Национальный Исследовательский Университет

Высшая Школа экономики

Факультет экономики

Математические методы анализа экономики
(Специализация или магистерская программа)

Математической экономики и эконометрики
(кафедра)

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

Реализация и сравнение робастных эконометрических методов в пакете "Эр"

Выполнил
Студент группы №71ММАЭ
Пузанов А.С.

Научный руководитель:
Демешев Б.Б
Демидова О.А.

.Москва 2013

Оглавление

Описание работы

Работа посвящена реализации, применению и сравнению нового метода оценки параметров регрессии – параметрической модальной регрессии с другими методами оценке. Мы приводим расширенные по сравнению с пионерской статьей Weixin Yao and Longhai Li (2011) результаты симуляций методом Монте-Карло для написанной нами функции в пакете R, и указываем на наличие преимущественных правил выбора ширины окна ядерной оценки плотности (вопроса не рассмотренного в статье [16])– «Botev» и «Seather Johnes». Показываем влияние параметра предназначенного для уменьшения ширины окна, и указываем рекомендуемые границы его использования. Сравниваются результаты OLS, медианной и модальной регрессии на данных по загрязнению воздуха и ценах на дома в Бостоне и зависимость цен на компьютеры от оборудования к ним по данным PC Magazine. В обоих случаях мы не находим значительных отличий в качестве полученной модели, но модальная регрессия демонстрирует отсутствие гетероскедастичности по тесту Tukey (Tukey test of additivity), и несколько отличающиеся оценки коэффициентов. Также в каждом случае модальная регрессия показывает небольшое уменьшение длины доверительного интервала при фиксированной доле покрытия (coverage rate). В качестве базы для сравнения модальной регрессии с другими робастными методами была выбрана медианная регрессия, как классический пример использования характеристики условного распределения отличного от среднего. Мы также привели несколько методических упражнений для наглядного объяснения разницы в результатах оценки различных характеристик условного распределения. В конце работы также

приводится общая структура написанной функции и ключевые принципы, лежащие в её основе.

Глава 1

Введение

Широко известно, что традиционный метод наименьших квадратов весьма чувствителен к выбросам и структурным отклонениям. Для решения данной проблемы существует большое разнообразие альтернативных, устойчивых методов оценки параметров – робастных оценок. Квантильные регрессии, метод взвешенных наименьших квадратов, группировка наблюдений, непараметрические методы и другие. все эти методы весьма популярны. Почти все они в той или иной форме имплементированы в эконометрический пакет R. Одним из недавних направлений развития робастных оценок стала оценка условного распределения моды. Методология и преимущества оценки в полной мере раскрыты в статье [16] (Weixin Yao and Longhai Li “A New Regression Model: Modal Linear Regression”). Авторы [16] впервые показывают параметрический метод оценки линейной модальной регрессии, и доказывают необходимые свойства оценки. Данная статья вышла в 2011 году, однако в течении прошедшего времени нами не был найден ни один пример имплементации данного метода оценки в пакете R. Статья [16] является основой для данной работы и написанной под R функции оценки модальной регрессии. Авторы статьи [16] также дали возможность изучить код , использованный ими для получения оценок в пакете MATLAB. Этот код лёг в основу устройства функции *param.mod()* созданной нами в пакете R. Однако гибкость и функциональность созданной нами функции значительно превосходит исходную функцию в MATLAB. Во многом это является следствием огромной инфраструктуры созданной вокруг R. Скрипт для загрузки функций доступен по

запросу на email (artem.puzanov2@gmail.com), и скоро появится на публичных ресурсах. После соответствующей документации также будет добавлена библиотека в репозиторий R.

Существует несколько работ, в той или иной форме посвященных оценке моды для данных низкой размерности. В частности авторы [16] упоминают Muller and Sawitzki (1991); Scott (1992); Friedman and Fisher (1999); Chaudhuri and Marron (1999); Fisher and Marron (2001); Davies and Kovac (2004);Hall, Minnotte, and Zhang (2004); Ray and Lindsay (2005); Yao and Lindsay (2009).

Мода популяции - достаточно часто используемая характеристика, наряду с медианой и средним. Однако до авторов[16] никто не делал попытку оценить моду условного распределения при предположении о некоторой параметрической структуре данных.

Пусть у нас есть выборка $\{(x_i, y_i), i = 1, \dots, n\}$, где x_i вектор размерности p. $f(y|x)$ – условная функция плотности Y при известных x. Целью большинства регрессионных методов оценки является оценка какой-либо из характеристик подобного распределения. Классические модель регрессии как правило оценивают условное среднее, но часто рассматриваются и другие характеристики, например медиана.

В данной работе мы рассматриваем применение метода оценки **моды условного распределения** $f(y|x)$ предложенном в [16].

В данном случае рассматривается оценка с предпосылкой о линейной зависимости моды $f(y|x)$ от x. Иными словами рассматривает не условное среднее, а наиболее «частые» значения. Приведем ключевые преимущества модальной линейной регрессии, перечисленные в статье [16].

1) В случае если функция плотности ошибок сильно смещена, оценка модальной регрессия дает более осмысленную (легко интерпретируемую) локальную оценку.

2) Так как модальная регрессия фокусируется на участке функции с наибольшей условной плотностью, предполагается, что предсказательный интервал для заданной доверительной вероятности у неё меньше.

3) Модальную регрессию можно использовать, если часть точек «плохие» (ошибочны или имеют несхожую с остальным точками структуру) без значительного смещения оценок

4) В силу робастности моды, модальная регрессия устойчива к выбросам и «тяжелым хвостам»

5) У модальной регрессии есть ряд преимуществ перед стандартной робастной регрессией: традиционные робастные регрессии (например медианная регрессия и) требуют симметричной плотности ошибки, и в случае невыполнения данной предпосылки её интерпретация довольно затруднительна. Модальная регрессия не требует подобной предпосылки, и легко интерпретируема в случае смещённой плотности ошибок.

Далее статья организована следующим образом: В первой части статьи мы более приводим рассматриваем исходную структуру EM(Expectation maximization) алгоритма, и проверяем чувствительность EM алгоритма к правилу выбору ширины окна на симуляциях методом Монте-Карло. Также рассматривается сходимость коэффициентов на увеличении размеров выборки. Вторая часть статьи посвящена примерам применения модальной регрессии к различным двум наборам данных и сравнению результатов применения модальной регрессии и иных методов оценки. Третья часть предлагает примеры упражнений для студентов, имеющих

целью показать как могут отличаться оценки различными методами при разной структуре исходных данных. Четвертая часть статьи посвящена описанию структуры моделей в R, и применению данной структуры применительно к внедренной функции.

Фундаметнальная идея оценок характеристик условной плотности

Задача модели – ответ на вопрос о зависимости некоторой характеристики плотности вероятности зависимой величины от независимых величин. Задача модели оценки условного среднего – ответ на вопрос «как распределено среднее зависимой величины в зависимости от независимых величин». Значительная часть моделей в эконометрике существует для уточнения ответа именно на данный вопрос. В частности значительная часть робастных методов сконцентрирована вокруг данной концепции. Несколько ближе по идее к модальной регрессии стоят квантильные регрессии. Их задача также оценка некоторой отличной от среднего характеристик распределения. Например медианная регрессия (50% квантитль) является очень популярным робастным методом оценки. Важно заметить, что логика построения модели, естественно, не сводится к исключительно поиску среднего – исследователь заинтересован в получении наибольшего количества информации о поведении зависимых и независимых величин. Однако часто исследователи фокусируются на условном среднем и вообще не рассматривают иные характеристики распределения. Это не является значительной проблемой , однако столь узкий фокус может повредить итоговому качеству получаемой модели, если исследователь просто не рассматривает того что возможно характеристики медианы, моды и иные могут нести не меньше полезной информации, чем условное среднее.

Естественно, при определенных предпосылках, ответы на вопросы о различных характеристиках имеющихся данных имеют одинаковый ответ – классические предпосылки МНК делают регрессию условного

среднего и условной медианы равноценными. Из этого ,однако, не следует, что получая оценку медианы, мы автоматом получаем оценку и среднего. Используя разные методы, можно узнать не только более точную информацию при ответе на поставленный вопрос, но и новую информацию.

Рассмотрим пример процесса приведенный в [16]:

Пусть $f(y|x)$ связаны следующим соотношением: $Y = m(x) + \sigma(x)\varepsilon$, где ε имеет плотность $h(.)$. Пусть $h(.)$ имеет смещенную плотность со средним = 0 и модой = 1.

Пример №1: Если $m(x) = X^T\beta$ а $\sigma(x) = X^T\alpha$, то

$$E(Y|x) = X^T\beta$$

$$Mode(Y|x) = X^T(\alpha + \beta).$$

В этом случае зависимость между X и Y линейна как с точки модальной, так и с точки зрения линейной регрессии, но зависимости разные. Если бы мы считали использовали только оценку линейного среднего, существующая зависимость $\sigma(x) = X^T\alpha$ была бы принята за гетероскедастичность в модели.

Пример №2: Пусть $f(y|x)$ связаны следующим соотношением: $Y = m(x) + \sigma(x)\varepsilon$где $m(x) = 0$ а $\sigma(x) = X^T\alpha$.В этом случае $E(Y|x) = 0$, но $Mode(Y|x) = X^T\alpha$. В случае если мы используем только оценки условного среднего, в данном случае будет сделан вывод, что зависимость между Y и X отсутствует. Оценка моды укажет на зависимость между Y и X.

Глава 2

Оценка параметров с помощью Modal Expectation Maximization

Суть метода оценки модальной регрессии состоит в итеративной оценке функции плотности остатков. Каждая итерация придает вес наблюдению равный оценке значения функции плотности остатка на данном наблюдении в предыдущей итерации. Таким образом чем ближе наблюдение к моде функции плотности остатков, тем больший вес ему придается.

Задачей такого алгоритма является максимизация суммы оценок плотности остатков. Работа алгоритма прекращается, как только разница между суммой значений функции плотности на текущей и предыдущей итерации становится достаточно мала. Авторы [16] предлагают оценивать β максимизируя следующую целевую функцию:

$$Q_h(\beta) = \frac{1}{n}\sum_{i=1}^{n} \phi_h(y_i - x_i^T\beta) \qquad (1)$$

где $\phi_h = h^{-1}\phi(th^{-1})$ и $\phi(t)$ - ядерная функция плотности. Итоговая оценка $\hat{\beta}$, максимизирующая функцию (1)- будет MODLR оценкой (в соответствии с [16]).

Важный момент: алгоритм использованный авторами для симуляций методом Монте-Карло использует не оцененную на данной итерации функцию плотности остатков, а оценку значений нормальной функции плотности в остатках(с нулевым математическим ожиданием и дисперсией, оценённой ранее по методами описанному авторами [16]). Для удобства будем обозначать оценку полученную с помощью взвешивания по стандартному нормальному распределению - MODLR , а оценки полученные с помощью взвешивания по ядерной оценки плотности – MODLRkern.

MEM (Modal expectation maximization) алгоритм.

Так как у оценки (1) нет явного решения авторами [16] было предложено использовать алгоритм MEM, разработанный в статье [11](Li, J., Ray, S., and Lindsay, B. G. (2007)). В алгоритме MEM, как и в большинстве стандартных EM алгоритмов есть два шага:

Шаг E:

$$\pi\left(j\middle|\beta^{(k)}\right) = \frac{\phi_h\left(y_i - x_i^T\beta\right)}{\sum_{i=1}^{n}\phi_h(y_i - x_i^T\beta)} \qquad (2)$$

Шаг M:

$$\beta^{k+1} = argmax \sum_{i=1}^{n}\left\{\pi(j|\beta^k)\log\phi_h\left(y_j - x_j^T\beta\right)\right\} X^T W_k y$$

$$= (X^T W_k X)^{-1} X^T W_k Y$$

где $X = (x_1, \dots, x_n)^T$,а $W_K -$ диагональная матрица $n\times n$, с диагональными элементами $w_{jj} = \pi(j|\beta^k)$, $y = (y_1, \dots, y_n)^T$. Далее приведем теоремы о характеристиках MEM алгоритма.

Теоремы о свойствах MEM алгоритма.

Доказательства данных теорем (Теоремы №1, №2 и №3) и их предпосылки в полном объёме приведены в статье[16].

Теорема №1:

Каждая итерация MEM алгоритма монотонно не уменьшает функцию (1). Как отмечается авторами [16] данный алгоритм не обязательн сходится к глобальному максимуму, и разумнее использовать несколько стартовых точек и выбрать результат с максимальном значением (1). Ключевым отличием MEM алгоритма от IRWLS является использование изменяющегося вектора весов.

Теорема №2:

При $h \to 0$ и $nh^5 \to \infty$, существует состоятельная оценка (1), такая, что:

$$\left|\left|\hat{\beta} - \beta_0\right|\right| = O_P\left\{h^2 + (nh^3)^{-\frac{1}{2}}\right\}$$

где β_0 – истинное значение коэффициента из (1).

Теорема №3:

При предпослыках аналогичных предпосылкам теоремы №2 для $\hat{\beta}$ асимптотически верно следующее выражение:

$$\sqrt{nh^3}\left[\hat{\beta} - \beta_0 - \frac{h^2}{2}J^{-1}K\{1 + o_p(1)\}\right] \xrightarrow{D} N\{0, v_2 J^{-1}LJ^{-1}\}$$

где $v_2 = \int t^2\phi^2(t)dt$

$J = E\{g''(0|x_i)x_i x_i^T; K = E\{g'''^{(0|x_i)}x_i\}$; $L = E\{g(0|x_i)x_i^T\}$

Зная асимптотическое смещение и асимптотическую дисперсию $\hat{\beta}$, мы можем выбрать теоретически оптимальную ширину окна h для оценки β минимизируя WMSE (weighted mean squared errors)

$$E\left\{\left(\hat{\beta} - \beta_0\right)^T W\left(\hat{\beta} - \beta_0\right)\right\} \approx \frac{K^T J^{-1} W J^{-1} K h^4}{4} + (nh^3)^{-1} tr(J^{-1} L J^{-1} W)$$

где W – функция весов отражающая важность коэффициентов, tr(A) – след матрицы A. Если принять $W = J^{-1}LJ^{-1}$ то оптимальный h должен оцениваться как

$$h_{opt} = \left[\frac{3v_2(p+1)}{K^T L^{-1} K}\right]^{-\frac{1}{7}} \qquad (3)$$

Доказательства и более подробное описание метода оценки можно найти в [16].

Глава 3

Тестирование модели на симуляциях методом Монте-Карло.

1) Пример генерированных данных

По примеру авторов [16] возьмем следующий пример генерирующего процесса $\{(x_i, y_i), i = 1, \dots, n\}$:

$$Y = 1 + 3X + (1 + 2X)\varepsilon \qquad (4)$$

где $X \sim UNIF(0,1);\ \varepsilon \sim 0.5N(-1, 2.5^2) + 0.5N(1, 0.5^2)$

Характеристики распределения ε:

$$E(\varepsilon) = 0;\ Mode(\varepsilon) = 1;\ Median(\varepsilon)$$

Рассмотрим плотность получившегося распределения ε:

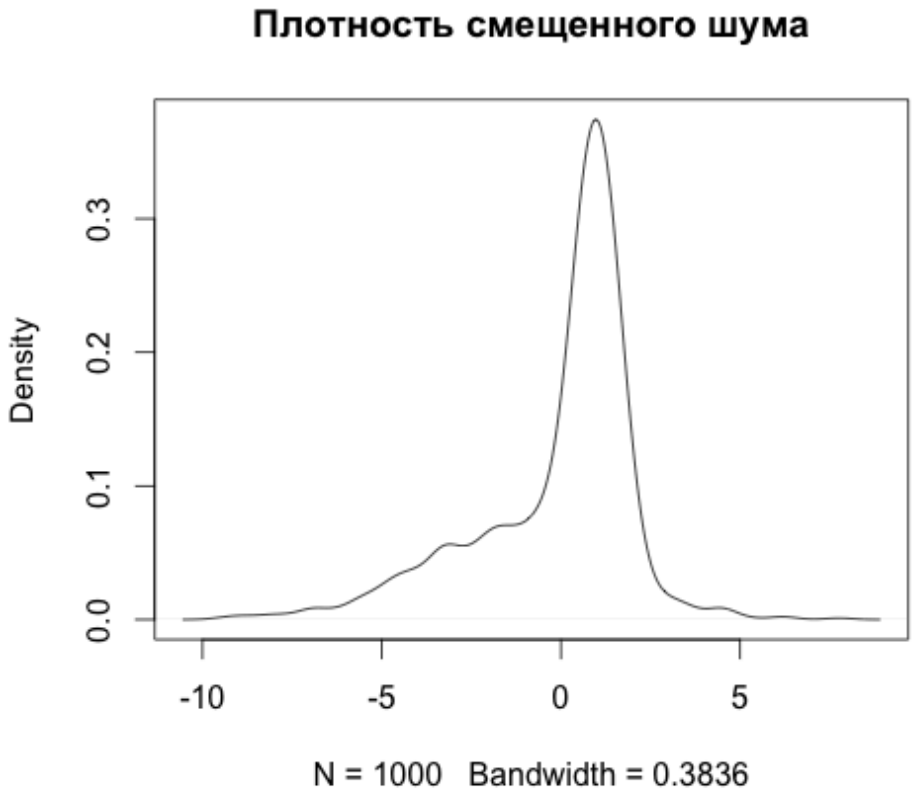


График 1 (По оси Y отложены значения функции плотности остатков)

Как мы видим, плотность распределения ε имеет заметное смещение. Теперь сравним результаты оценок различными методами на примере сгенерированных 400 наблюдений:

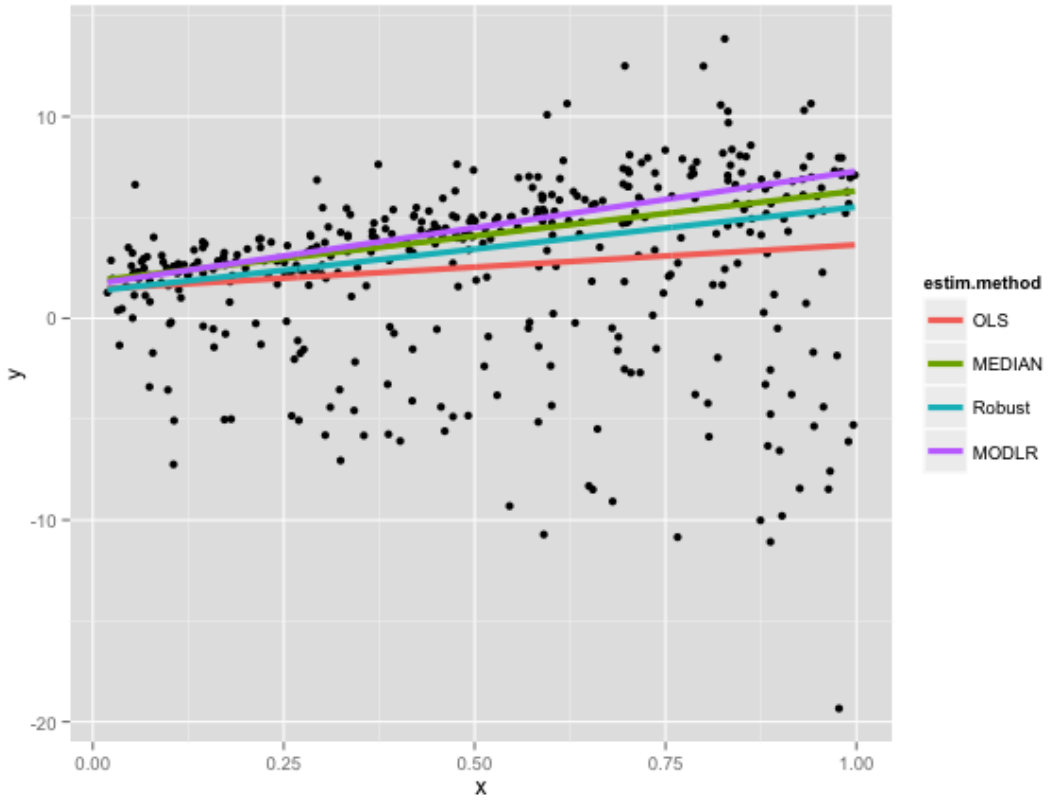


График 2 (по сгенерированным данным)

Как мы видим, MODLR оценка проходит выше остальных значений, адекватно показывая наиболее «частые» значения Y.

Отметим, что полученные в результате генерирующего процесса (4) уравнения регрессии должны сильно отличаться для среднего, моды и медианы.

$$E(Y|X) = 1 + 3X \qquad (5)$$

$$Median(Y|X) = 1.67 + 4.34X \qquad (6)$$

$$Mode(Y|X) = 2 + 5X \qquad (7)$$

Теперь рассмотрим средние оценки соответствующих параметров по результатам монте-карло для различных значениях n.

Симуляции ментодом Монте-Карло ;Таблица 1 (по сгенерированным данным)

Тип модели	Истинное значение оцениваемых параметров	n = 200		n = 400		n =1000	
		$\hat{\theta}$	$\hat{\sigma}$	$\hat{\theta}$	$\hat{\sigma}$	$\hat{\theta}$	$\hat{\sigma}$
MODLR (scale = 1)	$\alpha = 2$	1,62	0,32	1,72	0,23	1,80	0,19
	$\beta = 5$	5,19	0,56	5,16	0,38	5,08	0,25
MODLR (scale =2)	$\alpha = 2$	1,86	0,28	1,93	0,18	1,95	0,13
	$\beta = 5$	5,06	0,58	4,99	0,49	4,97	0,36
OLS	$\alpha = 1$	0,98	0,47	1,02	0,34	1,00	0,21
	$\beta = 3$	3,07	1,05	2,98	0,77	2,99	0,50
MEDIAN	$\alpha = 1.67$	1,64	0,42	1,63	0,29	1,67	0,19
	$\beta = 4.34$	4,31	0,97	4,36	0,67	4,32	0,46

$\hat{\theta}$ – среднее значение коэффициента по всем симуляциям.

$\hat{\sigma}$ – стандартная ошибка значения коэффициента по всем симуляциям.

MODLR (scale =1) – оценка модальной регрессии при стандартном выборе $h = h_{opt}$., где h_{opt}- выбор согласно уравнению (3).

MODLR (scale =2) – оценка модальной регрессии при выборе $h = h_{opt}/2$

Оценка MODLRkern не приводится, в силу необнаруженной эмпирической сходимости. Тем не менее мы продолжим её использовать, в силу того, что в терминах доли покрытия она в некоторых случаях превосходит MODLR(scale =1) и MODLR(scale =2).[1]

[1] Причины подобного рода несостоятельности не выяснены. Авторы[16] указали только результаты оценки с помощью весов от плотности стандартного нормального распределения. Ведущим предположением на данный момент является недооценка

На основе таблицы 1 можно сделать 3 вывода:

1) Оценки переменных MODLR стремятся к истинным оценкам на приемлемо малых объёмах выборки

2) Оценки переменных MODLR несколько смещены для малых объёмов выборки

3) Стандартные ошибки оценок MODLR заметно ниже, чем стандартные ошибки OLS и MEDIAN на небольших выборках. Дальнейшее схождение оценки MODLR(scale = 1) для больших объёмов выборки находится в приложении (Таблица 10) *Важно отметить, что полученные результаты не эквивалентны результатам авторов в [16], и сходятся к истинным значениям на большей выборке*[2]*. Результаты, аналогичные результатам полученным в статье [16], удалось получить с помощью уменьшения ширины окна, полученного в (3). За уменьшение параметра ширины окна отвечает параметр scale –в случае MODLR (scale =2) исходная ширина окна была уменьшена в два раза. Оценки с MODLR (scale =2) можно считать почти эквивалентным оценкам авторов [16]. Так как математического обоснования применения данного способа получить более состоятельные оценки нет, мы продолжим использовать оба метода, и выбирать наилучший на основе косвенных признаков Результаты применения различных значений параметра scale для n =* 400 *приводятся в приложении (Таблица 11).*

В целом ,можно заключить, что значение параметра scale > 2 не имеет смысла, так как это повышает стандартную ошибку оценки коэффициентов и при scale $\geq$ 4 *происходит сильное смещение оценок коэффициентов.*

ядерными функциями плотности весов остатков. При использовании квадратов весов оценки были значительно ближе к истинным.

[2] Код авторов (1) написан в программе MATLAB. Функция использованная для оценки модальной регрессии в данной работе была написана автором в пакете R. В данный момент причина расхождений оценок неизвестна.

Оптимальная ширина окна для ядерной оценки плотности, полученной в (3) все же требует использования некоторого изначального правила выбора ширины окна для оценки плотности остатков. Обозначим за h_{basic} базовое окно, используемое для оценки плотности остатков для регрессии, отдающей стартовые оценки $\hat{\beta}$. Авторы [16] в реализации MEM алгоритма использовали метод оценки разработанный Zdravko Botev в его статьях [1-3]. В функции написанной к данной работе мы также использовали код выложенный на сайте (http://www-etud.iro.umontreal.ca/~botev/). Как показывает чувствительность состоятельности оценок MODLR к параметру scale, правило выбора ширины окна - важно. На основе результатов из таблицы 1, была выдвинута гипотеза о важной роли правила выбора h_{basic} Ниже приведена таблица результатов симуляции методом Монте-Карло для оценки модальной регрессии MODLR (scale = 1) при различных правилах выбора ширины окна (Таблица №2).

Сравнение правил выбора ширины окна;Таблица 2 (по сгенерированным данным)

Правила выбора ширины окна	Истинные значения параметров	n = 200		n = 400		n =1000	
		$\hat{\theta}$	$\hat{\sigma}$	$\hat{\theta}$	$\hat{\theta}$	$\hat{\sigma}$	$\hat{\theta}$
Botev	$\alpha = 2$ $\beta = 5$	1,62	0,32	1,72	0,23	1,80	0,19
		5,19	0,56	5,16	0,38	5,10	0,25
Seather - Johnes		1,63	0,31	1,72	0,24	1,81	0,19
		5,19	0,56	5,16	0,38	5,10	0,26
Silverman rule of thumb		1,33	0,35	1,50	0,26	1,68	0,14
		5,32	0,62	5,31	0,38	5,22	0,23

Как мы видим правило выбора “Botev” и метода “Seather Johnes” дают наилучшие оценки.. В функции написанной авторами статьи [16] по умолчанию используется правило “Botev”

Правило Silverman’s работает хуже чем “Botev” и “Seather Johnes” . Далее во всей работе используется правило Botev для выбора ширины окна в регрессии, при генерации стартовых точек MEM алгоритма.

В целом, функция использованная в данной работе несколько смещенно оценивает коэффициенты, переоценивая угловой коэффициент и недооценивая константу на небольших выборках. В тоже время функция состоятельно оценивает коэффициенты при достаточно больших выборках.

Как упоминалось выше, стандартная ошибка для оценки коэффициентов модальной регрессии в симуляции методом Монте-Карло меньше, чем у остальных методов.

Объяснение столь точной оценки простое: значительная часть наблюдений для модальной регрессии находится в достаточно узком доверительном интервале. Чтобы показать как именно отличается число наблюдений попадающих в доверительный интервал , авторы [16] предлагают воспользоваться следующей стратегией:

Для каждой из 1000 симуляций методом Монте-Карло (для выбранного числа наблюдений) создается вектор зависимой переменной Xиз 1000 значений равномерно лежащих между 0.1 и 0.9 . Затем для известной нами величины ошибки ($\sigma(\varepsilon) = 2$) мы смотрим какая доля предсказанных значений попадает в $\theta\sigma(\varepsilon)$ от фактических значений, где $\theta = \{ 0.1 ; 0.2 ; 0.5\}$. Долю наблюдений, попадающих в заданный интервал будем называть *долей покрытия.*

Доля покрытия, Монте-Карло; Таблица 3 (по сгенерированным данным)

Величина интервала	Модель	n = 50		n = 100		n =200	
		$\hat{\theta}$	$\hat{\sigma}$	$\hat{\theta}$	$\hat{\theta}$	$\hat{\sigma}$	$\hat{\theta}$
$0.1\sigma(\varepsilon)$	LSE	0,035	0,016	0,032	0,012	0,030	0,009
	$MEDIAN$	0,074	0,019	0,077	0,015	0,079	0,013
	MODLR (scale = 1)	0,074	0,021	0,084	0,015	0,088	0,014
$0.2\sigma(\varepsilon)$	LSE	0,071	0,030	0,065	0,023	0,061	0,016
	$MEDIAN$	0,147	0,034	0,152	0,026	0,155	0,021
	MODLR (scale = 1)	0,146	0,037	0,165	0,025	0,174	0,022
$0.5\sigma(\varepsilon)$	LSE	0,188	0,065	0,179	0,049	0,173	0,038
	$MEDIAN$	0,342	0,059	0,353	0,042	0,359	0,032
	MODLR (scale = 1)	0,341	0,067	0,376	0,039	0,389	0,032

Как мы видим по Таблице №3, у модальной оценки для фиксированного интервала вокруг предсказанного значения доля в среднем больше, чем у остальных методов оценки. При этом на 50 наблюдениях у медианной регрессии доля покрытия приблизительно равна доля покрытия модальной регрессией. Это можно объяснить смещенностью оценок параметров в модальной регрессии с помощью нашей функции при малых выборках[3]. Для $n > 100$ доля покрытия у модальной регрессии становится заметно выше, в силу более точной оценки параметров при увеличении выборки. Значительное число

[3] В статье [16] при 50 наблюдениях оценка MODLR дает наибольшую долю покрытия для всех размеров выборки.

наблюдений в узком интервале объясняет сниженную ошибку оценок параметров у MODLR на фоне других методов.

Глава 4

Применение модальной регрессии и сравнение с другими методами оценки

В данном разделе мы рассмотрим примеры применения модальной регрессии, и сравним результаты оценки методом MODLR с другими методами оценки, включая классическую робастную оценку с помощью медианы. Для начала попробуем обозначить, какие характеристики получившихся оценок будут сравниваться. Проблема заключается в том, что некоторая часть популярных методов сравнения (в частности R^2) не является корректной (корректность в данный момент не доказана). Будем исходить из ключевых особенностей модальной регрессии:

1) Возможность корректной оценки при некоторых типах гетероскедастичности

2) При одинаковой величине доверительного интервала – большая доля наблюдений попадающих в доверительный интервал

Как мы видели, существует как минимум 3 спецификации MODLR, которые мы можем рассматривать:

MODLR(scale = 1), MODLR(scale = 2) –методы ограниченные способом пере-взвешивания с помощью стандартного нормального распределения; и MODLRkern – метод использующий пере взвешивание на основе ядерной оценки плотности (1). Так как на текущем этапе мы не можем отдать предпочтение одному из них, мы будем сравнивать плотности остатков в нуле и выбирать ту модель, у которой наибольше значение плотности остатков в нуле. Теоретически это позволит нам выбирать наиболее «модальную» из этих моделей.

Для оценки возможной гетероскедастичности, вызванной связью ошибки и переменных, мы используем тест Tukey (Tukey additivity test).

Как мы упоминали выше, значительная часть процедур может быть некорректна для оценки параметров модальной регрессии. Однако существуют процедуры, напрямую изучающие результат работы модели. Классический пример – качество предсказания Value-at-risk. В данном случае мы используем параметр схожий с параметром, оптимизируемом в Value-at-risk – долю покрытия. Под долей покрытия подразумевается доля наблюдений, попадающих в некоторый интервал вокруг предсказанного значения зависимой переменной.

В случае выбора между двумя моделями преимуществом в контексте доли покрытия обладает та модель, которая для фиксированной доли покрытия требует меньший интервал вокруг предсказанного значения.

О примерах

В статье [16] использовался пример данных о лесных пожарах. В качестве зависимой переменной использовался показатель площади выгоревшей территории в результате пожара. В случае если пожара не происходило, значение зависимой переменной принималось за 0. Данный пример действительно хорошо подходил для показа фундаментальной разницы между различными моделями, так как зависимая переменная была сильно смещена, и стандартные методы оценки не подходили. Однако, несмотря на наглядность примера, он все же не был проработан. Вполне разумным развитием такого подхода было бы использовать трансформацию типа Tobit, однако подобный глубокий разбор модели не входил в задачи авторов статьи

[16]. Очевидным недостатком подобного подхода является неясность отличия модальной регрессии от остальных подходов на данных с известной методологией и высокой объясняющей силой стандартных моделей.

В данной статье мы рассмотрим два примера применения модальной регрессии:

1) Гедонистический подход к оценке спроса на чистый воздух с использованием цен на дома (Статья David Harrison, Daniel Rubelfield: Hedonic Housing prices and the demand for clean air [16]). В статье [16] используется подход к оценке стоимости загрязнения воздуха, основанный на разнице в стоимости домов в различных районах Бостона.

2) Гедонистический подход к анализу рынка персональных компьютеров в 1993-1995 годах. Основано на статье [17] T. Stengos and E. Zacharias : Intertemporal Pricing and Price Discrimination: A Semiparametric Hedonic Analysis of the Personal Computer Market.;

В этой статье рассматривается зависимость стоимости компьютеров от стоимости компонент к ним на протяжении 2-ух лет.

Используя методологию авторов статей [16] и [17], мы рассматриваем разницу между оценками модальной регрессии и другими методами оценки, а также сравниваем долю покрытия с помощью leave-one-out cross validation.

Оценка влияния чистоты воздуха на стоимость домов в Бостоне

Статья [16] посвящена анализу спроса на чистый воздух с помощью разницы в ценах на дома/квартиры в различных районах Бостона. Как отмечают авторы [16], в то время как эффекты загрязнения воздуха на здоровье хорошо документированы, оценка денежного эквивалента стоимости изменения концентрации загрязняющих веществ в воздухе задача достаточно сложная. В данном случае стоимость загрязнения воздуха оценивается на основе стоимости домов, в предположении, что при прочих равных за дома в зоне с загрязненным воздухом домохозяйства готовы платит меньше. Ключевые предпосылки авторов $U(x,h)$ - полезность домохозяйства при условии бюджетного ограничения $y = x + p(h) + T$, где

h – группа характеристик дома, таких как транспортная доступность, характеристики района и иные

$x-$*количество комозитных частных благ, с единичной ценой*

$y -$ доход домохозяйства

$p(h) -$ функция оценки домохяйством стоимости

$T -$ издержки на передвижение

Более полное описание предпосылок и последствий их выбора приведены в [16]

Авторы использовали 4-шаговую процедуру оценки:

1) Оценка $p(h)$функции

2) Оценка $W_a = \frac{\partial U / \partial(-a)}{\partial U / \partial(x)} = \frac{\partial p(h)}{\partial a} = p_a(h)$

3) Оценка обратного компенсированного спроса на основе оценки W_a

4) Оценка влияния политики контроля за вредными выбросами

В данной статье мы приводим только результаты *первого шага,* так как хотим сравнить только результаты базовых моделей.

Авторы [16] использовали данные о медианной стоимости домов, для каждого дома выделены 13 характеристик:

Зависимая величина Median Value (MV). - медианная стоимость домов, в которых проживают хозяева. По примеру авторов [16] мы используем логарифм данной величины. Загрязнение воздуха измеряется в концентрации NO_2(оксид азота) в воздухе– переменная NOX.

Также в модель включены 8 переменных, характеризующих качество района и 2 переменных, характеризующих транспортную доступность.

Так как существовала возможность нелинейной зависимости MV от переменной NOX, авторы[16] использовали перебор по сетке степени переменной NOX. Наилучшим результатом с точки зрения качества модели был признан степень равная двум. Полное описание переменных - в приложении (Таблица [12])

Оценки моделями коэффициентов по стоимости домов; Таблица 4 (По данным о пригородах Бостона, 1970 Census)

Variable	OLS	MEDIAN	MODLRkern
Intercept	4,56	3,83	3,63
RMSQ	0,01	0,01	0,01
log(DIS)	-0,19	-0,14	-0,12
log(RAD)	0,10	0,06	0,05
CRIM	-0,01	-0,01	-0,01
BLACK	0,0004	0,0006	0,0006
TAX	-0,00042	-0,00037	-0,00031
PTRATIO	-0,031	-0,027	-0,028

log(STAT)	-0,37	-0,26	-0,23
NOX^2	-0,63	-0,45	-0,43
CHAS	0,09	0,07	0,06
ZN	0,0001	0,0005	0,0001

В таблице №4 приведены значения оценок коэффициентов различных моделей. Как мы видим, медианная и MODLR оценки значения коэффициента при NOX заметно ниже в абсолютном выражении, чем МНК оценка, и также заметно ниже (в абсолютном выражении), чем оценка WLS метода в [17] (-0.58). Из этого следует, что если наша спецификация корректна, то население городов чаще всего менее чувствительно, чем позволяет предположить МНК и WLS оценка. Возможно, это связано с тем, что жителей городов можно условно поделить на чувствительную и нечувствительную группы. Чаще могут встречаться нечувствительные, но в среднем мы видим результат взаимодействия этих групп. Впрочем, альтернативных причин снижения оценки коэффициента NOX в результате применения модальной и медианной регрессии может быть довольно много. R^2для OLS достаточно высокий – около 0.8 Полные результаты оценок (со стандартными ошибками и значимостями) приведены в приложении, Таблицы 12-17.

Также приведем пример, характеризующий поведение остатков итоговой модели при различных значениях параметра scale и разницу в результатах MODLRkern , MODLR(scale = 1) и MODLR(scale = 2), Построим график остатков для различных методов оценок

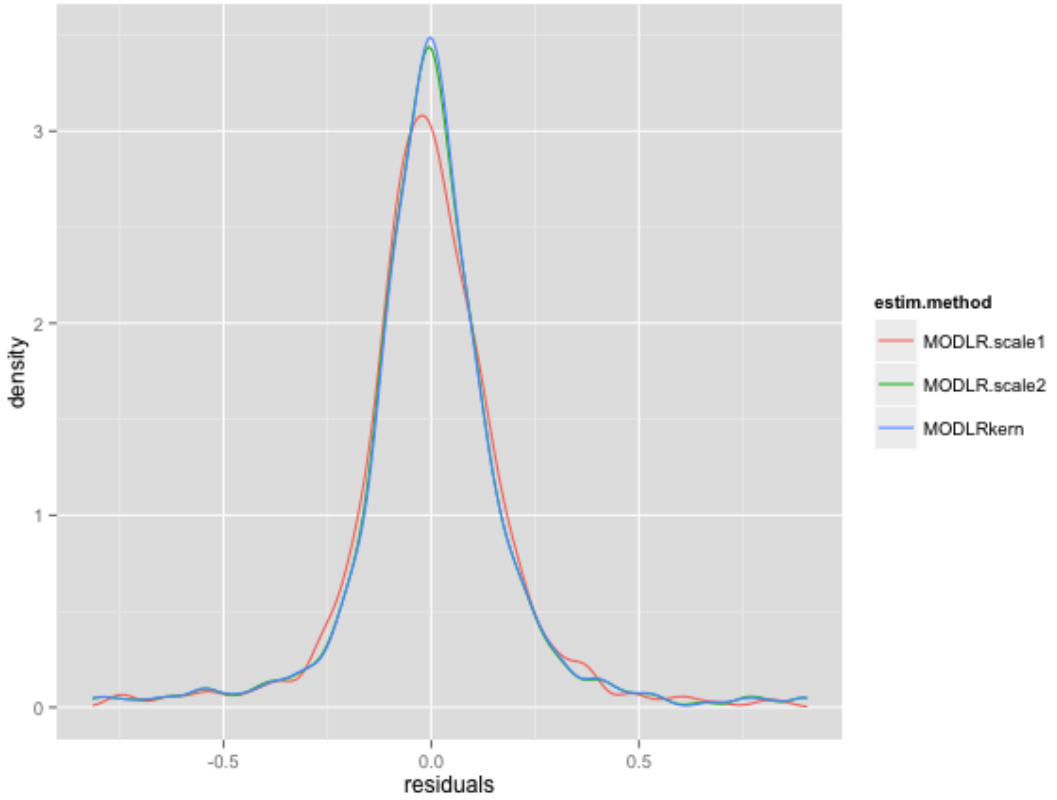


График 3 (График плотности вероятности остатков регрессий из таблицы 4)

Как мы видим на Графике №3, наибольшую плотность демонстрируют MODLR(scale =2) и MODLRkern. В качестве наилучшего выбора из вариантов модальной регрессии остановимся на MODLRkern.

Мы используем модель MODLRkern как наиболее точно соответствующую изначальному описанию MEM алгоритма и лучшую с точки зрения плотности остатков в нуле. (см. График 3). В силу схожести результатов оценок всех вариантов MODLR в отношении интересующего нас коэффициента при NOX, , мы приводим оценки коэффициентов только с помощью MODLRkern.

Теперь сравним характеристики модальной и линейной оценок в контексте возможной гетероскедастичности и возможности модальной регрессии её игнорировать при некоторых её формах: в

данной случае к переменным по очереди применяется тест Tukey на гетероскедастичность

Результаты теста Tukey; Таблица 5 (По данным о пригородах Бостона, 1970 Census)

Переменные	OLS p-values	MODLRkern p -values
rmsq	0.837	0.004
log(dis)	0.875	0.444
log(rad)	0.919	0.044
crim	0.765	0.000
black	0.963	0.020
tax	0.824	0.223
ptratio	0.925	0.083
log(lstat)	0.832	0.001
noxsq	0.980	0.763
chas	1.000	0.288
zn	0.829	0.216

Тест на аддитивность Tukey (Tukey's test of additivity, Таблица № 5) указывает на наличие гетероскедастичности при использовании некоторых переменных у OLS модели, но у модальной регрессии для всех регрессоров тест Tukey не отвергает гипотезу о гомоскедастичности. Рассмотрим характеристики доли покрытия. Для оценки предсказательного интервала использовался алгоритм предложенный авторами в [16], учитывающий возможность смещённости ошибки. Его суть сводится к перебору значений остатков, таким образом, чтобы при одинаковой оценке плотности для

остатков между ними находилась требуемая доля выборки (coverage rate).

Доля покрытия, Бостон ;Таблица 6(По данным о пригородах Бостона, 1970 Census)

Теоретическая доля покрытия	Модель	Интервал	Настоящая доля покрытия
10%	OLS	0,034	0,099
	MEDIAN	0,031	0,087
	MODLRkern	0,027	0,103
30%	OLS	0,105	0,338
	MEDIAN	0,097	0,271
	MODLRkern	0,086	0,310
50%	OLS	0,190	0,539
	MEDIAN	0,169	0,477
	MODLRkern	0,168	0,492

Ошибки для расчета смещенного предсказательного интервала были получены с помощью leave-one-out cross validation. Как мы видим в Таблице №6 , у модальной и медианной регрессии интервал в который попадает 10% выборки, несколько меньше чем у OLS и робастной OLS оценки, но это различие становится незначительным для больших % выборки. Таким образом, модальная регрессия не дает существенных преимуществ по сравнению с медианой и OLS моделями, но корректно решает задачу минимизации доверительного интервала для заданного процента выборки

Оценка зависимости стоимости компьютеров при гедонистической предпосылке о формировании цены.

Данный пример основан на статье « Intertemporal Pricing and Price Discrimination: A Semiparametric Hedonic Analysis of the Personal Computer Market» (T. Stengos and E. Zacharias)[17]. Статья посвящена анализу рынка компьютеров, и эволюции значений параметров во времени. Авторы используют полу-параметрических подход и сравнивают эволюцию значений параметров во времени. Однако в этой работе данный подход нами игнорируется, и из работы [17] мы в основном берем только методологию оценки базовой модели, без влияния времени на значения коэффициентов. Данная модель особенно интересна нам в силу большого объёма выборки, что позволяет предположить высокую точность оценки коэффициентов.

Данные

Мы используем данные о ценах на модель 486 PC показанных в рекламных объявлениях (данные получены из рекламы в журнале PC magazine США). Каждое наблюдение состоит из цены и группы характеристик, включая время объявления. Характеристики:

1) speed - Частота процессора в MHz (ожидаемый коэффициент положительный, в 486 PC частота прямо влияла на быстродействие)

2) hd - Объём жесткого диска в MB (ожидаемый коэффициент положительный)

3) ram - Объём оперативной памяти RAM((ожидаемый коэффициент положительный)

4) cd - Наличие CD

5) multi - Наличие устройств мультимедиа

6) screen - Размер экрана

7) premium -Премиальность (источник поставки IBM или COMPAQ).

8) pent - дамми переменная отвечающая за время запуска компьютеров Pentium

Так как в рекламных объявлениях часто указывалось несколько вариантов стоимости, каждый вариант считался за отдельное наблюдение. Более подробно об устройстве данных можно прочитать в статье [17].

Данный набор данных интересен также значительным объёмом – 6259 наблюдений.

Начнем со сравнения результатов всех стандартных моделей (полные версии MODLR регрессий – в приложении, таблицы 18-20.):

Оцени коэффициентов моделями, PC magazine; Таблица 7 (По данным PC magazine, 1993-1995)

Переменные	OLS	MEDIAN	MODLR (scale = 2)	MODLR (scale = 1)	MODLRkern
(Intercept)	4,2340	4,4442	4,1560	4,3930	4,4490
log(speed)	0,2082	0,1873	0,1494	0,1653	0,1868
log(hd)	0,1437	0,1414	0,2027	0,1535	0,1340
log(ram)	0,1811	0,1990	0,2116	0,2123	0,2021
log(screen)	0,7244	0,6678	0,7221	0,6927	0,6775
cdyes	0,0474	0,0389	0,0369	0,0335	0,0405
multiyes	0,0333	0,0238	0,0328	0,0263	0,0261
premiumyes	-0,2318	-0,2391	-0,3069	-0,2633	-0,2399
pent*	-0,0108	-0,0073	0,0120	0,0016	-0,0112
ads	0,0002	0,0002	0,0001	0,0002	0,0002
trend	-0,0222	-0,0219	-0,0244	-0,0222	-0,0214

*Коэффициент значим во всех моделях кроме MODLR(scale = 1).

По результатам в таблице №7 различные методы оценки модальной регрессии существенно различаются по оценкам влияния различных характеристик как между собой, так и от стандартной и медианной регрессии. Наибольшая схожесть со стандартной и медианной моделью наблюдается у MODLRkern модели. Наиболее интересные с точки зрения различий переменные это log(speed), log(ram) , multi и premium (заданы как дамми переменные, равные единице при наличии мультимедиа или принадлежности компании производителя к топ-сегменту). Как мы видим, логарифм скорости модель MODLR(scale =

2) оценивает значительно ниже, чем остальные регрессии. Однако та же модель оценивает влияние логарифма объёма жесткого диска в значительно большую величину. OLS и MODLR(scale = 2) оценивают эффект мультимедиа примерно на 50% выше, чем остальные модели. Эффект премиум-сегмента компании производителя также больше по абсолютной величине в MODLR(scale = 2). В силу значительных различий и отстутсвия согласнованных(одновременных) отличий MODLRот OLS и медианной регрессии, о причинах разницы в значениях коэффициентов в целом сказать сложно. Если оценить плотность остатков, то видно, что наибольшей плотностью остатков в нуле обладает модель MODLR(scale = 2) (График 4).

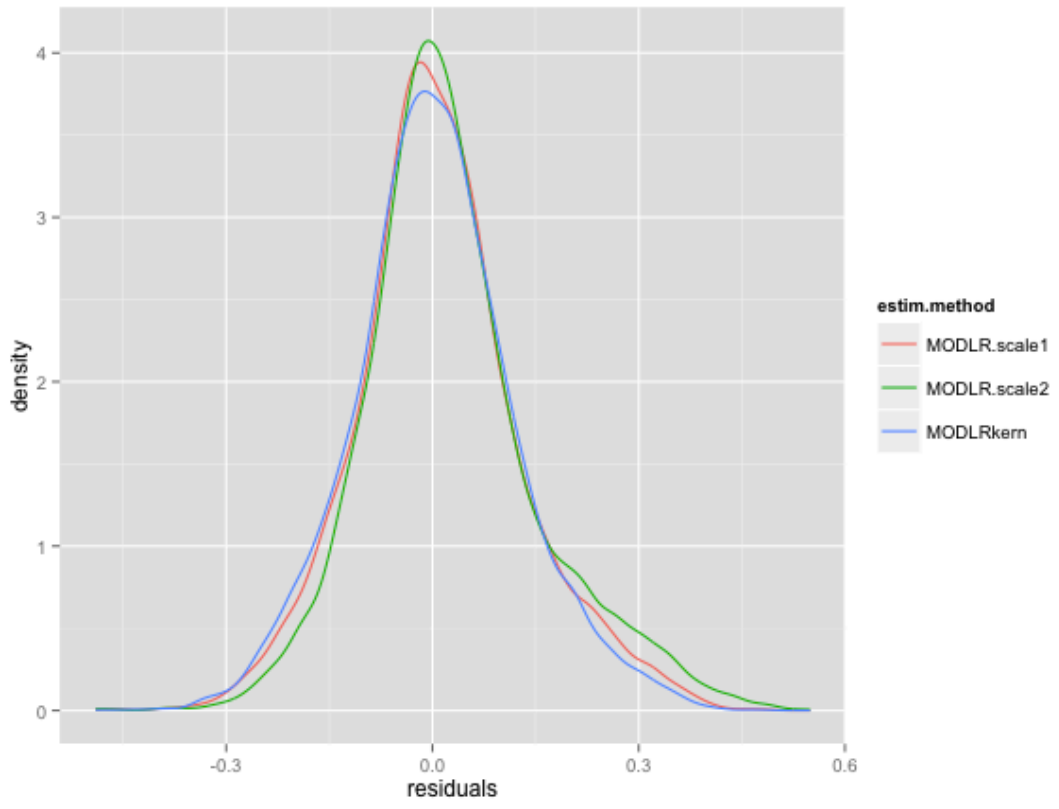


График 4 (График плотности остатков регрессий из таблицы 7)

Выбрав MODLR(scale = 2) в качестве основной модели, оценим тест Tukey на гетероскедастичность для всех численных переменных:

Тест Tukey, PC magazine; Таблица 8(По данным PC magazine, 1993-1995)

Переменные	OLS p-values	MODLR(scale = 2) p -values
log(speed)	0,000	0,866
log(hd)	0,000	0,849
log(ram)	0,000	0,492
log(screen)	0,000	0,629
pent	0,212	1,000
ads	0,535	0,993
trend	0,000	0,818

Как мы видим по таблице № 8 тест Tukey показывает наличие гетероскедастичности остатков при применении его к ряду переменных у OLS, однако оценка MODLR показывает отсутствие гетероскедастичности в формулировке теста Tukey. Далее приведем анализ доли покрытия у различных методов оценки.

Доля покрытия, PC magazine ; Таблица 9 (По данным PC magazine, 1993-1995)

Теоретическая доля покрытия	Модель	Интервал	Реальная доля покрытия
10%	OLS	0,026	0,095
	MEDIAN	0,025	0,098
	MODLR (scale =2)	0,026	0,095
30%	OLS	0,08	0,29
	MEDIAN	0,083	0,31
	MODLR (scale =2)	0,075	0,30
50%	OLS	0,150	0,50
	MEDIAN	0,142	0,49
	MODLR (scale =2)	0,137	0,50

Ошибки для расчета смещенного предсказательного интервала мы получили с помощью leave-one-out cross validation. У модальной и регрессии доверительный интервал для 10% выборки не отличается от медианной регрессии, но на 30 и 50 % доле покрытия доверительный интервал несколько меньше. Впрочем, различие незначительно. Скорее всего, столь малые различия связаны с большой выборкой и хорошей объясняющей силой OLS.

В результате применения оценки MODLR к большому объёму cross-section данных по продажам компьютеров, мы установили, что оценка MODLR в зависимости от различных методов выбора ширины окна и

различия в методах взвешивания остатков демонстрирует различные оценки. MODLR оценка убирает определенные эффекты гетероскедастичности, но при большой объясняющей силе модели не даёт значительных преимуществ при отсутствии смещения ошибок.

Глава 5

Методические упражнения

В качестве методических примеров к использованию робастных методов оценки, хорошо иметь задания показывающие, как разные методы оценки могут давать не только уточнения оценок, но и дополнительную информацию об имеющемся наборе данных. В качестве первого примера предлагается показать случай, когда OLS , медианная и модальная регрессия дают схожие оценки. После этого постепенно будем изменять случайный процесс таким образом, чтобы оценки различными методами начинали расходиться. Сначала используем процесс, отвечающий всем предпосылкам МНК:

Ошибка - $\varepsilon \sim N(0,1)$

Независимая переменная $X \sim UNIF(0,1)$

Генерирующий процесс: $Y=1+3X+\varepsilon$ *Type equation here.*

На данном этапе все три модели должны давать одинаковый результат.

На втором этапе заменим белый шум на смещенный белый шум:

$$\varepsilon \sim 0.5N(-1,2.5^2) + 0.5N(1,0.5^2)$$

$$Y = 1 + 3X + \varepsilon$$

Теперь оценки должны различаться

$$E(Y|X) = 1 + 3X$$

$$Median(Y|X) = 1.67 + 3X$$

$$Mode(Y|X) = 2 + 3X$$

На данном этапе также можно заменить часть ε на произвольные большие значения для генерации выбросов.

На третьем этапе используем аналогичный белый шум,

$$\varepsilon \sim 0.5N(-1,2.5^2) + 0.5N(1,0.5^2)$$

но новые данные генерируем по следующему алгоритму:

$$\varepsilon \sim 0.5N(-1,2.5^2) + 0.5N(1,0.5^2)$$
$$Y = 1 + 3X + (1 + 2X)\varepsilon$$

В этом случае разница между оценками должна быть значительно заметнее:

$$E(Y|X) = 1 + 3X$$

$$Median(Y|X) = 1.67 + 4.34X$$

$$Mode(Y|X) = 2 + 5X$$

Цель данного типа упражнений заключается в попытке показать, что вместе с оценкой условного среднего некоторые робастные методы дают оценки совершенно разных характеристик распределения. Получая различные оценки параметров от различных методов не обязательно делать вывод, что среднее неустойчиво. Возможно, что была найдена дополнительная информация об условном распределении. В некоторых случаях (3-е упражнение) эта информация весьма существенна.

Популярность OLS и связанных с ней моделей объясняется скорее простотой понимания условного среднего, чем тем, что исследователей целенаправленно интересует именно поведение условной средней. Проверка остальных характеристик распределения несложна, но может оказаться полезна.

Устройство функции

Данный раздел посвящен описанию устройства написанной в R функции для оценки модальной регрессии. Имплементация оценивающих процедур в языках программирования – процесс достаточно трудоемкий. Сложность состоит не в чисто вычислительной части, а в сочетании и проверке различных вариантов достижения одинаковых целей. Какие требование предъявляет пользователь, когда хочет использовать некоторый метод оценки? Автор данной статьи сформулировал для себя 4 базовых принципа

1) Скорость – функция должна быть проверена на узкие места, значительно замедляющие вычисления. Где возможно, желательно использовать функции, ссылающиеся на быстрые вычислительные рутины. У пользователя должен быть контроль и указания о том как различные параметры влияют на скорость вычисления

2) Гибкость – каждая процедура внутри функции должна быть почти полностью под контролем пользователя. Там где существует множество вариантов решения одной проблемы, желательно предоставлять доступ ко всем.

3) Сочетаемость – результат работы функции должен быть похож по структуре на стандартные оценки моделей, с добавлением в итоговый объект результатов, специфичных именно для данного метода оценки. Это также означает применимость и предсказуемость применения известных методов к результату оценки.

4) Удобство – функция должна работать при минимальном числе аргументов, в известной пользователю парадигме

Далее приведены описание результатов применения данных принципов в контексте функции оценки модальной регрессии:

Скорость

1) В силу итерационной природы поиска решения, любая значительная по времени процедура оценки может потенциально занять неприемлемо большой период времени если будет исполняться каждую итерацию. В силу этого число операций и их сложность минимизировалась, в том числе с помощью библиотек, использующих низкоуровневые языки программирования.

Гибкость

2) «R» – весьма популярный язык программирования для статистического анализа. Существует большое число библиотек, часто с пересекающимися или дублирующими функциями. Разница между ними часто состоит, скорее, в парадигме использования, нежели в целях. В контексте библиотеки, используемой в данной статье, это обозначает, что в случае, когда некоторые операции могут быть сделаны с помощью сторонних библиотек, опция их использования присутствует

Сочетаемость

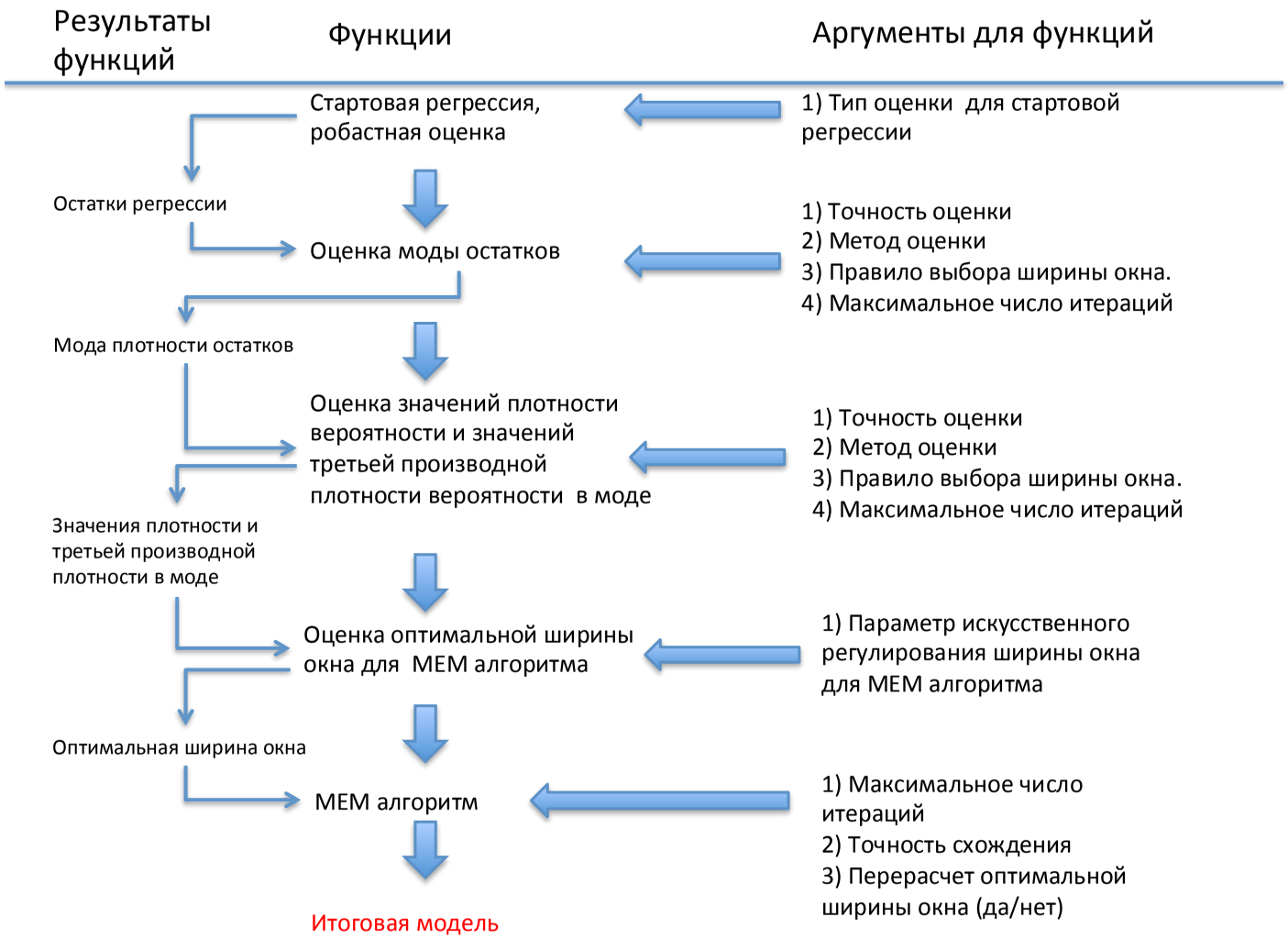
3) В «R» существует значительное число функций «верхнего уровня» - методов. Методы - это функции результат которых определяется авторами функций «нижнего уровня». Подобного рода функции нужны, чтобы пользователю не приходилось помнить отельную функцию для каждого типа объекта, используемую для некоторого одинакового действия. Классический пример такой функции : summary. Данная функция выводит ключевые характеристики для объекта к которому применяется. Результат предлагаемой функции, в

частности, (в силу природы МЕМ алгоритма) хорошо описывается как результат GLM (функция glm() в R) функции. Другие важные для оценки результата характеристики также включены в результаты регрессии, в добавок к стандартным результатам GLM.

Удобство

4) Удобство функции обеспечивается двумя характеристиками – легкостью доступа к внутренним механизмам (все возможные интересные изменения заложены в аргументы функции) и простота работы «из коробки». Чтобы функция работала «из коробки». требуется задать ряд предположительно оптимальных значений для аргументов по умолчанию. Разработанная функция требует только формулу и источник данных, для построения оценки. Оптимальные значения аргументов выбраны по симуляциям методом Монте-Карло.

Схема устройства созданной функции param.mod() в пакете R



Заключение

Мы рассмотрели поведение функции параметрической оценки модальной регрессии, и сравнили её поведение как при изменении параметров оценки внутри самой функции, так и на фоне классической OLS и популярным вариантом робастной оценки - медианной регрессией.

Были проведены симуляции методом Монте-Карло, с помощью которых проверялось поведение оценки в зависимости от ряда параметров, установленных нами. Мы показали, что оценка типа MODLR(scale = 1) состоятельна на сгенерированных данных. В силу неэквивалентности полученных результатов со статьей [16], мы установили чувствительность скорости сходимости при изменении параметра scale, отвечающего за уменьшение ширины окна в ядерной оценке плотности. Оценка типа MODLR(scale = 2) оказалась лучше с точки зрения скорости схождения к истинным значениям коэффициентов и близкой по результатам к результатам, полученными авторами [16]. Данное обстоятельство учтено в написанной к данной статье функции под R в виде соответствующего аргумента «scale». Была долю покрытия у оценок модальной, медианной и стандартной OLS, и установили преимущество модальной регрессии в отношении ширины предсказательного интервала.

Далее, для двух наборов реальных данных с известной методологией оценки, взятой из статей [16] и [17], мы сравнили результаты оценок коэффициентов модальной, медианной и МНК регрессий, и для 1-го набора данных из [16] установили интересное отличие ключевого коэффициента NOX, не рассмотренное авторами исходной статьи. Для набора данных по стоимости компьютеров [17] наши оценки привели к ряду разногласий в оценках коэффициентов между

различными методами MODLR, в силу чего разумной гипотезы о причинах отклонений MODLR оценок коэффициентов от OLS и MEDIAN оценок построить не удалось. В третьей части мы привели примеры упражнений для студентов, предназначенные для объяснения отличий между различными оцениваемыми характеристиками условной плотности распределения, и последствиями некорректной интерпретации полученных результатов, связанной с сильным акцентом текущих методов оценок на условном среднем. Данные упражнения легко задаются и решаются в любом эконометрическом пакете, и могут быть полезны для создания более объёмного понимания студентами последствий выбора различных характеристик условной плотности для оценки интересующих их зависимостей.

В конце статьи приведено краткое описание принципов, лежащих в основе написанной автором данной статьи функции для оценки моды условного распределения. Данная функция вскоре будет доступна на публичных ресурсах, и после подготовки документации будет превращена в часть библиотеки в репозитории «R». Направлений развития для дальнейшей работы достаточно много.

Во-первых, как замечают авторы [16], использование MEM алгоритма в принципе возможно для широкого ряда типов регрессий – нелинейных регрессий, частично-линейные регрессии с изменяющимися коэффициентам и иные.

Во-вторых, ключевым преимуществом MODLR является узкий интервал зоны покрытия, что эквивалентно преимуществу данной модели с точки зрения получения Value-at-risk оценок. В случае если распространить результаты MODLR оценок на временные ряды, возможно MODLR окажется весьма интересной альтернативой в ситуациях, когда стандартные методы не справляются с

гетероскедастичностью данных. MODLR или его эквивалент для временных рядов, в силу целей лежащих в его основе, может оказаться мощным методом оценки Value-at-risk. В-третьих, возможность корректно интерпретировать модели построенные с помощью MODLR может оказаться незаменимой в случаях, когда наблюдается уровень гетероскедастичности , делающий базовые методы оценки(даже для последующей коррекции с помощью известных процедур) весьма ненадежными. Поиск подобных примеров, впрочем, не является легкой задачей сам по себе.

И наконец, в-четвертых, требуется понять откуда возникает разница в скорости сходимости оценок на симуляциях методом Монте-Карло между функцией написанной автором данной статьи, и результатами [16]. Из этого также следует необходимость глубже оценить следствия выбора различных правил определения ширины окна для остатков стартовой регрессии в MEM алгоритме. В дополнение к вышеописанному, в [16] существует алгоритм более точного вычисления значений константы, в случае если известен угловой коэффициент. В данный момент эта опция MEM алгоритма находится на стадии программирования.

Список литературы

1. Botev, Z.I., Kroese, D.P. (2008). Non-asymptotic bandwidth selection for density estimation of discrete data. Methodology and Computing in Applied Probability. Volume 10, Number 3, 435-451,
2. Botev, Z.I., Kroese, D.P. (2009). The Generalized Cross Entropy Method, with Applications to Probability Density Estimation. Methodology and Computing in Applied Probability. DOI: 10.1007/s11009-009-9133-7
3. Botev. Z.I. and Kroese D. P. (2010). Efficient Monte Carlo simulation via the Generalized Splitting Method. Statistics and Computing. DOI: 10.1007/s11222-010-9201-4
4. Botev. Z.I., Grotowski J.F and Kroese D. P. (2010). Kernel density estimation via diffusion. Annals of Statistics. Volume 38, Number 5, Pages 2916--2957
5. Chaudhuri, P. and Marron, J. S. (1999). Sizer for Exploration of Structures in Curves. Journal of the American Statistical Association, 94, 807-823.
6. David Harrison, Daniel Rubelfield(1976): Hedonic Housing prices and the demand for clean air
7. Davies, P. L. and Kovac, A. (2004). Densities, Spectral Densities and Modality. Annals of Statistics, 32, 1093-1136.
8. Fisher, N. I. and Marron, J. S. (2001). Mode Testing via The Excess Mass Estimate. Biometrika, 88, 499-517.
9. Friedman J. H. and Fisher, N. I. (1999). Bump Hunting in High-Dimensional Data. Statistics and Computing, 9, 123-143.
10. Hall, P., Minnotte, M. C., and Zhang, C. (2004). Bump Hunting with Non-Gaussian Kernels. Annals of Statistics, 32, 2124-2141.

11. Li, J., Ray, S., and Lindsay, B. G. (2007). A Nonparametric Statistical Approach to Clustering via Mode Identification. Journal of Machine Learning Research, 8(8), 1687-1723
12. Muller, D. W. and Sawitzki, G. (1991). Excess Mass Estimates and Tests for Multimodality Journal of the American Statistical Association, 86, 738-746
13. Ray, S. and Lindsay, B. G. (2005). The Topography of Multivariate Normal Mixtures. Annals of Statistics, 33, 2042-2065.
14. Scott, D. W. (1992). Multivariate Density Estimation: Theory, Practice and Visualization.New York: Wiley.
15. T. Stengos and E. Zacharias(2005) : Intertemporal Pricing and Price Discrimination: A Semiparametric Hedonic Analysis of the Personal Computer Market.;
16. Weixin Yao and Longhai Li “A New Regression Model: Modal Linear Regression”
17. Yao, W. and Lindsay, B. G. (2009). Bayesian Mixture Labeling by Highest Posterior Density. Journal of American Statistical Association, 104, 758-767

Приложение

Симуляции методом Монте-Карло ;Таблица 10 (По сгенерированным данным)

Тип модели	Истинное значение оцениваемых параметров	n = 2000		n = 4000		n =10000	
		$\hat{\theta}$	$\hat{\sigma}$	$\hat{\theta}$	$\hat{\sigma}$	$\hat{\theta}$	$\hat{\sigma}$
MODLR (scale = 1)	$\alpha = 2$	1.86	0.14	1.89	0,12	1,93	0,06
	$\beta = 5$	5,07	0.19	5,06	0,15	5,02	0,11

Правило выбора параметра scale; Таблица 11(По сгенерированным данным)

Тип модели	Истинное значение оцениваемых параметров	Scale = 2		Scale = 3		Scale = 4	
		$\hat{\theta}$	$\hat{\sigma}$	$\hat{\theta}$	$\hat{\sigma}$	$\hat{\theta}$	$\hat{\sigma}$
MODLR (n = 400)	$\alpha = 2$	1.93	0.18	1.95	0.24	1.98	0.38
	$\beta = 5$	4.99	0.49	4.93	0.69	4.62	1.04

Описание переменных; Таблица 12(Данные о пригородах Бостона, 1970 Census)

MV	Медианная стоимость домов, занятых владельцами (в 1000$)
RM	Среднее количество комнат в доме (используется в форме RMSQ = RM^2, так как по итогам анализа [16] выяснилось что такая формулировка является наилучший спецификацией.
AGE	Доля домов построенных до 1940 года
BLACK	Доля чернокожего населения в форме 1000*(BLACK-0.63)^2, в силу параболической зависимости стоимости домов от доли чернокожего населения(более подробно в [16])
Log(STAT)	Доля населения с низким социальным статусом
CRIM	Уровень преступности
ZN	Доля участков площадью более 25 тысяч квадратных футов
INDUS	Доля занятая нерозничными бизнесами
TAX	Налог на собтственность с поправкой на индивидуальный город
PTRATIO	Отношение числа учеников к числу учителей
CHAS	Наличие выхода на реку у города (доступ к реке повышает стоимость недвижимости)
DIS	Взвешенные расстояния до 5 центров занятности
RAD	Индекс доступности радиальных дорог
NOX	Концентрация NO_2в воздухе (среднегодовая концентрация в долях на 100 миллионов).

Таблица 13 (Данные о пригородах Бостона, 1970 Census)

Оценка OLS регрессии

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.5600277  0.1534613  29.715  < 2e-16 ***
rmsq         0.0063754  0.0012526   5.090 5.11e-07 ***
log(dis)    -0.1945014  0.0290968  -6.685 6.27e-11 ***
log(rad)     0.0951725  0.0185873   5.120 4.38e-07 ***
crim        -0.0118862  0.0012347  -9.627  < 2e-16 ***
black        0.0003643  0.0001027   3.548 0.000425 ***
tax         -0.0004155  0.0001114  -3.730 0.000213 ***
ptratio     -0.0310270  0.0049725  -6.240 9.44e-10 ***
log(lstat)  -0.3692999  0.0228277 -16.178  < 2e-16 ***
noxsq       -0.6332648  0.1097092  -5.772 1.38e-08 ***
chas         0.0920751  0.0329154   2.797 0.005354 **
zn           0.0000605  0.0004935   0.123 0.902483

Multiple R-squared:  0.8059,    Adjusted R-squared:
0.8016
```

Таблица 14 (Данные о пригородах Бостона, 1970 Census)

Оценка медианной регрессии по стоимости домов в Бостоне
Coefficients:

	coefficients	lower bd	upper bd
(Intercept)	3.82591	3.50456	4.16052
rmsq	0.01257	0.01028	0.01594
log(dis)	-0.14216	-0.20665	-0.09498
log(rad)	0.06185	0.01947	0.08233
crim	-0.01148	-0.01795	-0.00783
black	0.00056	0.00033	0.00075
tax	-0.00037	-0.00056	-0.00014
ptratio	-0.02692	-0.03395	-0.02059
log(lstat)	-0.26274	-0.30139	-0.22334
noxsq	-0.45385	-0.64954	-0.26326
chas	0.06786	0.01591	0.10118
zn	**0.00046**	**-0.00050**	**0.00101**

Таблица 15 (Данные о пригородах Бостона, 1970 Census)

Оценка модальной регрессии MODLRkern по стоимости домов в Бостоне

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	3.628e+00	1.348e-01	26.910	< 2e-16	***
rmsq	1.476e-02	1.133e-03	13.023	< 2e-16	***
log(dis)	-1.205e-01	2.451e-02	-4.915	1.21e-06	***
log(rad)	4.696e-02	1.506e-02	3.118	0.001928	**
crim	-8.843e-03	1.180e-03	-7.496	3.07e-13	***
black	6.466e-04	9.372e-05	6.899	1.61e-11	***
tax	-3.124e-04	9.081e-05	-3.440	0.000631	***
ptratio	-2.787e-02	3.855e-03	-7.229	1.86e-12	***
log(lstat)	-2.339e-01	2.062e-02	-11.343	< 2e-16	***
noxsq	-4.321e-01	9.124e-02	-4.735	2.86e-06	***
chas	5.806e-02	2.646e-02	2.194	0.028687	*
zn	1.038e-04	3.801e-04	0.273	0.784973	

Таблица 16 (Данные о пригородах Бостона, 1970 Census)

Оценка модальной регрессии MODLR(scale =1) по стоимости домовв Бостоне

Coefficients

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	4.134e+00	1.293e-01	31.960	< 2e-16	***
rmsq	9.767e-03	1.075e-03	9.086	< 2e-16	***
log(dis)	-1.557e-01	2.399e-02	-6.488	2.12e-10	***
log(rad)	7.408e-02	1.511e-02	4.904	1.28e-06	***
crim	-1.064e-02	1.093e-03	-9.738	< 2e-16	***
black	5.262e-04	8.665e-05	6.072	2.52e-09	***
tax	-3.597e-04	9.028e-05	-3.984	7.79e-05	***
ptratio	-2.943e-02	3.965e-03	-7.422	5.09e-13	***
log(lstat)	-3.137e-01	1.943e-02	-16.144	< 2e-16	***
noxsq	-5.121e-01	8.943e-02	-5.726	1.79e-08	***
chas	7.711e-02	2.674e-02	2.884	0.0041	**
zn	2.366e-05	3.931e-04	0.060	0.9520	

Таблица 17 (Данные о пригородах Бостона, 1970 Census)

Оценка модальной регрессии MODLR(scale =2) по стоимости домов в Бостоне

Coefficients

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	3.646e+00	9.798e-02	37.206	< 2e-16	***
rmsq	1.434e-02	8.272e-04	17.340	< 2e-16	***
log(dis)	-1.166e-01	1.783e-02	-6.543	1.52e-10	***
log(rad)	4.945e-02	1.099e-02	4.499	8.52e-06	***
crim	-8.843e-03	8.508e-04	-10.394	< 2e-16	***
black	6.673e-04	6.693e-05	9.971	< 2e-16	***
tax	-3.153e-04	6.602e-05	-4.776	2.36e-06	***
ptratio	-2.806e-02	2.829e-03	-9.919	< 2e-16	***
log(lstat)	-2.404e-01	1.500e-02	-16.024	< 2e-16	***
noxsq	-4.190e-01	6.594e-02	-6.354	4.78e-10	***
chas	6.330e-02	1.937e-02	3.268	0.00116	**
zn	2.339e-05	2.792e-04	0.084	0.93327	

Таблица 18 (Данные PC magazine, 1993-1995)

Оценка модальной регрессии, MODLRkern по данным PC magazine

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	4.449e+00	4.956e-02	89.775	< 2e-16	***
log(speed)	1.868e-01	2.917e-03	64.055	< 2e-16	***
log(hd)	1.340e-01	3.869e-03	34.635	< 2e-16	***
log(ram)	2.021e-01	3.111e-03	64.970	< 2e-16	***
log(screen)	6.775e-01	1.805e-02	37.542	< 2e-16	***
cdyes	4.050e-02	2.878e-03	14.069	< 2e-16	***
multiyes	2.605e-02	3.299e-03	7.895	3.4e-15	***
premiumyes	-2.399e-01	3.884e-03	-61.778	< 2e-16	***
pent	-1.123e-02	4.022e-03	-2.791	0.00527	**
ads	2.337e-04	1.647e-05	14.184	< 2e-16	***
trend	-2.135e-02	2.987e-04	-71.461	< 2e-16	***

Таблица 19 (Данные PC magazine, 1993-1995)

Оценка модальной регрессии, MODLR(scale = 1) по данным PC magazine

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	4.393e+00	4.087e-02	107.495	<2e-16	***
log(speed)	1.653e-01	2.444e-03	67.651	<2e-16	***
log(hd)	1.535e-01	3.321e-03	46.211	<2e-16	***
log(ram)	2.123e-01	2.586e-03	82.089	<2e-16	***
log(screen)	6.927e-01	1.476e-02	46.919	<2e-16	***
cdyes	3.347e-02	2.423e-03	13.816	<2e-16	***
multiyes	2.632e-02	2.689e-03	9.787	<2e-16	***
premiumyes	-2.633e-01	3.328e-03	-79.111	<2e-16	***
pent	1.563e-03	3.435e-03	0.455	0.649	
ads	1.836e-04	1.388e-05	13.230	<2e-16	***
trend	-2.221e-02	2.547e-04	-87.175	<2e-16	***

Таблица 20 (Данные PC magazine, 1993-1995)

Оценка модальной регрессии, MODLR(scale = 2) по данным PC magazine

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	4.156e+00	2.427e-02	171.260	< 2e-16	***
log(speed)	1.494e-01	1.460e-03	102.341	< 2e-16	***
log(hd)	2.027e-01	2.007e-03	101.020	< 2e-16	***
log(ram)	2.116e-01	1.520e-03	139.177	< 2e-16	***
log(screen)	7.221e-01	8.682e-03	83.176	< 2e-16	***
cdyes	3.685e-02	1.480e-03	24.900	< 2e-16	***
multiyes	3.281e-02	1.562e-03	21.003	< 2e-16	***
premiumyes	-3.069e-01	2.023e-03	-151.725	< 2e-16	***
pent	1.203e-02	2.119e-03	5.679	1.42e-08	***
ads	9.702e-05	8.554e-06	11.342	< 2e-16	***
trend	-2.441e-02	1.565e-04	-155.961	< 2e-16	***