = 1,$$

так что оптимальная ширина окна равна

$$h_{opt} = (4\pi)^{-1/10}\left(\frac{3}{8}\right)^{-1/5}\pi^{1/10}\sigma n^{-1/5} = 1{,}059\sigma n^{-1/5},$$

[4] См. таблицу 3.1 на с. 43 в Silverman (1986).

откуда происходит правило «$1{,}06\sigma n^{-1/5}$». На практике применяется $\hat{\sigma}$, выборочное стандартное отклонение.

2.3.2 Подстановка

Методы подстановки, такие как в Sheather & Jones (1991), состоят в подстановке оценок неизвестной константы $\int f''(z)^2\,dz$ в формулу для оптимальной ширины окна на основе первоначальной оценки $f''(z)$, которая в свою очередь основана на «предварительной» ширине окна, например, найденной по правилу $1{,}06\sigma n^{-1/5}$. Все прочие константы в выражении для h_{opt} известны после выбора ядерной функции (то есть $\int K^2(z)dz$ и $\int z^2K(z)dz$ известны). Хотя такие правила популярны, заинтересованный читатель может обратиться к работе Loader (1999), где обсуждаются относительные достоинства методов подстановки по сравнению с другими методами выбора ширины окна, обсуждаемыми ниже.[5]

2.3.3 Кросс-валидация на основе наименьших квадратов

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 $\hat{f}(x)$ и $f(x)$ имеет вид

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

Можно заменить эти величины их выборочными аналогами, сделать поправку на смещение и получить целевую функцию, которую затем можно минимизировать с помощью численных методов. Этот подход был предложен в работах Rudemo (1982) и Bowman (1984).

Для понимания сущности комментариев в Loader (1999) на Рис. 3 изображены оценки бимодальной плотности – ядерная оценка при применении правила подстановки и кросс-валидации на основе наименьших квадратов. Рис. 3 показывает, что на самом деле правило подстановки чрезмерно сглаживает, приводя к существенному смещению в левой вершине. Кросс-валидация на основе наименьших квадратов исправляет это, как отмечает Loader (1999), но ценой дополнительной вариации в правой вершине.

Одна из проблем данного подхода – его чувствительность к наличию округленных или дискретизированных данных, а также к мелкомасштабным эффектам в данных.

Из примера следует, что, возможно, ядерную оценку с фиксированным параметром h можно улучшить, и существуют «адаптивные» ядерные оценки, которые позволяют h меняться в точке x или X_i; см. Abramson (1982) и Breiman, Meisel & Purcell (1977). Эти оценки, однако, способствуют введению ложного шума в оценку плотности. Поскольку метод с фиксированным h доминирует в прикладных исследованиях, продолжим изучать этот подход.

2.3.4 Кросс-валидация на основе правдоподобия

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

[5]Loader пишет: «Мы обнаруживаем, что свидетельства превосходства методов подстановки гораздо менее убедительны, чем заявлялось ранее. В свою очередь, мы рассматриваем примеры с реальными данными, симуляции и асимптотику. В числе результатов мы находим, что методы подстановки настраиваются произвольным выбором предварительных оценок и склонны к чрезмерному сглаживанию при использовании в сложных задачах сглаживания».

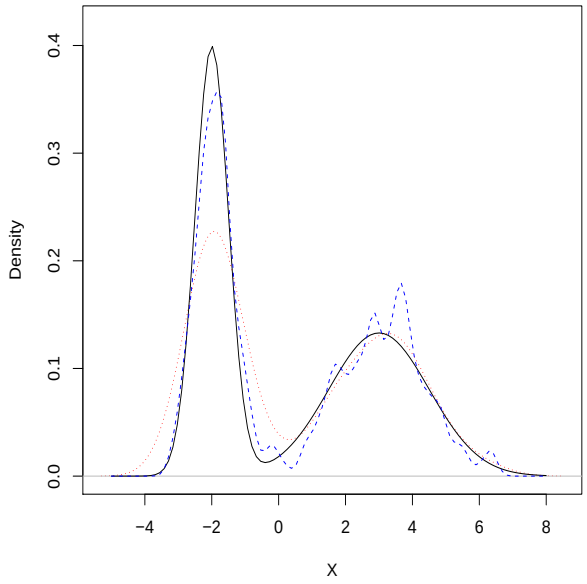

Рис. 3: Оценки плотности: правило подстановки и кросс-валидация на основе наименьших квадратов. Сплошная линия – истинная функция плотности, точечная линия – оценка при методе подстановки, пунктирная – при кросс-валидации на основе наименьших квадратов.

одного наблюдения), имеющей вид

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

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

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

Этот метод широко применим и был предложен в Stone (1974) и Geisser (1975). Один из недостатков этого метода в том, что он может давать чрезмерное сглаживание в случае распределений с тяжелыми хвостами, таких как распределение Коши.

2.3.5 Бутстраповские методы

Faraway & Jhun (1990) предложили метод выбора ширины окна h на основе бутстрапа путем оценивания IMSE из (5) для каждой фиксированной ширины окна и затем минимизации по всем значениям. Данный подход использует сглаженный бутстраповский метод на основе начальной оценки плотности. Один из недостатков этого подхода в том, что целевая функция является случайной, что может привести к проблемам при численной минимизации, а также его вычислительная сложность.

2.4 Частотная и ядерная оценки функции вероятности

До сих пор мы рассматривали оценивание одномерной функции плотности, предполагая, что анализируемые данные непрерывны по своей природе. Предположим теперь, что требуется оценить одномерную *функцию вероятности*, а данные дискретны по своей природе.

Непараметрический негладкий подход в данном случае предполагает частотную оценку, тогда как непараметрический гладкий подход дает ядерную оценку, весьма отличающуюся от приведенной в (2). Для незнакомых с термином «частотная» оценка поясним, что это просто оценка вероятности, подсчитанная по выборочной частоте наступления события. Например, если случайная величина – это результат испытания Бернулли (то есть 0 или 1 с постоянной вероятностью в каждом испытании), то частотная оценка вероятности нуля (единицы) – это просто число нулей (единиц), деленное на число испытаний.

Сначала рассмотрим оценивание функции вероятности, определенной для $X_i \in \mathcal{S} = \{0, 1, \ldots, c-1\}$. Негладкая «частотная» (неядерная) оценка $p(x)$ имеет вид

$$\tilde{p}(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}[X_i, x],$$

где $\mathbb{I}[\cdot]$ – индикаторная функция, определенная выше. Легко показать, что

$$\mathbb{E}[\tilde{p}(x)] = p(x),$$
$$\mathbb{V}[\tilde{p}(x)] = \frac{p(x)(1-p(x))}{n},$$

а значит,

$$\mathbb{MSE}\left[\tilde{p}(x)\right] = n^{-1} p(x)(1-p(x)) = O\left(n^{-1}\right),$$

откуда следует, что

$$\tilde{p}(x) - p(x) = O_p\left(n^{-1/2}\right).$$

Рассмотрим теперь ядерную оценку $p(x)$,

$$\hat{p}(x) = \frac{1}{n} \sum_{i=1}^{n} l(X_i, x), \tag{7}$$

где $l(\cdot)$ – ядерная функция, определенная, скажем, как

$$l(X_i, x) = \begin{cases} 1-\lambda & \text{если } X_i = x \\ \lambda/(c-1) & \text{иначе,} \end{cases}$$

где $\lambda \in [0, (c-1)/c]$ – «сглаживающий параметр», или «ширина окна». Требование, чтобы параметр λ лежал в отрезке $[0, (c-1)/c]$ обеспечивает то, что $\tilde{p}(x)$ является корректной оценкой вероятности, принадлежащей отрезку $[0, 1]$. Легко показать, что

$$\mathbb{E}[\hat{p}(x)] = p(x) + \lambda\left(\frac{1-cp(x)}{c-1}\right),$$
$$\mathbb{V}[\hat{p}(x)] = \frac{p(x)(1-p(x))}{n}\left(1 - \lambda\frac{c}{(c-1)}\right)^2.$$

Эта оценка была предложена в Aitchison & Aitken (1976) для дискриминантного анализа в случае многомерных бинарных данных. См. также Simonoff (1996).

Заметим, что, когда $\lambda = 0$, эта оценка превращается в частотную оценку $\tilde{p}(x)$, а если λ принимает наибольшее допустимое значение, $(c-1)/c$, ядерная оценка становится прямоугольной (то есть дискретной равномерной) оценкой, дающей одинаковые вероятности по всем исходам.

Используя ширину окна, балансирующую смещение и дисперсию, можно показать, что

$$\hat{p}(x) - p(x) = O_p\left(n^{-1/2}\right).$$

Заметим, что, в отличие от получения оценки Розенблатта–Парзена, в данном случае вместо приближений нам удалось использовать точные выражения для вывода результатов.

2.5 Ядерное оценивание плотности в случае смешанных дискретных и непрерывных данных

Предположим, в нашем распоряжении смесь дискретных и непрерывных данных, и требуется смоделировать их совместную функцию плотности.[6] В случае смеси дискретных и непрерывных данных исследователи традиционно используют ядерные методы, прибегая к «частотному» подходу. Этот подход подразумевает разбиение непрерывных данных на подмножества в соответствии с реализациями дискретных данных («клетки»). Конечно, это приводит к состоятельным оценкам. Тем не менее, по мере увеличения числа подмножеств количество данных в каждой клетке уменьшается, что ведет к проблеме «редких данных». В таких случаях данных в каждой клетке может оказаться недостаточно для того, чтобы дать разумные оценки плотности (они будут иметь большую дисперсию).

Подход, который рассматривается далее, использует концепцию «обобщенных мультипликативных ядер». Для непрерывных переменных используются стандартные непрерывные ядра, обозначаемые $W(\cdot)$ (ядро Епанечникова и др.). В случае неупорядоченной дискретной переменной $\bar{x}^d$ можно использовать ядро из Aitchison & Aitken (1976), имеющее вид

$$\bar{l}(\bar{X}_i^d, \bar{x}^d) = \begin{cases} 1-\lambda, & \text{если } \bar{X}_i^d = \bar{x}^d, \\ \frac{\lambda}{c-1}, & \text{иначе.} \end{cases}$$

В случае упорядоченной дискретной переменной $\tilde{x}^d$ можно использовать ядро из Wang & Vanryzin (1981), имеющее вид

$$\tilde{l}(\tilde{X}_i^d, \tilde{x}^d) = \begin{cases} 1-\lambda, & \text{если } \tilde{X}_i^d = \tilde{x}^d, \\ \frac{(1-\lambda)}{2}\lambda^{|\tilde{X}_i^d - \tilde{x}^d|}, & \text{если } \tilde{X}_i^d \neq \tilde{x}^d. \end{cases}$$

Обобщенное мультипликативное ядро для одной непрерывной, одной неупорядоченной и одной упорядоченной дискретных переменных определяется следующим образом:

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

Используя подобные мультипликативные ядра, можно модифицировать любой существующий ядерный метод для случая категориальных переменных, расширяя, таким образом, сферу применения ядерных методов.

Оценивание совместной функции плотности/вероятности, определенной на смешанных данных, осуществляется естественным образом путем применения этих обобщенных мультипликативных ядер. Например, для одной неупорядоченной дискретной переменной $\bar{x}^d$ и одной непрерывной переменной x^c, ядерная оценка функции плотности имеет вид

$$\hat{f}(\bar{x}^d, x^c) = \frac{1}{nh_{x^c}} \sum_{i=1}^{n} \bar{l}(\bar{X}_i^d, \bar{x}^d) W\left(\frac{X_i^c - x^c}{h_{x^c}}\right).$$

Этот метод естественным образом расширяется на случай смеси упорядоченных, неупорядоченных и непрерывных данных (то есть количественных и качественных переменных). Данная оценка особенно хорошо подходит для случая «редких данных». Чтобы не загромождать страницу обозначениями для формального определения оценки в случае p непрерывных, q упорядоченных и r неупорядоченных переменных, будем считать, что идея, лежащая в основе использования мультипликативных ядер, ясна, а подробности заинтересованный читатель может найти в Li & Racine (2003).

[6] Термин «плотность» подходит для функций распределения, определенных для смеси дискретных и непрерывных переменных. Это мера, определенная на дискретных переменных в соответствующей функции плотности.

Пример для смешанных дискретных и непрерывных данных

Рассмотрим набор данных “wage 1” из учебника Wooldridge (2002), содержащий $n = 526$ наблюдений, и смоделируем совместную функцию плотности двух переменных – непрерывной (“lwage”) и дискретной (“numdep”). Переменная “lwage” – это логарифм среднего почасового заработка индивида. Переменная “numdep” – количество иждивенцев, $(0, 1, \dots)$. Применим кросс-валидацию на основе правдоподобия (см. п. 2.3.4) для выбора ширины окна и получим оценку, представленную на Рис. 4.

Таблица 1: Данные по количеству иждивенцев в наборе данных из статьи Mroz (1987) (“numdep”) ($c = 0, 1, \dots, 6$)

c	n_c
0	252
1	105
2	99
3	45
4	16
5	7
6	2

Заметим, что мы имеем дело со случаем «редких» данных для некоторых клеток (см. таблицу 1), и традиционный подход потребовал бы непараметрической оценки одномерной функции плотности на основе только двух наблюдений в последней клетке ($c = 6$).

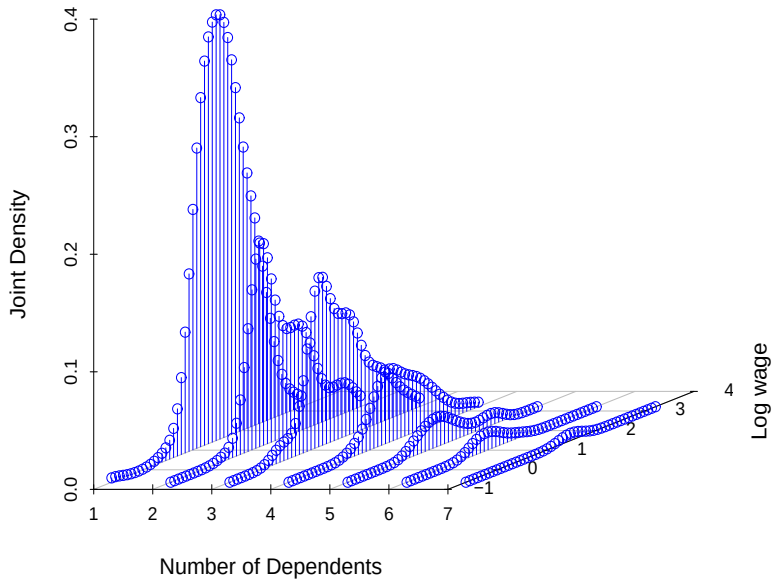

Рис. 4: Непараметрическая ядерная оценка совместной функции плотности, определенной для одной непрерывной и одной дискретной переменной.

2.6 Построение доверительных интервалов

Можно строить точечные и совместные доверительные интервалы для оценки плотности, и это обычно делают, используя либо асимптотическую формулу, приведенную в (4), где

неизвестные компоненты заменяются на их оценки, либо используя методы ресэмплинга, такие как бутстрап. Заметим, что асимптотическую нормальность ядерной оценки можно доказать, применив центральную предельную теорему Ляпунова для двойных массивов.

Точечные доверительные интервалы – это интервалы для фиксированной точки x, имеющие вид

$$\mathbb{P}\{\hat{f}_l(x) < f(x) < \hat{f}_u(x)\} = 1 - \alpha,$$

где α – вероятность ошибки первого рода. Совместные доверительные интервалы, с другой стороны, – это интервалы вида

$$\mathbb{P}\{\cap_{i=1}^{n}\{\hat{f}_l(X_i) < f(X_i) < \hat{f}_u(X_i)\}\} = 1 - \alpha.$$

Поскольку построение этих двух типов интервалов требует центрирования в $f(x)$, необходима корректировка смещения – либо путем оценки асимптотической формулы из (3), либо с помощью методов ресэмплинга, таких как бутстрап или метод «складного ножа».

С другой стороны, если исследователя интересует лишь оценка дисперсии оценки, доверительный интервал можно центрировать вокруг $\hat{f}(x)$ вместо несмещенной оценки $f(x)$. На Рис. 5 изображена оценка плотности из Рис. 2 и точечные 95%-ые доверительные интервалы (не скорректированные на смещение). Почему же скорректированные на смещение интервалы не являются нормой? Одна из причин в том, что оценка смещения, как известно, очень сложна, и получающиеся скорректированные на смещение оценки могут иметь большую дисперсию; см. подробности о скорректированных на смещение оценках в Efron (1982).

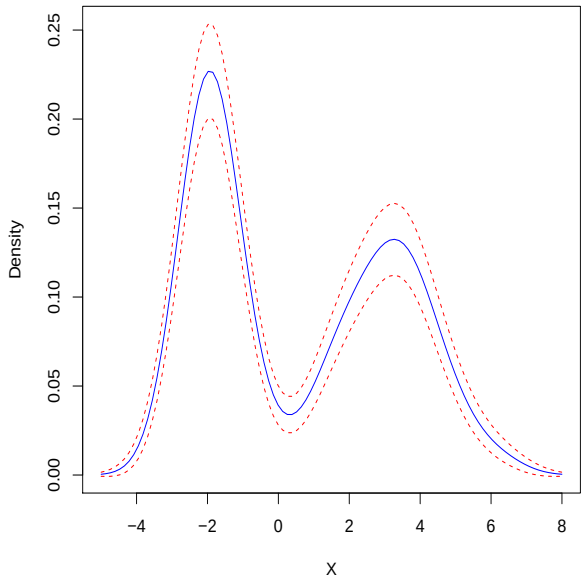

Рис. 5: Ядерная оценка плотности $\hat{f}(x) \pm 1{,}96 \times s$ с использованием асимптотических стандартных ошибок, s рассчитывается по формуле (4).

2.7 Проклятие размерности

По мере увеличения размерности пространства *непрерывной* переменной скорость сходимости ядерных методов падает, что хорошо известно как проблема «проклятия размерности». Если за p обозначить число непрерывных переменных, на которых определена функция плотности, можно показать, что

$$\hat{f}(x) - f(x) = O_p\left(n^{-2/(p+4)}\right);$$

см. вывод данного результаты в случае смешанных данных и кросс-валидации на основе наименьших квадратов в Li & Racine (2003a).

На с. 94 в работе Silverman (1986) представлена часто цитируемая таблица, содержащая размеры выборки, требуемые для обеспечения размера относительной среднеквадратической ошибки *правильно специфицированной параметрической оценки (многомерной нормальной)* по сравнению с многомерной ядерной оценкой плотности (только для непрерывных данных) менее 0,1 при оценивании в многомерном среднем, при определении относительной среднеквадратической ошибки как $\mathbb{E}[\hat{f}(\mu) - f(\mu)]^2/f(\mu)^2$, использовании гауссовского ядра и подсчете оптимальной поточечной ширины окна. Эту таблицу часто приводят те, кто из нее делают вывод, что ядерные методы бесполезны, когда размерность превышает две или три переменные.

Хотя, конечно, таблица в Silverman (1986) является верной, вывод о бесполезности ядерных методов в случае, когда размерность превышает несколько переменных, отсюда не следует по двум простым причинам. Во-первых, популярные параметрические модели редко, если вообще когда-либо, правильно специфицированы.[7] Конкуренция, таким образом, идет между *неправильно специфицированными*, а значит, *несостоятельными* параметрическими моделями и относительно *неэффективными*, но *состоятельными* непараметрическими моделями.[8] Во-вторых, проклятие размерности применимо лишь к числу имеющихся *непрерывных* переменных. В прикладных исследованиях не являются редкими случаи, когда анализируется небольшое число непрерывных переменных, или же часто данные состоят исключительно из категориальных переменных.

3 Оценивание условной плотности

Функции условной плотности лежат в основе многих важных статистических объектов, хотя их редко моделируют напрямую в рамках параметрических моделей и, возможно, еще реже в рамках ядерного оценивания. Тем не менее, как будет видно, они чрезвычайно полезны для целого ряда задач, будь то прямое оценивание функций условной плотности, моделирование счетных данных (см. детальное рассмотрение моделей счетных данных в Cameron & Trivedi, 1998), или же моделирование условных квантилей путем оценивания условной КФР. И, конечно, регрессионный анализ (то есть моделирование условного среднего) напрямую зависит от функции условной плотности, так что этот статистический объект на самом деле неявно составляет основу многих популярных статистических методов.

3.1 Ядерное оценивание функции условной плотности

Пусть $f(\cdot)$ и $\mu(\cdot)$ обозначают совместную и маржинальную плотности (X, Y) и X соответственно, где Y и X могут быть непрерывными, неупорядоченными и упорядоченными переменными. Далее будем считать Y зависимой переменной (то есть Y – объясняемая переменная), а X – независимой (то есть X – объясняющая переменная). Введем обозначения $\hat{f}$ и $\hat{\mu}$ для соответствующих ядерных оценок, и оценим условную плотность $g(y|x) = f(x, y)/f(x)$ как

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

Ядерные оценки совместной и маржинальной плотностей $f(x, y)$ и $f(x)$ описаны в предыдущей главе и не воспроизводятся здесь; см. подробности теоретических оснований диктуемого данными выбора ширина окна для этого метода в Hall, Racine & Li (2004).

[7] «Нормальность – это миф; никогда не было и никогда не будет нормального распределения» (Geary, 1947).

[8] Robinson (1988) называет параметрические модели $\sqrt{n}$-несостоятельными (обычно их называют $\sqrt{n}$-состоятельными), чтобы подчеркнуть это явление.

3.1.1 Наличие несущественных регрессоров

Hall, Racine & Li (2004) предложили оценку, приведенную в (9), но выбор подходящих параметров сглаживания в рамках этого подхода может быть сложен, не в последнюю очередь потому, что правила подстановки принимают особенно сложный вид в случае смешанных данных. Одна из сложностей состоит в том, что не существует общей формулы для оптимальных параметров сглаживания. Гораздо более серьезная проблема состоит в том, что может быть сложно определить, какие компоненты X существенны для задачи условной инференции. Например, если j-ая компонента X независима от Y, то она несущественна для оценки плотности Y при данном X, и в идеале должна быть удалена перед осуществлением инференции. Hall, Racine & Li (2004) показывают, что кросс-валидация на основе наименьших квадратов преодолевает эти трудности. Она автоматически определяет, какие компоненты существенны, а какие – нет, путем присвоения больших параметров сглаживания последним и, следовательно, сжимая их в сторону равномерного распределения. Эта процедура эффективно устраняет несущественные компоненты из анализа, удаляя их вклад в дисперсию оценки; смещение их и так очень мало вследствие их независимости от Y. Кросс-валидация также дает важную информацию о том, какие компоненты существенны: это именно те компоненты, которые процедура кросс-валидации выбрала сглаживать обычным способом, присвоив им параметры сглаживания обычного размера. Кросс-валидация дает асимптотически оптимальное сглаживание для существенных параметров и устраняет несущественные компоненты путем чрезмерного сглаживания.

Важность этого результата лучше всего видна при сравнении условий состоятельности, приведенных в п. 2.2.1, где упомянуты стандартные результаты для оценивания плотности: $h \to 0$ при $n \to \infty$ (смещение $\to 0$) и $nh \to \infty$ при $n \to \infty$ (дисперсия $\to 0$). Hall, Racine & Li (2004) показывают, что для несущественных переменных в X соответствующая ширина окна должна вести себя в точности противоположным образом, а именно: $h \to \infty$ при $n \to \infty$ при оптимальном сглаживании. То же самое было показано в рамках регрессионного анализа; см. дальнейшие подробности в Hall, Li & Racine (в печати).

3.1.2 Моделирование панельных данных по ВВП Италии

Рассмотрим панельные данные по динамике ВВП Италии (использованные Giovanni Baiocchi) для 21 региона за период 1951–1998 гг. (в млн. лир, базисный год – 1990). Панель содержит 1008 наблюдений и две переменные – "gdp" (ВВП) и "year" (соответствующий год). Исходя из природы данных, будем считать ВВП непрерывной переменной, а год $(1951, 1952, \dots)$ – упорядоченной дискретной переменной. Далее оценим плотность распределения ВВП условно на год. На Рис. 6 изображена оценка условной плотности, $\hat{f}(\text{gdp}|\text{year})$ при выборе ширины окна с помощью кросс-валидации на основе правдоподобия, которая дала ширину окна $\hat{h}_{\text{gdp}} = 0{,}715$ для ВВП и $\hat{\lambda}_{\text{year}} = 0{,}671$ для года.

Из Рис. 6 видно, что распределение дохода трансформировалось из унимодального в начале 1950-х годов в четко выраженное бимодальное в 1990-х. Данный результат устойчив к выбору ширины окна и наблюдается независимо от того, применяются ли простые эвристические правила или же диктуемые данными методы, такие как кросс-валидация на основе наименьших квадратов или правдоподобия. Ядерный метод сразу раскрывает эту тенденцию, которую легко упустить при применении параметрических моделей распределения дохода. Например, (унимодальное) логнормальное распределение является популярной параметрической моделью для распределений дохода, но она неспособна раскрыть мультимодальную структуру, имеющуюся в данных.

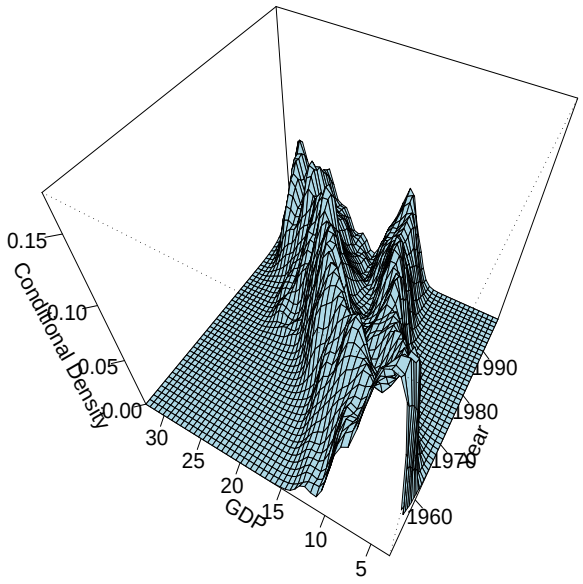

Рис. 6: Непараметрическая оценка функции условной плотности для панельных данных по ВВП Италии.

3.2 Ядерное оценивание функции условного распределения

Li & Racine (в печати) предложили непараметрическую ядерную оценку функции условного распределения, подходящую для смешанных дискретных и категориальных данных, а также соответствующую непараметрическую оценку условного квантиля. Выбор ширины окна для ядерной квантильной регрессии остается открытым вопросом, и авторы используют модификацию метода выбора ширины окна на основе функции условной плотности, предложенного в Hall, Racine & Li (2004).

Пусть $F(y|x)$ – функция условного распределения Y при $X = x$, а $f(x)$ – функция маржинальной плотности X. Можно оценить $F(y|x)$ как

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

где $G(\cdot)$ – ядерная функция распределения, выбранная исследователем, скажем, функция стандартного нормального распределения, h_0 – параметр сглаживания, соответствующий Y, а $K_h(X_i, x)$ – мультипликативное ядро, определенное в (8), где ради простоты обозначений каждое одномерное непрерывное ядро поделено на соответствующую ширину окна.

На Рис. 7 представлена эта оценка для данных по ВВП Италии из п. 3.1.2.

Функция условного распределения, представленная на Рис. 7, предоставляет информацию, содержащуюся на Рис. 6, только способом, более удобным для оценивания, скажем, функции условной квантили, к рассмотрению которой мы переходим.

3.3 Ядерное оценивание условной квантили

Оценивание функций регрессии – популярное занятие среди практиков. Иногда, однако, регрессионная функция не отражает влияние объясняющих переменных на зависимую. Например, если зависимая переменная является цензурированной слева (или справа), зависимость, отражаемая регрессионной функцией, искажается. В таких случаях условные квантили выше (или ниже) точки цензурирования являются устойчивыми к наличию цензурирования.

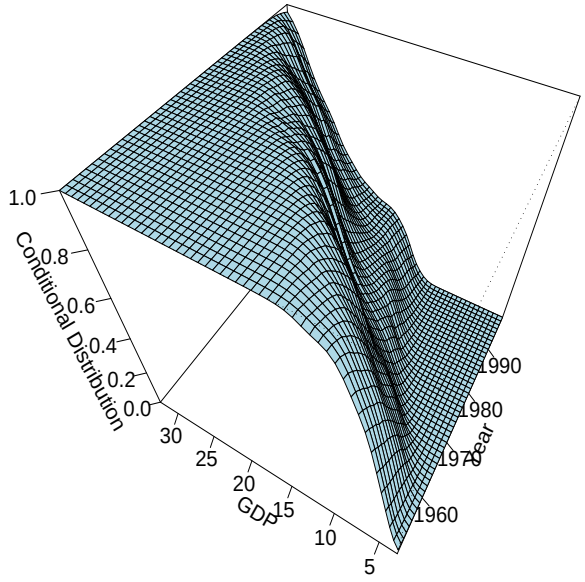

Рис. 7: Непараметрическая оценка функции условного распределения для панельных данных по ВВП Италии.

Более того, функция условной квантили дает более полную картину условного распределения зависимой переменной, чем функция условного среднего.

Если мы умеем оценивать функцию условного распределения, такую как представлена на Рис. 7, оценивание условных квантилей следует естественным образом. То есть, оценив функцию условного распределения, нужно просто обратить ее в нужной квантили, как описывается ниже. Условная α-квантиль ($\alpha \in (0,1)$) функции условного распределения $F(\cdot|x)$ определяется как

$$q_\alpha(x) = \inf\{y:\ F(y|x) \geq \alpha\} = F^{-1}(\alpha|x).$$

Или, что эквивалентно, $F(q_\alpha(x)|x) = \alpha$. Можно прямо оценить функцию условной квантили $q_\alpha(x)$, обращая оцененную функцию условного распределения, то есть

$$\hat{q}_\alpha(x) = \inf\{y:\ \hat{F}(y|x) \geq \alpha\} \equiv \hat{F}^{-1}(\alpha|x).$$

Теоретические подробности этого метода можно найти в Li & Racine (в печати).

На Рис. 8 представлены 0,25, 0,50 (медиана) и 0,75 условные квантили для данных по ВВП Италии из п. 3.1.2, совместно с блочными диаграммами[9] исходных данных. Приятные особенности данного примера в том, что объясняющая переменная является упорядоченной, и имеется несколько наблюдений за год. Негладкие оценки квантилей, полученные с помощью блочных диаграмм, можно напрямую сравнить с полученными прямым оцениванием гладкой функции распределения, и очевидно, что эти оценки согласуются.

[9]Диаграмма «ящик с усами» (box-and-whisker plot), иногда называемая просто блочной или ящичковой диаграммой (box plot) – это метод визуализации данных наподобие гистограммы, изобретенный Дж. Тьюки (J. Tukey). Для создания диаграммы «ящик с усами» следует нарисовать ящик с концами в первой и третьей квартилях распределения, Q_1 и Q_3. Далее рисуется статистическая медиана M в виде горизонтальной линии внутри ящика. Затем рисуются «усы» к наиболее дальним точкам, не являющимся выбросами (то есть расположенным внутри интервала, равного $3/2$, помноженным на расстояние между квартилями Q_1 и Q_3). Затем для каждой точки, находящейся на расстоянии более $3/2$, помноженных на межквартильный интервал, от конца ящика, рисуется точка.

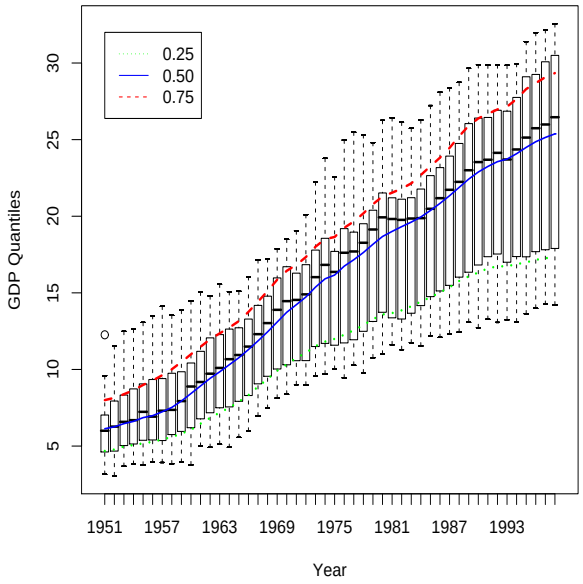

Рис. 8: Непараметрические оценки условных квантилей для панельных данных по ВВП Италии, $\alpha = (0{,}25, 0{,}50, 0{,}75)$.

3.4 Модели бинарного выбора и счетных данных

Другим приложением ядерного оценивания плотностей для смешанных данных является оценивание моделей условной моды. Для примера рассмотрим какую-нибудь дискретную величину, скажем, $Y \in \mathcal{S} = \{0, 1, \ldots, c-1\}$, которая обозначает число успешных заявок на получение патента. Определим условную моду распределения $y|x$ как

$$m(x) = \max_y g(y|x). \tag{11}$$

Для оценивания условной моды $m(x)$ необходимо смоделировать условную плотность. Обозначим за $\hat{m}(x)$ оценку условной моды, имеющую вид

$$\hat{m}(x) = \max_y \hat{g}(y|x), \tag{12}$$

где $\hat{g}(y|x)$ – ядерная оценка $g(y|x)$, определенная в (9). Рассмотрим в качестве примера моделирование низкого веса при рождении (бинарного показателя), используя этот метод.

Моделирование низкого веса при рождении (0/1)

Для этого примера возьмем данные о весе при рождении из библиотеки `MASS` (Venables & Ripley, 2002) программной среды `R` и оценим параметрическую логит-модель и непараметрическую модель условной моды, используя (12), где условная плотность оценивается по формуле (9) на основе метода Hall, Racine & Li (2004). Затем сравниваем их таблицы расхождения[10] и оцениваем их классификационную способность. Исход y – бинарный показатель низкого веса новорожденного ребенка (“low”), определяемый ниже. Этот метод можно без изменений применять к неупорядоченным и упорядоченным мультиномиальным исходам. Данный пример содержит $n = 189$ наблюдений и 7 объясняющих переменных в векторе x – “smoke”, “race”, “ht”, “ui”, “ftv”, “age”, и “lwt” – определяемых ниже.

Переменные означают следующее:

[10] «Таблица расхождения» – это таблица реальных исходов и предсказанных моделью. Диагональные элементы содержат верно предсказанные исходы, а внедиагональные – неверно предсказанные.

1. “low” – индикатор веса при рождении менее 2,5 кг
2. “smoke” – индикатор курения во время беременности
3. “race” – расовая принадлежность матери (1 = белая, 2 = черная, 3 = другая раса)
4. “ht” – история гипертонии
5. “ui” – наличие раздражительности в утробе
6. “ftv” – количество визитов к врачу в течение первого триместра
7. “age” – возраст матери, в годах
8. “lwt” – вес матери в фунтах во время последнего менструального периода

Заметим, что все переменные, кроме “age” и “lwt”, по своей природе являются категориальными.

Рассчитаем таблицы расхождения для каждой модели, используя кросс-валидацию на основе правдоподобия для получения ширины окна в непараметрической модели условной моды.

Таблица 2: Таблицы расхождения для данных о низком весе при рождении. В таблице слева представлен результат для параметрической логит-модели, справа – для ядерной модели.

		Предсказано	
Реально		0	1
	0	119	11
	1	34	25

		Предсказано	
Реально		0	1
	0	127	1
	1	27	32

Как видно, непараметрическая модель верно классифицирует $(127 + 32)/189 = 84{,}1\%$ наблюдений с низким/высоким весом при рождении, тогда как логит-модель правильно классифицирует лишь $(119 + 25)/189 = 76{,}1\%$ наблюдений.

4 Регрессия

Один из наиболее популярных методов непараметрической ядерной оценки регрессии был предложен в Nadaraya (1965) и Watson (1964) и известен как оценка Надараи–Уотсона, а также как «локально постоянная» оценка по причинам, которые станут ясны при описании «локально полиномиальной» оценки (Fan, 1992). Начнем с краткого введения в метод локально постоянного оценивания регрессионных функций и их производных, а затем приступим к локально полиномиальной регрессии. Напомним, что дальнейший материал основан на многих объектах, рассмотренных в разделах *Оценивание плотности и функции вероятности* и *Оценивание условной плотности*, таких как обобщенное мультипликативное ядро и др.

4.1 Локально постоянная ядерная регрессия

Ради простоты обозначений начнем с рассмотрения парной регрессии.[11]

[11] Как будет видно, случай многомерных смешанных данных вытекает естественным образом; о необходимых модификациях будет сказано в соответствующих разделах.

4.1.1 Локально постоянная оценка условного среднего ($\hat{g}(x)$)

По определению, условное среднее непрерывной случайной величины Y равно

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

где $g(y|x)$ – функция условной плотности, определенная в главе *Оценивание условной плотности*, а $m(x) = \int y f(y,x)\, dy$.

Локально постоянная оценка условного среднего получается заменой неизвестных функций плотности совместного и маржинального распределений, $f(y,x)$ и $f(x)$, их ядерными оценками, определенными в главе *Оценивание плотности и функции вероятности*, что дает

$$\hat{g}(x) = \int y \frac{\hat{f}(y,x)}{\hat{f}(x)}\, dy.$$

После незначительных преобразований локально постоянная оценка $\hat{g}(x)$ упрощается до

$$\hat{g}(x) = \int y \frac{\hat{f}(y,x)}{\hat{f}(x)}\, dy = \frac{\sum_{i=1}^{n} Y_i K\left(\frac{X_i - x}{h_x}\right)}{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h_x}\right)}. \tag{13}$$

Заметим, что знак интеграла исчезает из-за использования мультипликативной ядерной функции и замены переменной.

Заметим, что при условиях, которые будут даны далее, $\hat{g}(x)$ является состоятельной оценкой условного среднего. В сущности, мы локально усредняем те значения зависимой переменной, которые «близки» в смысле значений, принимаемых регрессорами. Контролируя объем локальной информации, используемой для построения оценки («размер локальной выборки»), позволяя локальному усреднению становиться более информативным по мере увеличения размера выборки и в то же время уменьшая окрестность, в которой происходит усреднение, можно обеспечить состоятельность оценок при стандартных условиях регулярности.

4.1.2 Асимптотическое смещение и дисперсия

Хотя локально постоянная оценка широко используется, она подвержена «смещению на краях», что легко увидеть при рассмотрении формулы для асимптотического смещения, которая в двумерном случае имеет вид

$$h^2 \left(\frac{1}{2} g''(x) + \frac{g'(x) f'(x)}{f(x)}\right) \kappa_2$$

(см. вывод на с. 101 в Pagan & Ullah, 1999). При прочих равных, по мере приближения к границе носителя распределения данных, $f(x)$ стремится к нулю и смещение растет. Класс локально полиномиальных оценок, рассматриваемый в главе 4.2 ниже, не подвержен смещению на краях, хотя для него свойственны проблемы, связанные с численной нестабильность, которые вскоре будут описаны. Асимптотическое смещение локально линейной оценки имеет вид

$$\frac{h^2}{2} g''(x) \kappa_2,$$

и, как легко видеть, член, ведущий к смещению на краях в выражении для локально постоянной оценки, а именно $g'(x) f'(x)/f(x)$, отсутствует в случае локально линейной оценки.

В главе 4.2 описывается локально линейная оценка для двумерного случая, а здесь мы отметим лишь, что локально постоянная и локально линейная оценки имеют одинаковую асимптотическую дисперсию, которая в двумерном случае имеет вид

$$\frac{\sigma^2(x)}{f(x)nh}\int K^2(z)\,dz,$$

где $\sigma^2(x)$ – условная дисперсия y.

4.1.3 Оптимальная и диктуемая данными ширина окна

Оптимальная с точки зрения IMSE ширина окна для локально постоянной оценки,

$$h_{opt}=\left[\frac{\sigma^2(x)\int f^{-1}(x)\,dx\int K^2(z)\,dz}{\int[2g'(x)f'(x)f^{-1}(x)+g''(x)]^2\,dx\kappa_2^2}\right]^{1/5}n^{-1/5},$$

выводится точно таким же способом, как и в п. 2.2.1, и, как и в случае оценивания плотности, зависит от неизвестных величин, определяемых процессом, порождающим данные.

Хотя можно применить методы подстановки, в многомерном случае они недостижимы из-за необходимости оценивания, среди прочего, производных высших порядков наряду со смешанными частными производными, тогда как для случая смешанных данных вообще не существует общих формул. На практике применяются альтернативные диктуемые данными подходы.

Два популярных диктуемых данными подхода к выбору ширины окна, обладающему желаемыми свойствами, – это кросс-валидация на основе наименьших квадратов и AIC_c-метод из Hurvich, Simonoff & Tsai (1998), основанный на минимизации модифицированного информационного критерия Акаике.

Кросс-валидация на основе наименьших квадратов для регрессии основана на минимизации функции

$$CV(h)=n^{-1}\sum_{i=1}^{n}(Y_i-\hat{g}_{-i}(X_i))^2,$$

где $\hat{g}_{-i}(X_i)$ – оценка $g(X_i)$ при удалении из выборки i-го наблюдения перед формированием прогноза для наблюдения i.

Подход в Hurvich, Simonoff & Tsai (1998) основан на минимизации

$$\mathrm{AIC}_c=\ln(\hat{\sigma}^2)+\frac{1+\mathrm{tr}(H)/n}{1-(\mathrm{tr}(H)+2)/n},$$

где

$$\hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}[Y_i-\hat{g}(X_i)]^2=Y'(I-H)'(I-H)Y/n,$$

$\hat{g}(X_i)$ – непараметрическая оценка, а H – $n\times n$ взвешивающая функция (то есть матрица ядерных весов) с элементом (i,j), имеющим вид $H_{ij}=K_h(X_i,X_j)/\sum_{l=1}^{n}K_h(X_i,X_l)$, где $K_h(\cdot)$ – обобщенное мультипликативное ядро.

Доказано, что кросс-валидация и AIC_c-метод асимптотически эквивалентны; см. подробности в Li & Racine (2004).

4.1.4 Существенные и несущественные регрессоры

Для существенного x условия на состоятельность те же самые, что и при оценке плотности, а именно: $h \to 0$ при $n \to \infty$ и $nh \to \infty$ при $n \to \infty$. Однако, если на самом деле x не является существенным, можно показать, что условие $h \to \infty$ при $n \to \infty$, а не $h \to 0$, дает оптимальное сглаживание. Доказано, что применение кросс-валидации на основе наименьших квадратов для выбора ширины окна ведет к оптимальному сглаживанию как для существенных, так и для несущественных x; подробности см. в Hall, Li & Racine (2006).

Для локально постоянной оценки условного среднего величины y при $h \to \infty$ имеем

$$\hat{g}(x) = \frac{\sum_{i=1}^{n} Y_i K(0)}{\sum_{i=1}^{n} K(0)} = n^{-1} \sum_{i=1}^{n} Y_i = \bar{y},$$

что есть *безусловное среднее* y. В таком случае говорится, что x «полностью сглаживается» из регрессионной функции, что разумно в случае, когда x не содержит полезной информации для предсказания y.

Интуиция, стоящая за желательностью полного сглаживания несущественных регрессоров, весьма проста. Присутствие несущественного x означает, что смещение $\hat{g}(x)$ равно нулю при любом значении h. Значит, можно использовать относительно малые значения h, но оценки с относительно малыми h обязательно будут более вариабельны, чем оценки с относительно большим h. Поскольку кросс-валидация дает аппроксимацию для MSE оценки, ясно, что MSE в этом случае минимизируется, если минимизируется дисперсия $\hat{g}(x)$, что имеет место при таком h, когда $\hat{g}(x) = \bar{y}$, то есть при $h \to \infty$. Повторим, что кросс-валидация способна дать подходящее значение h как в случае существенных, так и в случае несущественных регрессоров. Наконец, заметим, что скорость сходимости локально постоянной ядерной оценки при использовании оптимального сглаживания (обратно) пропорциональна $\sqrt{n}$ при наличии несущественных регрессоров, что соответствует параметрической скорости, тогда как в случае существенных регрессоров скорость сходимости пропорциональна $\sqrt{n^{4/5}}$ при использовании ядер второго порядка, то есть более низкая, чем параметрическая. Этот факт, возможно, ценится меньше, чем того заслуживает, и имеет важные выводы для автоматического снижения размерности в многомерном случае, что способно смягчить проблему проклятия размерности в некоторых ситуациях.

Расширение на случай множественной регрессии вытекает естественным образом, а версия для случая смешанных данных получается простой заменой ядра на обобщенное мультипликативное ядро, определенное в главе *Оценивание плотности и функции вероятности*; см. теоретические основания этого метода в Racine & Li (2003).

4.1.5 Локально постоянная оценка отклика ($\hat{\beta}(x)$)

Вдобавок к оценке условного среднего часто требуется оценить маржинальные эффекты («производные», или «отклик»).

Неизвестный отклик $\beta(x)$ в двумерном случае, рассмотренном выше, определяется следующим образом:

$$\beta(x) \equiv \frac{d\,g(x)}{dx} = g'(x) = \frac{f(x)m'(x) - m(x)f'(x)}{f^2(x)} = \frac{m'(x)}{f(x)} - g(x)\frac{f'(x)}{f(x)}.$$

Локально постоянная оценка получается заменой неизвестных $f(x)$, $m'(x)$, $g(x)$ и $f'(x)$ их ядерными оценками и имеет вид

$$\hat{\beta}(x) \equiv \frac{d\,\hat{g}(x)}{dx} = \frac{\hat{f}(x)\hat{m}'(x) - \hat{m}(x)\hat{f}'(x)}{\hat{f}^2(x)} = \frac{\hat{m}'(x)}{\hat{f}(x)} - \hat{g}(x)\frac{\hat{f}'(x)}{\hat{f}(x)},$$

где

$$\hat{m}(x) = \frac{1}{nh}\sum_i Y_i K\left(\frac{X_i - x}{h}\right),$$
$$\hat{f}(x) = \frac{1}{nh}\sum_i K\left(\frac{X_i - x}{h}\right),$$
$$\hat{m}'(x) = -\frac{1}{nh^2}\sum_i Y_i K'\left(\frac{X_i - x}{h}\right),$$
$$\hat{f}'(x) = -\frac{1}{nh^2}\sum_i K'\left(\frac{X_i - x}{h}\right).$$

Вновь многомерная версия вытекает естественным образом, а в случае смешанных данных применяются обобщенные мультипликативные ядра, определенные выше, и, конечно, эта оценка определена только для непрерывных регрессоров.

4.2 Локально полиномиальная ядерная регрессия

Оценка, приведенная в (13), называется локально постоянной оценкой, поскольку она минимизирует следующее выражение:

$$\hat{g}(x) \equiv \min_a \sum_{i=1}^{n} (Y_i - a) K\left(\frac{X_i - x}{h}\right).$$

Рассмотрим ее популярное расширение, которое не имеет смещения на краях, хотя может иметь другие проблемы, такие как потенциальная вырожденность, которая часто возникает в случае редких данных. Наиболее распространенным локально полиномиальным методом является локально линейный подход, который описывается ниже и который снова рассматривается для двумерного случая ради простоты обозначений.

Предполагая, что существует вторая производная $g(x)$, в малой окрестности точки x, $g(x_0) \approx g(x) + (\partial g(x)/\partial x)(x_0 - x) = a + b(x_0 - x)$. Задача оценивания $g(x)$ эквивалента задаче оценивания константы a в рамках локально линейной регрессии. Задача оценивания отклика $\partial g(x)/\partial x$ эквивалента задаче оценивания угла наклона b в рамках локально линейной регрессии.

Выберем a и b, минимизирующие

$$\mathcal{S} = \sum_{i=1}^{n} (Y_i - a - b(X_i - x))^2 K\left(\frac{X_i - x}{h}\right) = \sum_{i=1}^{n} (Y_i - a - b(X_i - x))^2 K(Z_i).$$

Решения $\hat{a}$ и $\hat{b}$ являются локально линейными оценками $g(x)$ и $\beta(x)$, соответственно. Решая, получаем

$$\begin{pmatrix} \hat{g}(x) \\ \hat{\beta}(x) \end{pmatrix} = \left[\sum_{i=1}^{n} \begin{pmatrix} 1 & X_i - x \\ X_i - x & (X_i - x)^2 \end{pmatrix} K(Z_i)\right]^{-1} \sum_{i=1}^{n} \begin{pmatrix} 1 \\ X_i - x \end{pmatrix} K(Z_i) y_i.$$

Одна из особенностей данного подхода в том, что он сразу дает оценки среднего и отклика, что не так в случае локально постоянной оценки. Формулы для асимптотического смещения и дисперсии приведены в п. 4.1.2 выше. Для оценивания маржинальных эффектов (то есть $\beta(x)$) обычно используют полином более высокого порядка (то есть для оценивания первых производных применяется локальная квадратичная регрессия) как средство снижения смещения (см. Fan & Gijbels, 1996).

Одна из проблем, часто выплывающих на поверхность при использовании этой оценки, заключается в вырожденности, возникающей из-за наличия редких данных, особенно при малой ширине окон, и для ее решения были предложены различные формы риджинга. Методы риджинга являются техникой решения плохо обусловленных задач линейной регрессии. Впервые этот подход был предложен в Hoerl & Kennard (1970). См. подробности использования риджинга в контексте локально линейной оценки в Cheng, Hall & Titterington (1997) и Seifert & Gasser (2000).

Поведение локально линейной оценки по отношению к h заметно отличается от случая локально постоянной оценки. При $h \to \infty$ локально линейная оценка $\hat{g}(x)$ стремится к $\hat{\beta}_0 + \hat{\beta}_1 x$, где $\hat{\beta}_0$ и $\hat{\beta}_1$ – МНК-оценки линейной регрессии y на x. То есть при $h \to \infty$ локально линейная подгонка приближается к глобально линейной подгонке точно так же, как локально постоянная подгонка приближается к глобально постоянной, а именно $\bar{y}$. Тем не менее, хотя локально постоянная оценка характеризуется тем, что несущественные переменные могут полностью сглаживаться, это свойство не сохраняется для локально линейной оценки, что может привести к излишней вариабельности при наличии несущественных регрессоров.

Формулы для смещения и дисперсии этой оценки приведены в главе 4.1. Многомерная версия локально линейной оценки для случая смешанных данных естественным образом следует при применении обобщенных мультипликативных ядер; см. подробности в Li & Racine (2004).

Симулированный двумерный пример

Рассмотрим пример, в котором генерируется выборка размера $n = 50$, x равномерно распределен, и $y = \sin(2\pi x) + \epsilon$, где ϵ – нормально распределенная величина с $\sigma = 0{,}25$. Сначала рассмотрим случай, когда для выбора ширины окна применяется кросс-валидация на основе наименьших квадратов. На Рис. 9 представлены данные, истинный DGP, а также локально постоянная и локально линейная оценки $g(x) = \sin(2\pi x)$.

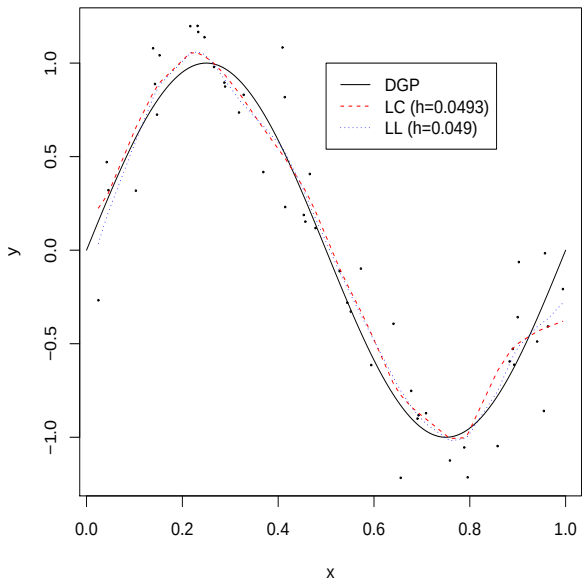

Рис. 9: Локально постоянная и локально линейная оценки при кросс-валидации на основе наименьших квадратов, $n = 50$.

Из Рис. 9 видно, что локально постоянная оценка характеризуется некоторым очевидным смещением на краях, так как оценка сдвигается немного вниз на правом краю и немного вверх на левом краю, как и следовало ожидать после изучения асимптотического смещения.

Тем не менее, обе оценки дают правдоподобное описание соответствующего DGP.

Далее рассмотрим различия в поведении локально постоянной и локально линейной оценок при $h \to \infty$. Положим соответствующую ширину окон равной $h = 100000$. На Рис. 10 представлены данные, истинный DGP, а также локально постоянная и локально линейная оценки.

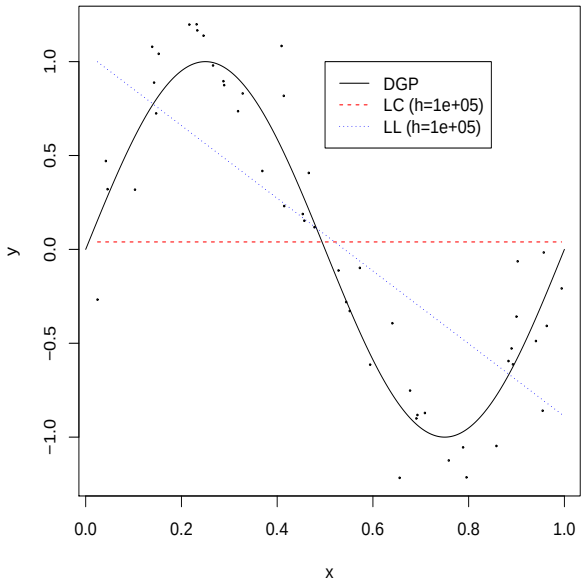

Рис. 10: Локально постоянная и локально линейная оценки при чрезмерном сглаживании, $h = 100000$, $n = 50$.

Рис. 10 ясно иллюстрирует четкие различия в свойствах каждой оценки при большом значении h и подчеркивает тот факт, что локально линейная оценка неспособна полностью удалить переменную с помощью чрезмерного сглаживания.

Предположим, интерес представляют маржинальные эффекты. В этом случае можно рассмотреть локально постоянную и локально линейную оценки $\beta(x)$. На Рис. 11 представлены соответствующие оценки отклика на основе полученной при кросс-валидации ширины окон.

Читатель может подумать, что эти оценки вовсе не такие гладкие, и, конечно, будет прав. Вспомним, что используется маленькая выборка ($n = 50$), стохастическая ширина окна, а при росте n оценки постепенно становятся гладкими. Тем не менее, здесь стоит отметить, что обычные параметрические спецификации, применяемые во многих прикладных эконометрических работах, вовсе неспособны отловить даже простое среднее и отклик, рассмотренные в этом примере. Напомним, что это и есть конкуренция, о которой говорилось выше, и, хотя оценки могут показаться некоторым читателям не столь приятными, на самом деле они высоко информативны.

Иллюстративное сравнение методов выбора ширины окна

Для оценки того, как различные методы выбора ширины окна работают на реальных данных, рассмотрим следующий пример, использующий данные из библиотеки `car` (Fox, 2002) программной среды `R`. Набор данных состоит из 102 наблюдений, каждое из которых соответствует конкретному роду занятий. Зависимая переменная – это престижность видов деятельности в Канаде, измеренная с помощью шкалы Пинео–Портера (Pineo–Porter) для профессий, взятых из социологического опроса середины 1960-х годов. Объясняющая переменная – средний доход для каждой профессии в канадских долларах 1971 г. На Рис. 12

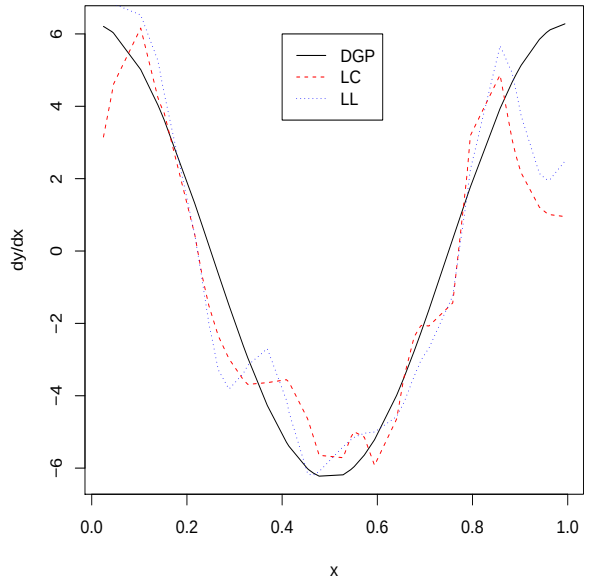

Рис. 11: Локально постоянная и локально линейная оценки отклика $\beta(x)$ при использовании кросс-валидации на основе наименьших квадратов, $n = 50$, $dy/dx = 2\pi\cos(2\pi x)$.

изображены данные и пять локально линейных оценок, соответствующих ширине окон при недостаточном сглаживании, чрезмерном сглаживании, полученной по методу прямой подстановки из Ruppert, Sheather & Wand (1995), AIC_c-методу из Hurvich, Simonoff & Tsai (1998) и при кросс-валидации. Во всех случаях использовалось гауссовское ядро второго порядка.

Как видно из рисунка, локально линейная оценка при чрезмерном сглаживании глобально линейна и в действительности соответствует простой линейной регрессии y на x, как и ожидалось, тогда как AIC_c-метод и кросс-валидация дают наиболее правдоподобную подгонку для имеющихся данных. Как уже отмечалось, в случае смешанных данных не существует правил подстановки. Мы получили разумные результаты при использовании кросс-валидации и AIC_c-метода в целом ряде ситуаций.

Приложение для многомерных смешанных данных

Рассмотрим множественный регрессионный анализ при наличии качественных переменных. Этот пример взят со с. 226 учебника Wooldridge (2003).

Рассмотрим уравнение почасовой заработной платы, для которого зависимой переменной является log(wage) (lwage), а объясняющие переменные включают три непрерывных величины, а именно educ (число лет образования), exper (число лет опыта работы) и tenure (число лет работы у текущего работодателя) и две качественные величины – female (женский/мужской пол) и married (женат, замужем/не женат, не замужем). В этом примере имеется $n = 526$ наблюдений. Для выбора ширины окна используется AIC_c-метод из Hurvich, Simonoff & Tsai (1998), чьи результаты представлены в таблице 3.

На Рис. 13 изображены графики частных регрессий. График частной регрессии – это просто двумерный график зависимости y от одного из регрессоров x_j, когда все остальные регрессоры фиксированы на уровне соответствующих медиан/мод. Также изображены бутстраповские доверительные интервалы, которые часто предпочитают интервалам, полученным с помощью асимптотической аппроксимации.[12]

[12] Асимптотические формулы основаны на приближениях для малых значений h. Как отмечалось, иногда оптимальное сглаживание может справедливо требовать $h \to \infty$. Поскольку этого нельзя знать заранее,

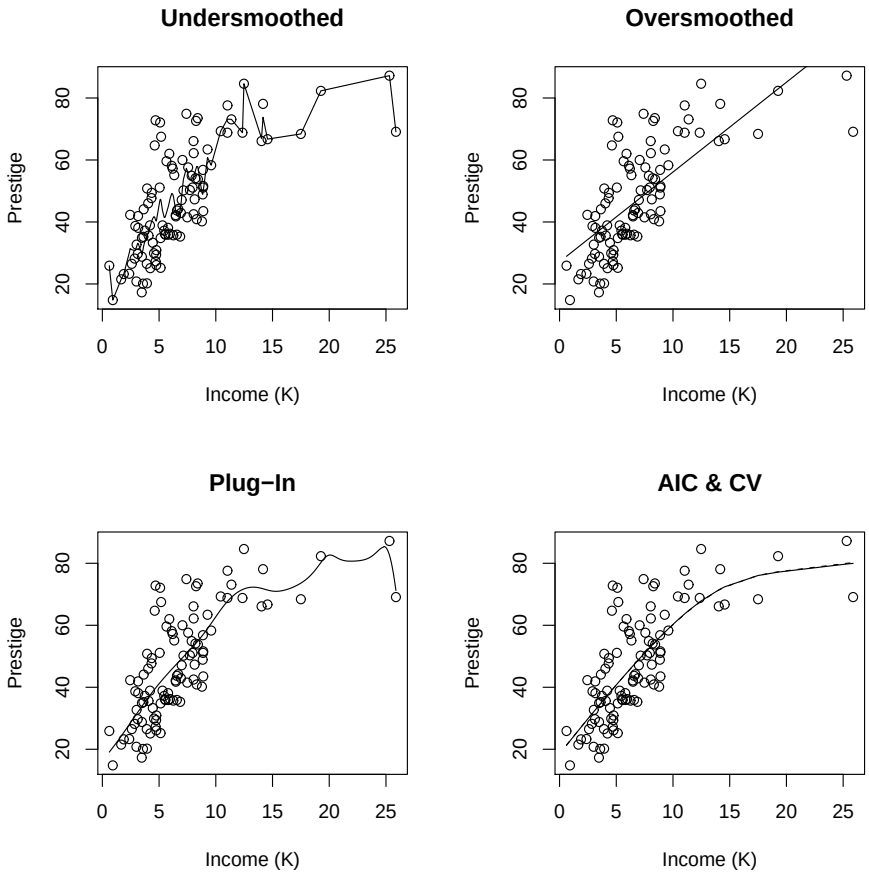

Рис. 12: Локально линейные ядерные оценки при различной ширине окна. Ширина окон соответствует недостаточному сглаживанию ($0{,}1\sigma n^{-1/5}$), чрезмерному сглаживанию ($10^3\sigma n^{-1/5}$), AIC_c-методу и кросс-валидации ($3{,}54\sigma n^{-1/5}$, $3{,}45\sigma n^{-1/5}$), методу подстановки ($1{,}08\sigma n^{-1/5}$).

На Рис. 14 представлены графики частных откликов с соответствующими бутстраповскими доверительными интервалами.

Отметим, что для двух категориальных переменных градиент вычисляется как разность заработных плат при остальных переменных, фиксированных на уровнях соответствующих медиан/мод, когда один из регрессоров меняет значение, скажем, с «незамужем» на «замужем». Заметим, что для крайнего левого значения каждой из характеристик («женский пол» и «женат/замужем») эта разность равна нулю, так как разность берется между значением, принимаемым переменной, и первым уровнем для каждой переменной; см. построение отклика в случае категориальных переменных в Racine, Hart & Li (2003).

4.3 Оценка качества подгонки

Необходим безразмерный показатель качества подгонки для моделей непараметрической регрессии, который был бы сравним с соответствующей мерой для параметрических регрессионных моделей, а именно с R^2. Отметим, что, естественно, им будет *внутривыборочная* мера качества подгонки. Учитывая недостатки подсчета R^2 на основе разложения суммы квадратов (такие как возможные отрицательные значения), существует альтернативное определение и метод расчета R^2, который напрямую применим к любой модели, линейной или нелинейной. Полагая, что Y_i обозначает исход, а $\hat{Y}_i$ – подогнанное значение для наблюдения

асимптотические аппроксимации в этом случае, естественно, будут низкого качества.

Таблица 3: Результаты выбора ширины окон для уравнения почасовой заработной платы.

```
Regression Data (526 observations, 5 variable(s)):

Regression Type: Local Linear
Bandwidth Selection Method: Expected Kullback-Leibler Cross-Validation
Formula: lwage ~ factor(female)+factor(married)+educ+exper+tenure
Bandwidth Type: Fixed
Objective Function Value: -0.8570284 (achieved on multistart 5)

factor(female)  Bandwidth:  0.01978275   Lambda Max:  0.500000
factor(married) Bandwidth:  0.15228887   Lambda Max:  0.500000
educ            Bandwidth:  7.84663015 Scale Factor:  6.937558
exper           Bandwidth:  8.43548175 Scale Factor:  1.521636
tenure          Bandwidth: 41.60546059 Scale Factor: 14.099208

Continuous Kernel Type: Second-Order Gaussian
No. Continuous Explanatory Vars.: 3

Unordered Categorical Kernel Type: Aitchison and Aitken
No. Unordered Categorical Explanatory Vars.: 2
```

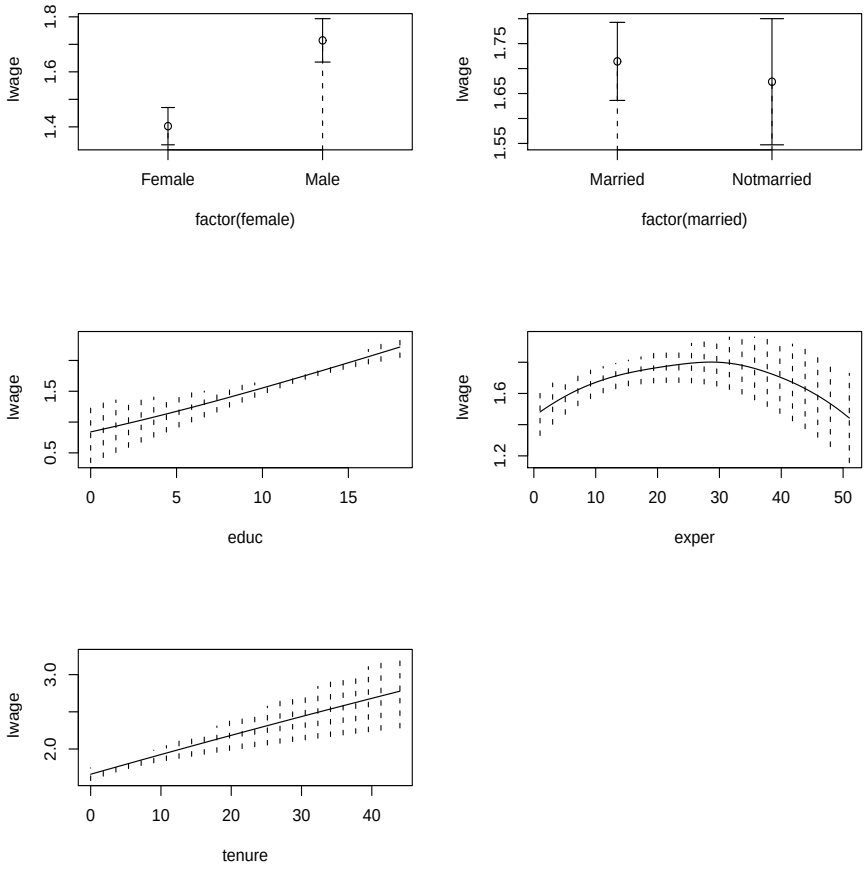

Рис. 13: Графики частных локально линейных непараметрических регрессий с бутстраповскими точечными доверительными интервалами для данных из Mroz (1987).

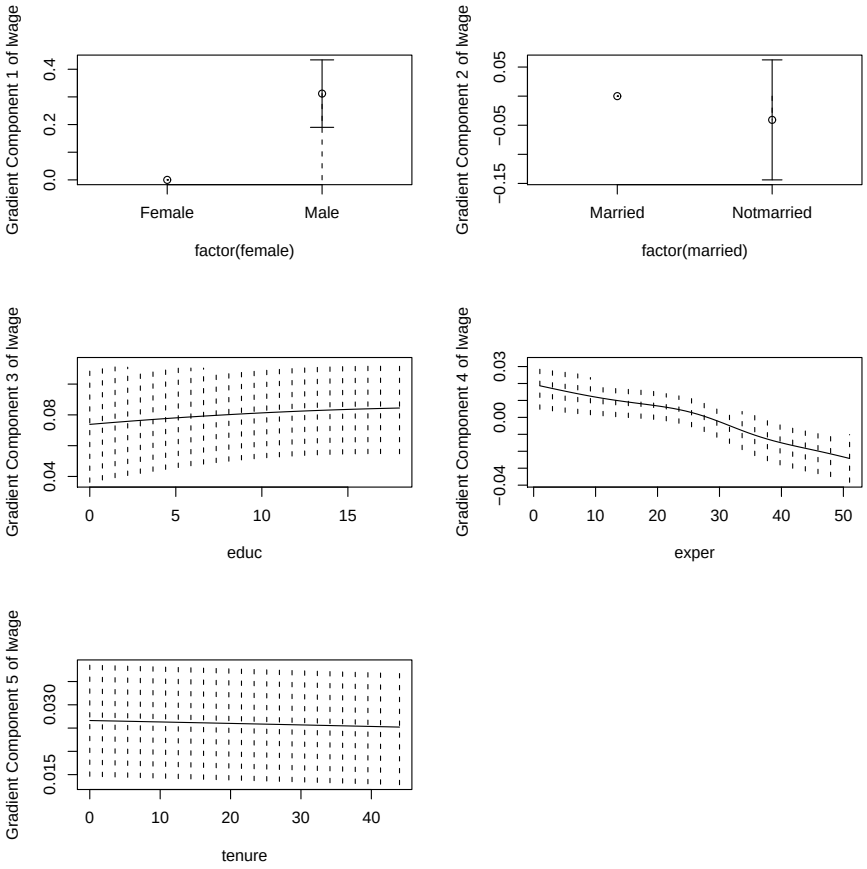

Рис. 14: Графики частных локально линейных непараметрических откликов с бутстраповскими точечными доверительными интервалами для данных из Mroz (1987).

i, можно определить R^2 следующим образом:

$$R^2 = \frac{\left[\sum_{i=1}^n (Y_i - \bar{y})(\hat{Y}_i - \bar{y})\right]^2}{\sum_{i=1}^n (Y_i - \bar{y})^2 \sum_{i=1}^n (\hat{Y}_i - \bar{y})^2},$$

и эта мера будет *всегда* принадлежать отрезку $[0, 1]$ и при равенстве единице означать точную подгонку данных, а при равенстве нулю – отсутствие предсказательной силы выше безусловного среднего зависимой переменной. Можно показать, что этот метод подсчета R^2 идентичен стандартному показателю, равному $\sum_{i=1}^n (\hat{Y}_i - \bar{y})^2 / \sum_{i=1}^n (Y_i - \bar{y})^2$, *если модель линейна, включает константу и для подгонки применяется МНК*. Эта полезная мера позволяет прямое сравнение качества подгонки внутри выборки, с очевидной оговоркой, что это ни в коем случае не критерий выбора модели, а просто показатель, который можно приводить в качестве результата оценивания. Этот показатель, конечно, можно также рассчитать, используя прогнозы и реализации вне выборки. Если рассматривать модели, оцениваемые на случайно выбранном подмножестве данных, а затем примененные к независимой выборке оставленных в стороне данных, эту меру, рассчитанную для отложенных данных, можно использовать для оценки различных моделей, особенно при усреднении по множеству независимых выборок отложенных данных.[13]

[13]Существует множество альтернативных показателей качества подгонки, рассчитываемых пакетом `npreg`. Для выяснения подробностей используйте команду `?npreg` в программной среде `R`.

Для рассмотренного выше примера со с. 226 учебника Wooldridge (2003) в случае локально линейной модели R^2 равен 51,5% при использовании этой меры, что можно напрямую сравнивать с *нескорректированным* R^2 в случае параметрической модели.

4.4 Устойчивая локально постоянная регрессия

Непараметрические ядерные методы часто (не без основания) критикуют из-за отсутствия робастности в традиционном смысле, а именно, к присутствию загрязнения в данных, которое может возникать из-за ошибок измерения, ввода данных и т.п. Методы, которые являются робастными в этом смысле, часто называют «устойчивыми», поскольку они «устойчивы» к присутствию небольшого числа плохих данных. Leung (2005) недавно предложил новый метод устойчивой робастной ядерной регрессии. Это новая захватывающая разработка, которая заслуживает внимания.

4.4.1 Устойчивая ядерная регрессия Леунга (2005)

Пусть $\{X_i, Y_i\}_{i=1}^n$ обозначает выборку. Рассмотрим регрессию Y на X в n выбранных точках $\{X_i\}_{i=1}^n$,

$$Y_i = g(X_i) + \epsilon_i, \qquad i = 1, \ldots, n, \tag{14}$$

где $g(\cdot)$ – неизвестная функция от X, а $\{\epsilon_i\}_{i=1}^n$ – IID случайные ошибки, имеющие распределение $F(\cdot)$.

Локально постоянная ядерная оценка $g(x)$, обозначаемая как $\hat{g}_h(x)$, имеет вид

$$\hat{g}_h(x) \equiv \arg\min_a \sum_{i=1}^n (Y_i - a)^2 K\left(\frac{X_i - x}{h}\right), \tag{15}$$

где h – ширина окна, определяющая степень локального сглаживания, а $K(\cdot)$ – ядерная функция, такая что $\int K(z)\,dz = 1$, $\int zK(z)\,dz = 0$ и $\int z^2K(z)\,dz = \mu < \infty$, например. Главный вопрос прикладных исследований – как лучше выбрать h.

Устойчивую локально постоянную ядерную оценку, с другой стороны, можно получить следующим образом:

$$\tilde{g}_{h|c}(x) \equiv \arg\min_a \sum_{i=1}^n \rho_c(Y_i - a) K\left(\frac{X_i - x}{h}\right). \tag{16}$$

где ρ_c – это, например, функция Хубера (Huber, 1964), на которой основаны M-оценки (см. с. 26 в Maronna, Martin & Yohai, 2006):

$$\rho_c(u) = \begin{cases} u^2, & \text{если } |u| \le c, \\ 2c|u| - c^2, & \text{если } |u| > c. \end{cases} \tag{17}$$

Чтобы вычислить $\tilde{g}_{h|c}(x)$, необходимо выбрать параметр устойчивости c. Одно популярное прикладное правило – $c = 1{,}345 \times s$, где s – робастная мера масштаба, такая как медианное абсолютное отклонение от медианы (MAD). Это популярное правило обеспечивает 95%-ую эффективность относительно гомоскедастичной нормальной модели в проблеме местоположения. Конечно, этот подход вычислительно более требовательный, чем методы, рассмотренные в главе *Регрессия*. Однако убедительные симуляции и приложения, приведенные в работе Leung (2005), указывают на то, что этот метод заслуживает внимания тех исследователей, которые обеспокоены наличием выбросов в данных.

Связанные по тематике работы включают Stone (1977) и Cleveland (1979), в которых рассматривается устойчивая локально полиномиальная подгонка на основе взвешенного МНК,[14] Cantoni & Ronchetti (2001), которые обсуждают сглаживание сплайнами при устойчивом выборе параметра сглаживания по схеме из Leung (2005), Fan & Jiang (2000), которые рассматривают устойчивые одношаговые локально полиномиальные оценки, но не обсуждают проблему выбора ширины окна, и Wang & Scott (1994), которые исследуют локально взвешенные полиномы, подогнанные с помощью методов линейного программирования. См. также Čížek & Härdle (2006), которые рассматривают устойчивое оценивание понижающих размерность регрессионных моделей.

Литература об устойчивых ядерных методах – важное продвижение, которое способно обновить методы ядерного сглаживания в интересном направлении, что привело бы к множеству по-настоящему устойчивых методов.

5 Полупараметрическая регрессия

Полупараметрические методы являются одними из наиболее популярных методов гибкого оценивания. Полупараметрические модели формируются при комбинировании параметрических и непараметрических моделей определенным способом. Такие модели полезны в ситуациях, когда полностью непараметрические модели срабатывают не очень хорошо, например, если проклятие размерности ведет к большой вариабельности оценок или исследователь хочет использовать параметрическую модель регрессии, но функциональная форма для части регрессоров или, возможно, плотность распределения ошибок неизвестна. Можно также представить себе ситуацию, в которой некоторые регрессоры входят в уравнение линейно (то есть, есть линейность по переменным), но функциональная форма параметров относительно других переменных неизвестна, или, возможно, регрессионная функция непараметрическая, а структура ошибок имеет параметрическую форму.

Полупараметрические модели являются компромиссом между полностью непараметрическими и полностью параметрическими спецификациями. Они основаны на параметрических предположениях и, следовательно, могут быть неправильно специфицированными и несостоятельными, так же как и их параметрические аналоги.

Было предложено множество полупараметрических методов. Далее ограничим внимание моделями регрессионного типа и рассмотрим три популярных метода, а именно: частично линейные, одноиндексные модели и модели с переменными коэффициентами.

5.1 Частично линейные модели

Частично линейная модель – одна из простейших полупараметрических моделей, применяемых на практике, – была предложена Робинсоном (Robinson, 1988), а Racine & Liu (2007) расширили его подход для случая категориальных регрессоров. Многие считают, что, поскольку модель очень проста, связанные с ней вычисления также должны быть простыми. Однако эта простота скрывает тот возможно недооцениваемый факт, что выбор ширины окна для частично линейных моделей может быть на несколько порядков сложнее вычислительно, чем для полностью непараметрических моделей, по одной простой причине. Как станет видно, в данном контексте применяются диктуемые данными методы выбора ширины окна, такие как кросс-валидация, и частично линейная модель требует от кросс-валидации регрессировать Y на Z (Z многомерный), а затем *каждый столбец* X на Z, тогда как кросс-валидация в случае полностью непараметрической модели подразумевает только регрессию

[14] Их метод “lowess” означает «локально взвешенную регрессию». Устойчивость следует из итеративной подгонки, где присваиваемые веса обратно пропорциональны остаткам из предыдущей подгонки, что ведет к снижению весов выбросов.

Y на X. Таким образом, вычислительные требования, связанные с частично линейными моделями, являются гораздо более серьезными, чем для полупараметрических моделей, о чем следует знать заранее.

Полупараметрическая частично линейная модель задается следующим образом:

$$Y_i = X_i'\beta + g(Z_i) + u_i, \quad i = 1, \ldots, n, \tag{18}$$

где X_i – $p \times 1$ вектор, β – $p \times 1$ вектор неизвестных параметров, и $Z_i \in \mathbb{R}^q$. Функциональная форма $g(\cdot)$ не специфицирована. Конечномерный параметр β составляет параметрическую часть модели, а неизвестная функция $g(\cdot)$ – непараметрическую часть. Предполагается, что данные являются IID с $\mathbb{E}[u_i|X_i, Z_i] = 0$, а процесс для ошибок условно гетероскедастичен с $\mathbb{E}[u_i^2|x, z] = \sigma^2(x, z)$ неизвестной формы. Обсудим, каким образом получить $\sqrt{n}$-состоятельную оценку β, поскольку, сделав это, оценку $g(\cdot)$ легко получить с помощью непараметрической регрессии $Y_i - X_i\hat{\beta}$ на z.

Взяв математическое ожидание (18) условно на Z_i, получим

$$\mathbb{E}[Y_i|Z_i] = \mathbb{E}[X_i|Z_i]'\beta + g(Z_i). \tag{19}$$

Вычитая (19) из (18), имеем

$$Y_i - \mathbb{E}[Y_i|Z_i] = (X_i - \mathbb{E}[X_i|Z_i])'\,\beta + u_i. \tag{20}$$

Обозначая $\tilde{Y}_i = Y_i - \mathbb{E}[Y_i|Z_i]$ и $\tilde{X}_i = X_i - \mathbb{E}[X_i|Z_i]$ и применяя МНК к (20), получаем оценку β:

$$\hat{\beta}_{\text{inf}} = \left[\sum_{i=1}^{n} \tilde{X}_i\tilde{X}_i'\right]^{-1} \sum_{i=1}^{n} \tilde{X}_i\tilde{Y}_i. \tag{21}$$

Данная оценка $\hat{\beta}_{\text{inf}}$ недоступна, так как $\mathbb{E}[Y_i|Z_i]$ и $\mathbb{E}[X_i|Z_i]$ неизвестны. Известно, однако, что эти условные матожидания можно состоятельно оценить, используя ядерные методы, описанные в главе *Регрессия*, так что можно заменить неизвестные условные матожидания в формуле для $\hat{\beta}_{\text{inf}}$ их ядерными оценками, получая, таким образом, доступную оценку β. Необходимы также некоторые условия для идентификации вектора параметров β, подробности см. в Robinson (1988).

Частично линейный пример

Рассмотрим снова набор данных “wage1” из учебника Wooldridge (2002), но предположим теперь, что исследователь не хочет специфицировать природу взаимосвязи между переменными exper и lwage и, следовательно, переносит переменную exper в непараметрическую часть полупараметрической линейной модели. В таблице 4 представлены результаты оценивания частично линейной модели.

Интересно сравнить эти результаты с результатами для линейной модели, которая квадратична по опыту, приведенными в таблице 5, и для локально линейной модели, представленными в главе *Регрессия*. Вначале отметим, что оценки параметров и соответствующие стандартные ошибки сравнимы по величине с результатами для полностью параметрической спецификации, которые содержатся в таблице 5. Во-вторых, в терминах внутривыборочной подгонки, полупараметрическая частично линейная спецификация ($R^2 = 44{,}9\%$) немного лучше параметрической спецификации ($R^2 = 43{,}6\%$), тогда как полностью непараметрическая спецификация ($R^2 = 51{,}5\%$) превосходит как полностью параметрическую, так и частично линейную спецификации.

Таблица 4: Результаты оценивания частично линейного уравнения почасовой заработной платы.

```
Partially Linear Model
Regression data: 526 training points, in 5 variable(s)
With 4 linear parametric regressor(s), 1 nonparametric regressor(s)

                 y(z)
Bandwidth(s): 2.050966

                  x(z)
Bandwidth(s): 4.1943673
              1.3531783
              3.1605552
              0.7646561

               factor(female) factor(married)       educ     tenure
Coefficient(s):      0.2902499     -0.03722828 0.07879512 0.01662935
Standard error(s):   0.0359527      0.04230253 0.00676465 0.00308927

Kernel Regression Estimator: Local Constant
Bandwidth Type: Fixed

Residual standard error: 0.1553021
R-squared: 0.4493789
```

Таблица 5: Результаты оценивания полностью линейного уравнения почасовой заработной платы.

```
Coefficients:
                            Estimate Std. Error
(Intercept)                0.1811615  0.1070747
factor(female)Male         0.2911303  0.0362832
factor(married)Notmarried -0.0564494  0.0409259
educ                       0.0798322  0.0068273
tenure                     0.0160739  0.0028801
exper                      0.0300995  0.0051931
I(exper^2)                -0.0006012  0.0001099

Multiple R-Squared: 0.4361,     Adjusted R-squared: 0.4296
```

5.2 Индексные модели

Полупараметрическая индексная модель имеет вид

$$Y = g(X'\beta_0) + u, \tag{22}$$

где Y – зависимая переменная, $X \in \mathbb{R}^q$ – вектор объясняющих переменных, β_0 – $q \times 1$ вектор неизвестных параметров, и u – ошибка, удовлетворяющая $\mathbb{E}[u|X] = 0$. Член $X'\beta_0$ называется одиночным индексом, так как это одномерная величина, хотя X – вектор. Функциональная форма $g(\cdot)$ неизвестна исследователю. Модель является полупараметрической по природе, так как функциональная форма линейного индекса специфицирована, а $g(\cdot)$ – нет.

Ichimura (1993), Manski (1988) и с. 14–20 в Horowitz (1998) дают превосходные интуитивные объяснения условий идентификации для полупараметрических одноиндексных моделей (то есть условий, при которых неизвестный вектор параметров β_0 и неизвестную функцию $g(\cdot)$ можно разумно оценить), заинтересованный читатель может обратиться к этим источникам за подробностями.

5.2.1 Метод Ичимуры

Рассмотрим случай непрерывного Y. Если бы функциональная форма $g(\cdot)$ была известна, мы получили бы стандартную модель нелинейной регрессии, и могли бы использовать нелинейный МНК для оценки β_0 путем минимизации

$$\sum_i (Y_i - g(X_i'\beta))^2 \tag{23}$$

по параметру β.

В случае неизвестной функции $g(\cdot)$, сначала требуется оценить $g(\cdot)$. Однако ядерные методы не способны напрямую оценить $g(X_i'\beta_0)$, поскольку неизвестны ни функция $g(\cdot)$, ни β_0. Тем не менее, для фиксированного значения β можно оценить

$$G(X_i'\beta) \stackrel{\text{def}}{=} \mathbb{E}[Y_i|X_i'\beta] = \mathbb{E}[g(X_i'\beta_0)|X_i'\beta] \tag{24}$$

с помощью ядерного метода, где последнее равенство следует из того, что $\mathbb{E}[u_i|X_i'\beta] = 0$ для всех β, поскольку $\mathbb{E}[u_i|X_i] = 0$.

Заметим, что при $\beta = \beta_0$, $G(X_i'\beta_0) = g(X_i'\beta_0)$, тогда как в общем случае $G(X_i'\beta) \neq g(X_i'\beta_0)$, если $\beta \neq \beta_0$. Ichimura (1993) предложил оценивать $g(X_i'\beta_0)$ как $\hat{G}_{-i}(X_i'\beta)$, выбирая β с помощью (полупараметрического) нелинейного МНК, где $\hat{G}_{-i}(X_i'\beta)$ – непараметрическая ядерная оценка $G(X_i'\beta)$ по всей выборке, исключая наблюдение i.

Одноиндексный пример для непрерывного Y

Рассмотрим далее применение одноиндексного метода Ичимуры (Ichimura, 1993), который подходит для ситуации непрерывной зависимой переменной, в отличие от рассматриваемого ниже метода Клейна и Спэйди (Klein & Spady, 1993). Снова используем набор данных “wage1” из учебника Wooldridge (2002). В таблице 6 представлены результаты анализа.

Таблица 6: Результаты оценивания полупараметрической индексной модели уравнения почасовой заработной платы.

```
Single Index Model
Regression Data: 526 training points, in 6 variable(s)

      factor(female) factor(married)     educ    exper    expersq   tenure
Beta:              1       -2.783907 9.947963 3.332755 -0.0750266 2.310801
Bandwidth: 2.457583
Kernel Regression Estimator: Local Constant

Residual standard error: 0.1552531
R-squared: 0.4501873
```

Интересно сравнить эти результаты с результатами для параметрической и непараметрической моделей, рассмотренных ранее, поскольку в данном случае мы получаем меру внутривыборочной подгонки равной 45,1%, которая лежит между соответствующими величинами

для параметрической (43,6%) и полностью непараметрической локально линейной (51,5%) моделей.

5.2.2 Оценка Клейна–Спэйди

Рассмотрим случай, когда переменная Y бинарна. Предполагая независимость ϵ_i и X_i, Klein & Spady (1993) предложили оценивать β методом максимального правдоподобия. Оценка логарифмической функции правдоподобия имеет вид

$$\mathcal{L}(\beta, h) = \sum_i (1 - Y_i) \ln(1 - \hat{g}_{-i}(X_i'\beta)) + \sum_i Y_i \ln(\hat{g}_{-i}(X_i'\beta)), \tag{25}$$

где $\hat{g}_{-i}(X_i'\beta)$ – оценка по всей выборке за исключением i-го наблюдения. Максимизация (25) по β и h дает полупараметрическую ММП-оценку β, предложенную в Klein & Spady (1993). Как и в случае метода Ичимуры, максимизацию необходимо проводить численными методами.

Одноиндексный пример для бинарного Y

Рассмотрим вновь данные о весе при рождении, взятые из библиотеки `MASS` (Venables & Ripley, 2002) программной среды `R`, и оценим одноиндексную модель (результаты оценивания логит-модели и модели непараметрической условной моды даны в главе *Оценивание условной плотности*). Исход является индикатором низкого веса младенца при рождении (0/1), так что подход Клейна и Спэйди применим. Таблица расхождения приведена в таблице 7.

Таблица 7: Таблица расхождения для данных по низкому весу при рождении при использовании одноиндексной модели.

		Предсказано	
		0	1
Реально	0	125	5
	1	37	22

Можно увидеть, что при внутривыборочной классификации эта модель ведет себя несколько лучше, чем параметрическая логит-модель, для этого набора данных. Одноиндексная модель правильно классифицирует $(125 + 22)/189 = 77{,}8\%$ случаев низкого/высокого веса новорожденных, тогда как логит-модель правильно классифицирует $(119 + 25)/189 = 76{,}1\%$.

5.3 Модели с гладкими (переменными) коэффициентами

Модель с гладкими коэффициентами была предложена в работе Hastie & Tibshirani (1993) и имеет следующий вид:

$$Y_i = \alpha(Z_i) + X_i'\beta(Z_i) + u_i = (1\, X_i') \begin{pmatrix} \alpha(Z_i) \\ \beta(Z_i) \end{pmatrix} + u_i = W_i'\gamma(Z_i) + u_i, \tag{26}$$

где X_i – $k \times 1$ вектор, а $\beta(z)$ – вектор неспецифицированных гладких функций от z. Умножая слева на W_i и беря матожидание по Z_i, имеем

$$\mathbb{E}[W_i Y_i | Z_i] = \mathbb{E}[W_i W_i' | Z_i]\gamma(Z_i) + \mathbb{E}[W_i u_i | Z_i]. \tag{27}$$

Можно выразить $\gamma(\cdot)$ как

$$\gamma(Z_i) = (\mathbb{E}[W_i W_i' | Z_i])^{-1} \mathbb{E}[W_i Y_i | Z_i]. \tag{28}$$

Li & Racine (2007b) рассматривают ядерный подход, который годится как для дискретных, так и для непрерывных регрессоров. Авторы предлагают использовать локально постоянную оценку вида

$$\hat{\gamma}(z) = \left[\sum_{j=1}^{n} W_j W_j' K\left(\frac{Z_j - z}{h}\right)\right]^{-1} \sum_{j=1}^{n} W_j Y_j K\left(\frac{Z_j - z}{h}\right)$$

и предлагают вариант кросс-валидации для выбора ширины окна; см. подробности в Li & Racine (2007b). Подогнанная модель имеет вид

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

Пример применения модели с гладкими коэффициентами

Рассмотрим снова набор данных “wage1” из учебника Wooldridge (2002), но предположим теперь, что исследователь не хочет предполагать, что коэффициенты при непрерывных переменных неизменны относительно категориальных переменных female и married. В таблице 8 представлены результаты оценивания модели с гладкими коэффициентами.

Таблица 8: Результаты оценивания уравнения почасовой заработной платы в рамках модели с гладкими коэффициентами.

```
Smooth Coefficient Model
Regression data: 526 training points, in 2 variable(s)

              factor(female) factor(married)
Bandwidth(s):    0.001813091       0.1342957

Bandwidth Type: Fixed

Residual standard error: 0.1470017
R-squared: 0.4787102

Average derivative(s):
    Intercept          educ        tenure         exper       expersq
 0.3402224978  0.0786499683  0.0142981775  0.0300505722 -0.0005950969
```

Сравнивая эти результаты с результатами для линейной модели, квадратичной по опыту работы, представленными в таблице 5, получаем, что средние значения параметров сравнимы по величине с полученными из полностью параметрической спецификации, приведенными в таблице 5. Однако полупараметрическая модель с гладкими коэффициентами превосходит параметрическую спецификацию в терминах внутривыборочной подгонки данных ($R^2 = 47{,}8\%$ против $R^2 = 43{,}6\%$). Это означает, что дополнительная гибкость, получаемая за счет возможности варьирования параметров относительно непрерывных переменных, ведет к улучшению качества подгонки.

6 Модели панельных данных

Непараметрическое и полупараметрическое оценивание моделей панельных данных получило меньше внимания, чем оценивание стандартных моделей регрессии. Панельные данные представляют собой выборки, состоящие из наблюдений за N пространственными единицами в течение T последовательных периодов времени, что дает набор данных в форме

$\{Y_{it}, X_{it}\}_{i=1,t=1}^{N,T}$. Панель, таким образом, – это просто набор N индивидуальных временных рядов, которые могут быть короткими (при малом T) или длинными (при большом T).

Непараметрическое оценивание моделей временных рядов само по себе является развивающейся областью. Однако, если T велико и N мало, для каждой единицы имеется длинный временной ряд и в таких случаях можно избежать оценивания модели панельных данных, просто оценивая по отдельности непараметрические модели для каждой индивидуальной единицы, используя индивидуальные временных ряды, доступные для каждой из них. Если имеет место такая ситуация, заинтересованный читатель может обратиться к главе 18 в Li & Racine (2007) за ссылками на литературу о непараметрических методах оценивания временных рядов.

При рассмотрении непараметрического оценивания моделей панельных данных одна из проблем, которая тут же возникает, состоит в том, что стандартные (параметрические) подходы, часто применяемые для моделей панельных данных (такие как взятие первых разностей для удаления так называемых фиксированных эффектов), более не являются корректными, если только не предполагать аддитивно сепарабельные эффекты, что во многом лишает смысла использование непараметрических методов.

В литературе было предложено множество подходов, включая работу Wang (2003), который предложил новый метод оценивания непараметрических моделей панельных данных, использующий информацию, содержащуюся в структуре ковариационной матрицы ошибок модели, работу Wang, Carroll & Lin (2005), которые предложили частично линейную модель со случайными эффектами, и работу Carroll, Henderson & Li (2006), которые рассматривают методы профиля правдоподобия для непараметрического оценивания моделей с аддитивными фиксированными эффектами, удаляемыми путем взятия первых разностей. Далее рассмотрим прямое непараметрическое оценивание моделей с фиксированными эффектами, используя методы, описанные в разделе *Регрессия*.

6.1 Непараметрическое оценивание панельных моделей с фиксированными эффектами

Рассмотрим следующую непараметрическую регрессионную модель панельных данных:

$$Y_{it} = g(X_{it}) + u_{it}, \quad i = 1, 2 \ldots, N,\ t = 1, 2, \ldots, T,$$

где $g(\cdot)$ – неизвестная гладкая функция, $X_{it} = (X_{it,1}, \ldots, X_{it,q})$ имеет размерность q, все другие переменные являются одномерными, и $\mathbb{E}[u_{it}|X_{i1}, \ldots, X_{iT}] = 0$.

Говорят, что панельные данные «однородны», если можно объединить все данные, в сущности игнорируя временную размерность, то есть суммируя по i и t, не обращая внимания на временную размерность и помещая таким образом все данные в один пул, а затем напрямую применяя методы, скажем, из главы *Регрессия*. Конечно, если данные неоднородны, это, очевидно, не будет разумным выбором.

Однако, чтобы допустить возможность того, что данные на самом деле *потенциально* однородны, можно ввести *неупорядоченную* дискретную переменную, скажем $\delta_i = i$, $i = 1, 2, \ldots, N$, и оценить $\mathbb{E}[Y_{it}|Z_{it}, \delta_i] = g(Z_{it}, \delta_i)$ непараметрически, используя ядерный подход для случая смешанных дискретных и непрерывных данных, рассмотренный в главе *Оценивание плотности и функции вероятности*. Переменная δ_i сродни включению пространственных фиктивных переменных (как поступают, например, при МНК-оценивании моделей линейной регрессии на панельных данных при включении фиктивных переменных). Если обозначить за $\hat{\lambda}$ полученный при кросс-валидации параметр сглаживания, связанный с δ_i, то, если $\hat{\lambda}$ достигает своей верхней границы, получаем $g(Z_{it}, \delta_i) = g(Z_{it})$, и данные, таким образом, объединяются в получаемой оценке $g(\cdot)$. Если, с другой стороны, $\hat{\lambda} = 0$ (или близка

к 0), получаем оценку каждой $g_i(\cdot)$, используя только временной ряд для i-й единицы наблюдения. Наконец, случай $0 < \hat{\lambda} < 1$, можно интерпретировать как частичную однородность данных.

Стоит отметить, что вдобавок к вопросу об однородности существует также проблема корректировки инференции на возможную серийную корреляцию ошибок u_{it}. То есть, даже если данные однородны, нельзя слепо применять асимптотический подход; адекватный бутстраповский подход, вероятно, на практике сработает лучше всего.

Применение к панельным данным об издержках авиакомпаний в США

Рассмотрим панель ежегодных наблюдений по шести американским авиакомпаниям за период с 1970 по 1984, взятую из пакета `Ecdat` (Croissant, 2006) программной среды `R`, см. таблицу F7.1 на с. 949 в учебнике Greene (2003). Переменными являются авиакомпания ("airline"), год ("year"), логарифм общих издержек в \$1000 ("lcost"), логарифм показателя выпуска, коммерческого оборота в пассажиро-милях ("loutput"), логарифм цены топлива ("lpf"), и коэффициент нагрузки, то есть средний коэффициент использования мощностей самолетного парка ("lf"). Переменная "airline" воспринимается как неупорядоченный показатель, а "year" – упорядоченный, и используется локально линейная оценка на основе AIC_c-метода из Hurvich, Simonoff & Tsai (1998).

В таблице 9 представлена информация о ширине окон, а на Рис. 15 – графики частных регрессий.

Таблица 9: Информация о ширине окон в локально линейной модели панельных данных по американским авиакомпаниям.

```
Var.: loutput          Bandwidth: 1020484   Scale Factor: 1.696225e+06
Var.: lpf              Bandwidth: 1417256   Scale Factor: 3.336533e+06
Var.: lf               Bandwidth: 0.0130355 Scale Factor: 0.472229
Var.: ordered(year)    Bandwidth: 0.1107695 Lambda Max:   1.000000
Var.: factor(airline) Bandwidth: 0.0024963 Lambda Max:   1.000000
```

Осмотр таблицы 9 показывает, что ширина окна для неупорядоченной переменной "airline" равна 0,0025, то есть модель не является однородной по авиакомпаниям (больше подходит модель отдельных временных рядов для каждой авиакомпании). Рис. 15 показывает, что издержки растут с объемом выпуска и ценой топлива и убывают по коэффициенту нагрузки.

Для сравнения в таблице 10 представлены результаты оценивания линейной модели панельных данных с фиксированными эффектами при использовании пакета `plm` (Croissant & Millo, 2007) программной среды `R`.

Сравнение Рис. 15 и таблицы 10 показывает, что как параметрическая, так и непараметрическая модели согласуются в том, что издержки растут с выпуском и ценой топлива и падают по коэффициенту нагрузки при прочих равных.

7 Состоятельное тестирование гипотез

Литература о применении непараметрических ядерных методов для тестирования гипотез испытала невероятный рост и породила множество новых подходов к тестированию ряда гипотез. Существуют непараметрические методы для тестирования на правильность спецификации параметрических моделей, тесты на совпадение распределений и регрессионных функций, среди прочих.

Параметрические тесты обычно требуют спецификации множества параметрических альтернатив, для которых нулевая гипотеза отвергается. Если, однако, нулевая гипотеза неверна

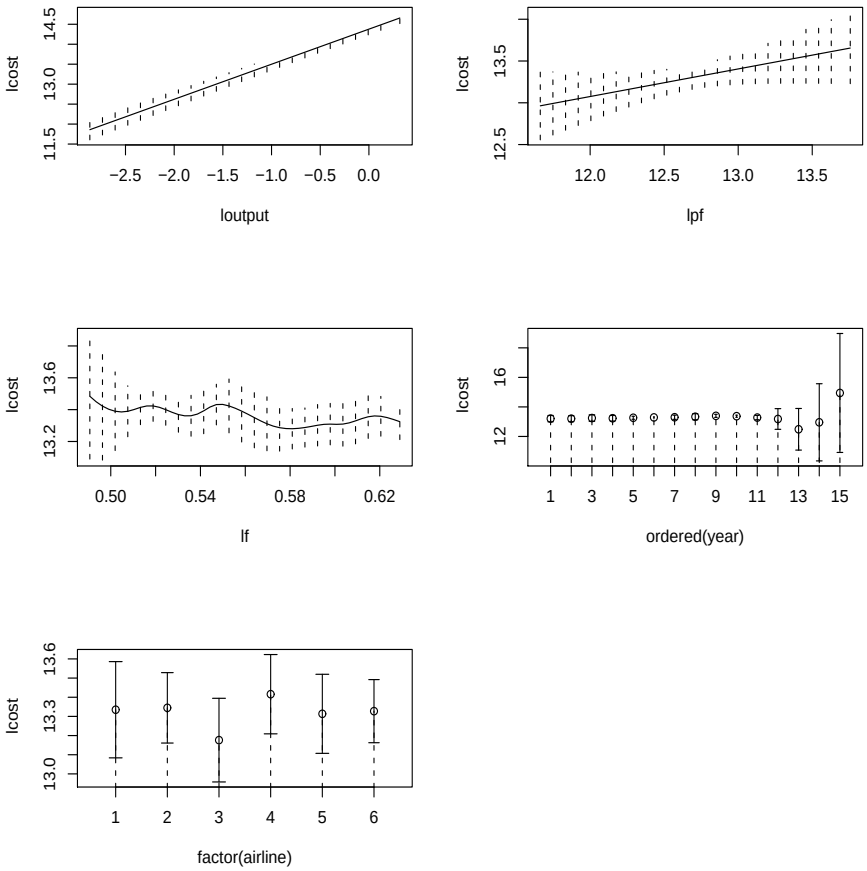

Рис. 15: Графики частных регрессий с бутстраповскими доверительными интервалами для панельных данных по американским авиалиниям.

Таблица 10: Результаты оценивания параметрической модели с фиксированными эффектами для панельных данных по американским авиалиниям.

```
--------------------------- Model Description ----------------------------
Oneway (individual) effect
Within Model
Model Formula            : log(cost) ~ log(output) + log(pf) + lf

------------------------------ Coefficients ------------------------------
             Estimate Std. Error z-value  Pr(>|z|)
log(output)  0.919285   0.028841 31.8743 < 2.2e-16 ***
log(pf)      0.417492   0.014666 28.4673 < 2.2e-16 ***
lf          -1.070396   0.194611 -5.5002 3.794e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

и тем не менее существуют альтернативные модели, которые тест не может обнаружить, тогда такой тест называют несостоятельным, то есть у него не хватает мощности. Для построения состоятельных тестов можно вместо параметрических использовать непараметрические методы.

Для точности определим, что имеется в виду под «состоятельным тестом». Пусть H_0 обо-

значает нулевую гипотезу, чью достоверность требуется проверить. Тест называют *состоятельным*, если

$$\mathbb{P}\{\text{отвергнуть } H_0 \mid H_0 \text{ неверна}\} \to 1 \text{ при } n \to \infty.$$

Мощность теста определяется как $\mathbb{P}\{\text{ отвергнуть } H_0 \,|\, H_0 \text{ неверна }\}$. Следовательно, состоятельный тест имеет асимптотическую мощность, равную единице.

Далее рассмотрим несколько тестов, которые могут пригодиться практикам.

7.1 Тестирование правильности спецификации параметрической модели

Существует множество методов тестирования правильности спецификации параметрических моделей регрессии, включая методы Härdle & Mammen (1993), Horowitz & Härdle (1994), Horowitz & Spokoiny (2001), Hristache, Juditsky & Spokoiny (2001) и Hsiao, Li & Racine (2007), среди прочих. Опишем кратко тест из Hsiao, Li & Racine (2007), поскольку он распространяется на случай смешанных непрерывных и категориальных данных, часто встречающийся в прикладных задачах.

Предположим, необходимо протестировать правильность параметрической модели регрессии. Нулевую гипотезу можно сформулировать следующим образом:

$$H_0 : \mathbb{E}[Y|x] = m(x, \gamma_0), \text{ почти для всех } x \text{ и для некоторого } \gamma_0 \in \mathcal{B} \subset \mathbb{R}^p, \tag{29}$$

где $m(x, \gamma)$ – известная функция, γ – $p \times 1$ вектор неизвестных параметров (что, естественно, включает модель линейной регрессии как частный случай), а $\mathcal{B}$ – компактное подмножество $\mathbb{R}^p$. Альтернативная гипотеза – отрицание H_0, то есть H_1: $\mathbb{E}[Y|x] \equiv g(x) \neq m(x, \gamma)$ для всех $\gamma \in \mathcal{B}$ на множестве (точек x) положительной меры. Если определить $u_i = Y_i - m(X_i, \gamma_0)$, нулевую гипотезу можно эквивалентно записать в виде

$$\mathbb{E}[u_i|X_i = x] = 0 \text{ почти для всех } x. \tag{30}$$

Состоятельный тест на спецификацию модели можно построить на основе непараметрического оценивания (30) и усреднения по u_i определенным способом, который мы кратко опишем. Во-первых, заметим, что $\mathbb{E}[u_i|X_i = x] = 0$ эквивалентно $(\mathbb{E}[u_i|X_i = x])^2 = 0$. Также, поскольку требуется тестировать нулевую гипотезу о том, что $\mathbb{E}[u_i|X_i = x] = 0$ почти для всех x, необходимо рассматривать матожидание $\mathbb{E}[\mathbb{E}[u_i|X_i = x]]$ или, эквивалентно, $\mathbb{E}[(\mathbb{E}[u_i|X_i = x])^2]$. По закону повторных матожиданий $\mathbb{E}[(\mathbb{E}[u_i|X_i = x])^2] = \mathbb{E}[u_i\mathbb{E}[u_i|X_i = x]]$. Следовательно, можно построить состоятельную тест-статистику на основе взвешенной плотностью версии $\mathbb{E}[u_i\mathbb{E}[u_i|X_i = x]]$, а именно $\mathbb{E}[u_i\mathbb{E}[u_i|X_i]f(X_i)]$, где $f(x)$ – совместная плотность X. Взвешивание плотностью используется просто для избежания случайного знаменателя, который в противном случае появился бы в ядерной оценке.

Выборочным аналогом $\mathbb{E}[u_i\mathbb{E}[u_i|X_i]f(X_i)]$ является выражение $n^{-1}\sum_{i=1}^n u_i\mathbb{E}[u_i|X_i]f(X_i)$. Для получения доступной тест-статистики заменим u_i на $\hat{u}_i$, где $\hat{u}_i = Y_i - m(X_i, \hat{\gamma})$ – остаток, полученный из параметрической модели при нулевой гипотезе, а $\hat{\gamma}$ – $\sqrt{n}$-состоятельная оценка γ из нулевой модели (скажем, оценка нелинейного МНК). Выражение $\mathbb{E}[u_i|X_i]f(X_i)$ оценим с помощью ядерного метода по всей выборке за исключением одного наблюдения $(n-1)^{-1}\sum_{j\neq i}^n \hat{u}_j K_{ij}$. Полагая, что X_i – вектор смешанных дискретных и непрерывных данных, и используя обобщенные мультипликативные ядра, получаем тест-статистику в виде

$$I_n \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^n \hat{u}_i \left\{ \frac{1}{n-1}\sum_{j=1, j\neq i}^n \hat{u}_j K_{ij} \right\} = \frac{1}{n(n-1)}\sum_{i=1}^n \sum_{j=1, j\neq i}^n \hat{u}_i\hat{u}_j K_{ij}. \tag{31}$$

Стьюдентизированная версия этого теста обозначается за J_n. Для получения распределения I_n (J_n) можно использовать бутстраповские методы и получать бутстраповские p-значения; см. подробности в Hsiao, Li & Racine (2007).

Тест на правильность спецификации наивной линейной модели

Оценив, скажем, простую параметрическую модель заработной платы, линейную по переменным, можно тестировать нулевую гипотезу о правильной спецификации параметрической модели, используя подход Hsiao, Li & Racine (2007), описанный выше. Для выбора ширины окон используется кросс-валидация, а для формирования распределения J_n при нулевой гипотезе используется простой бутстраповский метод.

Таблица 11: Результаты теста на правильность спецификации параметрического уравнения почасовой заработной платы.

```
Consistent Model Specification Test
Parametric null model: lm(formula = lwage ~
                                    factor(female) +
                                    factor(married) +
                                    educ +
                                    exper +
                                    tenure,
                                    data = wage1,
                                    x = TRUE,
                                    y = TRUE)

Number of regressors: 5
IID Bootstrap (399 replications)

Test Statistic 'Jn': 5.542416   P Value: < 2.22e-16
```

Таблица 11 дает понять, что, как неудивительно, наивная спецификация, линейная по всем переменным, не включающая переменные взаимодействия, и только лишь позволяющая константе меняться относительно категориальных переменных, отвергается. Заметим, что мы не продвигаем данную параметрическую спецификацию как идеального кандидата, а просто демонстрируем, что тест способен обнаруживать неверно специфицированные параметрические модели в конечных выборках.

7.2 Тест на значимость в непараметрических моделях регрессии

Оценив параметрическую модель регрессии, исследователи часто переходят к тесту на значимость. Тест на значимость часто применяется для подтверждения или опровержения экономических теорий. Тем не менее, в рамках параметрической модели регрессии, надежность параметрической инференции зависит от правильной функциональной спецификации соответствующего процесса, порождающего данные, а тесты на значимость для неверно специфицированных параметрических моделей будут иметь неверные размер и мощность, приводя, таким образом, к ошибочной инференции. В литературе предложено множество подходов, включая подход Lavergne & Vuong (1996), которые рассмотрели проблему выбора непараметрических регрессоров в рамках невложенных регрессионных моделей, работу Donald (1997), в которой предложен непараметрический тест для выбора факторов в многомерной непараметрической зависимости, работу Racine (1997), в которой рассмотрен тест на значимость для непрерывных регрессоров, и статью Racine, Hart & Li (2006), где рассмотрен состоятельный тест на значимость для категориальных регрессоров. См. также альтернативный непараметрический тест на значимость непрерывных переменных в непараметрических моделях регрессии в Delgado & Manteiga (2001).

7.2.1 Категориальные регрессоры

Предположим, мы оценили непараметрическую модель регрессии, в которой некоторые регрессоры категориальные, а некоторые непрерывные, и хотим протестировать, являются ли некоторые из категориальных регрессоров несущественными, то есть избыточными. Для этого можно применить тест из Racine, Hart & Li (2006), который коротко опишем. Пусть Z – категориальные объясняющие переменные, которые могут быть избыточными, X – остальные объясняющие переменные в модели регрессии, а Y – зависимая переменная. Тогда нулевую гипотезу можно записать как

$$H_0 : \mathbb{E}[Y|x,z] = \mathbb{E}[Y|x] \text{ почти везде.}$$

Альтернативная гипотеза – отрицание нулевой гипотезы H_0, то есть H_1: $\mathbb{E}[Y|x,z] \neq \mathbb{E}[Y|x]$ на множестве положительной меры.

Положим $g(x) = \mathbb{E}[Y|x]$ и $m(x,z) = \mathbb{E}[Y|x,z]$, тогда нулевая гипотеза примет вид: $m(x,z) = g(x)$ почти везде. Предположим, что одномерная величина Z может принимать c различных значений, $\{0,1,2,\ldots,c-1\}$. При $c = 2$ величина Z – это 0-1 фиктивная переменная, что, возможно, является наиболее часто встречающимся на практике случаем.

Заметим, что нулевая гипотеза H_0 эквивалентна $m(x, z = l) = m(x, z = 0)$ для всех X и для $l = 1,\ldots,c-1$. Тест-статистикой является оценка

$$I = \sum_{l=1}^{c-1} \mathbb{E}\left[m(x,z=l) - m(x,z=0)\right]^2.$$

Очевидно, $I \geq 0$, и $I = 0$ тогда и только тогда, когда H_0 верна. Значит, I служит адекватной мерой для тестирования H_0. Доступная тест-статистика имеет вид

$$I_n = \frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{c-1}\left[\hat{m}(X_i, Z_i = l) - \hat{m}(X_i, Z_i = 0)\right]^2, \tag{32}$$

где $\hat{m}(X_i, Z_i = l)$ – локально постоянная или локально линейная регрессионная оценка, описанная в главе *Регрессия*.

Легко показать, что I_n является состоятельной оценкой I. Следовательно, $I_n \to 0$ по вероятности при H_0, и $I_n \to I > 0$ по вероятности при H_1. Для получения распределения этой статистики при нулевой гипотезе или ее стьюдентизированной версии Racine, Hart & Li (2007) предложили две бутстраповские процедуры, обе из которых имеют хорошие свойства в конечных выборках; см. подробности в Racine, Hart & Li (2007).

7.2.2 Непрерывные регрессоры

Похожим образом нулевую гипотезу при тестировании на значимость непрерывного регрессора можно записать в виде

$$H_0 : \mathbb{E}[y|x,z] = \mathbb{E}[Y|z] \text{ почти везде,}$$

что эквивалентно

$$H_0 : \frac{\partial \mathbb{E}[y|x,z]}{\partial x} = \beta(x) = 0 \text{ почти везде}$$

Тест-статистикой является оценка

$$I = \mathbb{E}[\beta(x)^2]. \tag{33}$$

Тест-статистику можно получить, подсчитав выборочное среднее I с заменой неизвестных производных их непараметрическими оценками $\hat{\beta}(x_i)$, как описано в Racine (1997), то есть

$$I_n = \frac{1}{n}\sum_{i=1}^{n}\hat{\beta}(X_i)^2, \tag{34}$$

где $\hat{\beta}(X_i)$ – локально постоянная или локально линейная оценка частной производной, которые описаны в разделе *Регрессия*.

Легко показать, что I_n является состоятельной оценкой I. Следовательно, $I_n \to 0$ по вероятности при H_0, и $I_n \to I > 0$ по вероятности при H_1. Для получения распределения этой статистики при нулевой гипотезе или ее стьюдентизированной версии можно использовать бутстраповские процедуры; см. подробности в Racine (1997).

Иллюстрация

Рассмотрим набор данных "wage1" ($n = 526$) из учебника Wooldridge (2002), который содержит смесь непрерывных и категориальных регрессоров, и применим тесты на значимость, описанные выше. Результаты представлены в таблице 12, а p-значения указывают на значимость всех переменных на обычном 5%-ом уровне.

Таблица 12: Результаты теста на значимость для непараметрического локально линейного уравнения почасовой заработной платы.

```
Kernel Regression Significance Test
Type I Test with IID Bootstrap (399 replications)
Explanatory variables tested for significance:
factor(female) (1), factor(married) (2), educ (3), exper (4), tenure (5)

             factor(female) factor(married)    educ    exper   tenure
Bandwidth(s):     0.01978275       0.1522889 7.84663 8.435482 41.60546

Significance Tests
P Value:
factor(female)  < 2.22e-16 ***
factor(married) 0.0150376 *
educ            < 2.22e-16 ***
exper           < 2.22e-16 ***
tenure          0.0075188 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

8 Соображения по поводу вычислений

Ирония данной области эконометрики заключается в том, что ядерные методы по своей природе часто являются настолько вычислительно трудными, что большинство исследователей не стремятся применять их в тех случаях, для которых они идеально подходят, а именно когда наблюдается чрезвычайное изобилие данных, например, при оценке панельных микроданных, высокочастотных финансовых данных и т.д.

Для небольших наборов данных вычислительные трудности, связанные с ядерными методами, редко бывают проблемой. Тем не менее, для средних и больших наборов данных

вычисления, требуемые для реализации диктуемых данными методов выбора ширины окна, легко выходят из-под контроля. Естественно, интерес представляет общая задача ядерного оценивания, включающая оценивание безусловной и условной плотности, регрессии и производных как для категориальных, так и для непрерывных данных, и такая, которая использует ряд ядерных функций и методов выбора ширины окна.

Существует множество подходов к снижению вычислительного бремени, связанного с ядерными методами. На настоящий момент нам неизвестны методы, позволяющие осуществлять вычисления в рамках общей задачи ядерного оценивания в режиме реального времени или близкие к тому. Обсудим кратко некоторые подходы, существующие на данный момент, и другие, дающие надежду на прорыв, который позволит проводить ядерное оценивание в режиме реального времени с помощью обычного настольного или переносного компьютера.

Использование биннинговых методов

Использование биннинга дало множество вычислительно очень привлекательных оценок. Биннингом называют метод приближенных вычислений, когда данные предварительно сортируют на равномерной сетке, а затем к этим данным применяют подходящим образом модифицированную оценку. Например, биннинговые методы были предложены в Scott (1985), где для гладкого непараметрического оценивания плотности применяются усредненные сдвинутые гистограммы (ASH), а Scott & Sheather (1985) исследуют точность биннинговых методов при ядерном оценивании плотности.

Использование преобразований

На с. 61–66 в Silverman (1986) описано применение быстрых преобразований Фурье (FFT) для эффективного вычисления оценок (одномерной) плотности. Этот подход ограничивает оценивание сеткой точек (например, 512), чтобы ускорить скорость вычислений. Elgammal, Duraiswami & Davis (2003) обсуждают использование быстрых преобразований Гаусса (FGT) для эффективного вычисления ядерных оценок плотности с гауссовским ядром.

Использование параллелизма

Racine (2002) использует как параллельную природу большинства непараметрических методов, так и наличие нескольких вычислительных сред для достижения существенного снижения времени вычислений.

Использование многополюсных методов и методов ветвления

Две недавние разработки в области быстрых многополюсных методов и методов ветвления дают надежду на возможность общего ядерного оценивания в режиме реального времени. Чтобы расширить эти недавние разработки на случай общей задачи ядерного оценивания, надо, однако, проделать значительный объем работы. Многополюсные методы и методы ветвления были разработаны только для оценивания безусловной плотности и только для непрерывных данных.

Многополюсные методы представляют собой группу методов приближенных вычислений, обычных для задач в теории потенциального поля, когда n точек взаимодействуют в соответствии с некоторой потенциальной функцией, а целью является вычисление поля в произвольных точках (Greengard, 1988; Greengard & Strain, 1991). Для ускорения работы эти алгоритмы используют тот факт, что все вычисления требуется выполнить лишь с определенной степенью точности.

Методы ветвления можно представить себе как более мощные обобщения сетки, представляющие собой множество связанных сеток, построенных на разных разрешениях. Эта техника позволяет применение подхода «разделяй и властвуй», который может интегрировать локальную информацию для получения глобального решения, имеющего заданную точечную точность (Gray & Moore, 2003). Если *kd*-деревьям свойственно то, что представление в виде сетки имеет сложность, экспоненциально растущую с размерностью q, это не так для шаровидных методов ветвления (ball-trees), которые применялись в ситуациях с буквально тысячами размерностей.

Вызовы будущего

Святым граалем прикладного ядерного оценивания является разработка и реализация библиотеки, которая стала бы основой программного пакета, имеющего возможности, скажем, пакета `np`, но кроме этого использующего вычислительные преимущества лучших из перечисленных выше методов. Это, конечно, грандиозный проект, но он был бы тепло принят исследовательским сообществом.

9 Программное обеспечение

Для желающих заняться непараметрическим моделированием существует ряд возможностей. Ни один из известных нам программных пакетов не является универсальным, у них у всех не хватает функциональности. Во многих реализованы методы для двумерной плотности и регрессии, но недоступно оценивание объектов более высоких размерностей, тогда как другие пакеты подходят для более высоких размерностей, но в остальном имеют узкую сферу применения. Список, приведенный ниже, не представляет собой рекламу и не является исчерпывающим. Это лишь отправная точка для заинтересованного читателя.

- `EasyReg` (econ.la.psu.edu/∼hbierens/EASYREG.HTM) – программа для регрессионного анализа в среде Microsoft Windows, содержащая модуль для оценивания непараметрической ядерной регрессии с одной или двумя объясняющими переменными.
- `Limdep` (www.limdep.com) содержит, среди прочего, модули для ядерного оценивания плотности.
- `R` (www.r-project.org) содержит библиотеки, реализующие ряд ядерных методов, включая базовую библиотеку `stats`, библиотеки `KernSmooth` и `np`.
- `SAS` (www.sas.com) содержит модули для оценивания ядерной регрессии, локально взвешенного сглаживания и ядерного оценивания плотности.
- `Stata` (www.stata.com) содержит некоторые модули для оценивания одномерной и двумерной плотности и некоторые модули для оценивания локально постоянной ядерной регрессии.
- `TSP` (www.tspintl.com) содержит процедуры для оценивания одномерной плотности и простой ядерной регрессии.

10 Заключение

Непараметрические методы ядерного сглаживания за последние годы испытали огромный рост. Они применяются исследователями-практиками в целом ряде дисциплин. Непараметрические ядерные подходы представляют собой множество потенциально полезных методов для тех, кто вынужден сталкиваться с неприятной проблемой неправильной спецификации

параметрической модели. Привлекательность непараметрических методов состоит, главным образом, в их устойчивости к неправильной спецификации функциональных форм, в отличие от параметрических моделей. Хотя лежащая в основе многих из этих методов теория может показаться устрашающей для практика, мы попытались показать, как применять ряд непараметрических методов весьма непосредственным образом. Мы избегали всякой попытки энциклопедического покрытия всей области и вместо этого попытались направить заинтересованного читателя к учебникам, упомянутым во введении, и, конечно, оригинальным статьям из научных журналов. Представляя ряд полупараметрических и параметрических моделей, подходящих для множества прикладных задач, мы надеемся, что нам удалось воодушевить заинтересованных читателей на применение некоторых из этих методов в конкретных сферах их интересов.

Мы постарались подчеркнуть тот факт, 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. Тем не менее, как отмечено в разделе *Соображения по поводу вычислений*, существуют методы приближенных вычислений, которые имеют возможность значительно снизить объем необходимых вычислений, и мы поддерживаем всех тех, кто хочет внести свой вклад в эту область, так как их усилия были бы особенно полезны для практиков.

Благодарности

Я выражаю благодарность Биллу Грину (Bill Greene), Заку Ролнику (Zac Rolnick) и анонимному рецензенту за их образцовую редакторскую помощь и руководство. Я также хотел бы выразить признательность своим многочисленным соавторам и в особенности моему другу и соавтору Кви Ли (Qi Li), с которым я имел удовольствие успешно сотрудничать.

Наконец, я также хотел бы поблагодарить Единую иерархическую вычислительную сеть для академических исследований (SHARCNET:www.sharcnet.ca) за непрекращающуюся поддержку. Я также глубоко благодарен за финансовую поддержку Канадской программе поддержки исследований в области естественных и инженерных наук (NSERC:www.nserc.ca) и Канадской программе поддержки исследований в области общественных и гуманитарных наук (SSHRC:www.sshrc.ca).

Список литературы

Abramson, I.S. (1982). On bandwidth variation in kernel estimates – a square root law. *Annals of Statistics* 10, 1217–1223.

Aitchison, J. & C.G.G. Aitken (1976). Multivariate binary discrimination by the kernel method. *Biometrika* 63, 413–420.

Azzalini, A. & A.W. Bowman (1997). *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-plus Illustrations*. New York: Oxford University Press.

Bickel, P.J., C.A.J. Klaassen, Y. Ritov & J.A. Wellner (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Baltimore: Johns Hopkins University Press.

Bowman, A.W. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika* 71, 353–360.

Bowman, A.W., P. Hall & T. Prvan (1998). Bandwidth selection for the smoothing of distribution functions. *Biometrika* 85, 799–808.

Breiman, L., W. Meisel & E. Purcell (1977). Variable kernel estimates of multivariate densities. *Technometrics* 19, 135–144.

Cameron, A.C. & P.K. Trivedi (1998). *Regression Analysis of Count Data*. New York: Cambridge University Press.

Cantoni, E. & E. Ronchetti (2001). Resistant selection of the smoothing parameter for smoothing splines. *Statistics and Computing* 11, 141–146.

Cheng, M.-Y., P. Hall & D. Titterington (1997). On the shrinkage of local linear curve estimators. *Statistics and Computing* 7, 11–17.

Čížek, P. & W. Härdle (2006). Robust estimation of dimension reduction space. *Computational Statistics & Data Analysis* 51, 545–555.

Cleveland, W.S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of American Statistical Association* 74, 829–836.

Croissant, Y. (2006). *Ecdat: Data sets for econometrics*. R package version 0.1–5. URL: `http://www.r-project.org`

Croissant, Y. & G. Millo (2007). *plm: Linear models for panel data*. R package version 0.2–2. URL: `http://www.r-project.org`

Delgado, M.A. & W.G. Manteiga (2001). Significance testing in nonparametric regression based on the bootstrap. *Annals of Statistics* 29, 1469–1507.

Devroye, L. & L. Györfi (1985). *Nonparametric Density Estimation: The L^1 View*. New York: Wiley.

Donald, S.G. (1997). Inference concerning the number of factors in a multivariate nonparametric relationship. *Econometrica* 65, 103–131.

Draper, N. (1987). *Empirical Model-Building and Response Surfaces*. New York: Wiley.

Efromovich, S. (1999). *Nonparametric Curve Estimation: Methods, Theory and Applications*. New York: Springer Verlag.

Efron, B. (1982). *The Jackknife, the Bootstrap, and Other Resampling Plans*. Philadelphia: Society for Industrial and Applied Mathematics.

Elgammal, A., R. Duraiswami & L. Davis (2003). The fast gauss transform for efficient kernel density evaluation with applications in computer vision. *IEEE transactions on Pattern Analysis and Machine Intelligence.*

Epanechnikov, V.A. (1969). Nonparametric estimation of a multidimensional probability density. *Theory of Applied Probability* 14, 153–158.

Eubank, R.L. (1999). *Nonparametric Regression and Spline Smoothing*, 2nd edition. New York: Marcel Dekker.

Fan, J. (1992). Design-adaptive nonparametric regression. *Journal of American Statistical Association* 87, 998–1004.

Fan, J. & I. Gijbels (1996). *Local Polynomial Modelling and Its Applications*. London: Chapman and Hall.

Fan, J. & J. Jiang (2000). Variable bandwidth and one-step local m-estimator. *Science in China, Series A* 43, 65–81.

Fan, J. & Q.W. Yao (2005). *Nonlinear time series: Nonparametric and Parametric Methods*. New York: Springer Verlag.

Faraway, J. & M. Jhun (1990). Bootstrap choice of bandwidth for density estimation. *Journal of the American Statistical Association* 85, 1119–1122.

Fix, E. & J.L. Hodges (1951). Discriminatory analysis. Nonparametric estimation: Consistency properties. Technical Report, USAF School of Aviation Medicine, Randolph Field.

Fox, J. (2002). *An R and S-PLUS Companion to Applied Regression*. Thousand Oaks: Sage.

Geary, R. (1947). Testing for normality. *Biometrika* 34, 209–242.

Geisser, S. (1975). A predictive sample reuse method with application. *Journal of American Statistical Association* 70, 320–328.

Gray, A. & A.W. Moore (2003). Very fast multivariate kernel density estimation via computational geometry. In: *Proceedings Joint Statistical Meeting, Aug 3–7,2003.*

Greene, W.H. (2003). *Econometric Analysis*, 5th edition. Upper Saddle River: Prentice Hall.

Greengard, L. (1988). *The Rapid Evaluation of Potential Fields in Particle Systems*. Cambridge: MIT Press.

Greengard, L. & J. Strain (1991). The fast gauss transform. *Journal of Science and Computation* 12, 79–94.

Hall, P., Q. Li & J.S. Racine (в печати). Nonparametric estimation of regression functions in the presence of irrelevant regressors. *Review of Economics and Statistics.*

Hall, P., J.S. Racine & Q. Li (2004). Cross-validation and the estimation of conditional probability densities. *Journal of American Statistical Association* 99, 1015–1026.

Härdle, W. (1990). *Applied Nonparametric Regression.* New York: Cambridge University Press.

Härdle, W. & E. Mammen (1993). Comparing nonparametric versus parametric regression fits. *Annals of Statistics* 21, 1926–1947.

Härdle, W., M. Müller, S. Sperlich & A. Werwatz (2004). *Nonparametric and Semiparametric Models.* Berlin: Springer Verlag.

Hart, J.D. (1997). *Nonparametric Smoothing and Lack-of-Fit Tests.* New York: Springer Verlag.

Hastie, T. & A. Tibshirani (1993). Varying-coefficient models. *Journal of Royal Statistical Society, Series B* 55, 757–796.

Hayfield, T. & J.S. Racine (2007). *np: Nonparametric kernel smoothing methods for mixed datatypes.* R package version 0.13–1.

Henderson, D., R.J. Carroll & Q. Li (2006). Nonparametric estimation and testing of fixed effects panel data models. Manuscript, Texas A&M University.

Hodges, J.L. & E.L. Lehmann (1956). The efficiency of some nonparametric competitors of the t-test. *Annals of Mathematical Statistics* 27, 324–335.

Hoerl, A. & R. Kennard (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* 12, 55–67.

Horowitz, J.L. (1998). *Semiparametric Methods in Econometrics.* New York: Springer-Verlag.

Horowitz, J.L. & W. Härdle (1994). Testing a parametric model against a semiparametric alternative. *Econometric Theory* 10, 821–848.

Horowitz, J.L. & V.G. Spokoiny (2001). An adaptive, rate-optimal test of a parametric mean-regression model against a nonparametric alternative. *Econometrica* 69, 599–631.

Hristache, M., A. Juditsky & V. Spokoiny (2001). Direct estimation of the index coefficient in a single-index model. *Annals of Statistics* 29, 595–623.

Hsiao, C., Q. Li & J.S. Racine (2007). A consistent model specification test with mixed categorical and continuous data. *Journal of Econometrics* 140, 802–826.

Huber, P.J. (1964). Robust estimation of a location parameter. *Annals of Statistics* 35, 73–101.

Hurvich, C.M., J.S. Simonoff & C.L. Tsai (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of Royal Statistical Society, Series B* 60, 271–293.

Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics* 58, 71–120.

Johnston, J. & J. DiNardo (1997). *Econometric Methods*, 4th edition. McGraw-Hill.

Klein, R.W. & R.H. Spady (1993). An efficient semiparametric estimator for binary response models. *Econometrica* 61, 387–421.

Lavergne, P. & Q. Vuong (1996). Nonparametric selection of regressors: The nonnested case. *Econometrica* 64, 207–219.

Leung, D. (2005). Cross-validation in nonparametric regression with outliers. *Annals of Statistics* 33, 2291–2310.

Li, Q. & J.S. Racine (2007a). *Nonparametric Econometrics: Theory and Practice.* Princeton University Press.

Li, Q. & J.S. Racine (2007b). Smooth varying-coefficient nonparametric models for qualitative and quantitative data. Manuscript, Texas A&M University.

Li, Q. & J.S. Racine (2003). Nonparametric estimation of distributions with categorical and continuous data. *Journal of Multivariate Analysis* 86, 266–292.

Li, Q. & J.S. Racine (2004). Cross-validated local linear nonparametric regression. *Statistica Sinica* 14, 485–512.

Li, Q. & J.S. Racine (в печати). Nonparametric estimation of conditional CDF and quantile functions with mixed categorical and continuous data. *Journal of Business & Economic Statistics.*

Loader, C.R. (1999). Bandwidth selection: Classical or plug-in? *Annals of Statistics* 27, 415–438.

Manski, C.F. (1988). Identification of binary response models. *Journal of American Statistical Association* 83, 729–738.

Maronna, A., R. Martin & V. Yohai (2006). *Robust Statistics: Theory and Methods*. Wiley.

Nadaraya, E.A. (1965). On nonparametric estimates of density functions and regression curves. *Theory of Applied Probability* 10, 186–190.

Pagan, A. & A. Ullah (1999). *Nonparametric Econometrics*. New York: Cambridge University Press.

Parzen, E. (1962). On estimation of a probability density function and mode. *Annals of Mathematical Statistics* 33, 1065–1076.

Prakasa Rao, B.L.S. (1983). *Nonparametric Functional Estimation*. Orlando: Academic Press.

R Development Core Team (2007). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna. URL: `http://www.R-project.org`.

Racine, J. (2002). Parallel distributed kernel estimation. *Computational Statistics and Data Analysis* 40, 293–302.

Racine, J.S. (1997). Consistent significance testing for nonparametric regression. *Journal of Business & Economic Statistics* 15, 369–379.

Racine, J.S., J.D. Hart & Q. Li (2006). Testing the significance of categorical predictor variables in nonparametric regression models. *Econometric Reviews* 25, 523–544.

Racine, J.S. & Q. Li (2004). Nonparametric estimation of regression functions with both categorical and continuous data. *Journal of Econometrics* 119, 99–130.

Racine, J.S. & L. Liu (2007). A partially linear kernel estimator for categorical data. Manuscript, McMaster University.

Robinson, P.M. (1988). Root-n consistent semiparametric regression. *Econometrica* 56, 931–954.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics* 27, 832–837.

Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics* 9, 65–78.

Ruppert, D., R.J. Carroll & M.P. Wand (2003). *Semiparametric Regression Modeling*. New York: Cambridge University Press.

Ruppert, D., S.J. Sheather & M.P. Wand (1995). An effective bandwidth selector for local least squares regression. *Journal of American Statistical Association* 90, 1257–1270.

S original by Matt Wand. R port by Brian Ripley. (2007). KernSmooth: Functions for kernel smoothing for Wand & Jones (1995). R package version 2.22–21. URL: `http://www.maths.unsw.edu.au/wand`

Scott, D.W. (1985). Averaged shifted histograms: Effective nonparametric density estimators in several dimensions. *Annals of Statistics* 13, 1024–1040.

Scott, D.W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization*. New York: Wiley.

Scott, D.W. & S.J. Sheather (1985). Kernel density estimation with binned data. *Communication in Statistics: Theory and Methods* 14, 1353–1359.

Seifert, B. & T. Gasser (2000). Data adaptive ridging in local polynomial regression. *Journal of Computational and Graphical Statistics* 9, 338–360.

Sheather, S. & M. Jones (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of Royal Statistical Society, Series B* 53, 683–690.

Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*. New York: Chapman and Hall.

Simonoff, J.S. (1996). *Smoothing Methods in Statistics*. New York: Springer Verlag.

Stone, C.J. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). *Journal of Royal Statistical Society* 36, 111–147.

Stone, C.J. (1977). Consistent nonparametric regression. *Annals of Statistics* 5, 595–645.

Venables, W.N. & B.D. Ripley (2002). *Modern Applied Statistics with S*, 4th edition. New York: Springer Verlag.

Wand, M.P. & M.C. Jones (1995). *Kernel Smoothing*. London: Chapman and Hall.

Wang, F. & D. Scott (1994). The l_1 method for robust nonparametric regression. *Journal of American Statistical Association* 89, 65–76.

Wang, M.C. & J. van Ryzin (1981). A class of smooth estimators for discrete distributions. *Biometrika* 68, 301–309.

Wang, N. (2003). Marginal nonparametric kernel regression accounting for within-subject correlation. *Biometrika* 90, 43–52.

Wang, N., R.J. Carroll & X. Lin (2005). Efficient semiparametric marginal estimation for longitudinal/clustered data. *Journal of American Statistical Association* 100, 147–157.

Watson, G.S. (1964). Smooth regression analysis. *Sankhya* 26, 359–372.

Wooldridge, J.M. (2002). *Econometric Analysis of Cross Section and Panel Data*. Cambridge: MIT Press.

Wooldridge, J.M. (2003). *Introductory Econometrics*. Mason: Thompson South-Western.

Yatchew, A.J. (2003). *Semiparametric Regression for the Applied Econometrician*. New York: Cambridge University Press.

Nonparametric econometrics: a primer

Jeffrey S. Racine

McMaster University, Hamilton, Canada

This article is a primer for those who wish to familiarize themselves with nonparametric econometrics. Though the underlying theory for many of these methods can be daunting for some practitioners, this article will demonstrate how a range of nonparametric methods can in fact be deployed in a fairly straightforward manner. Rather than aiming for encyclopedic coverage of the field, we shall restrict attention to a set of touchstone topics while making liberal use of examples for illustrative purposes. We will emphasize settings in which the user may wish to model a dataset comprised of continuous, discrete, or categorical data (nominal or ordinal), or any combination thereof. We shall also consider recent developments in which some of the variables involved may in fact be irrelevant, which alters the behavior of the estimators and optimal bandwidths in a manner that deviates substantially from conventional approaches.