МИНОБРНАУКИ РОССИИ

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

Кафедра теории функций и стохастического анализа

# НЕПАРАМЕТРИЧЕСКОЕ ОЦЕНИВАНИЕ КРИВЫХ ЭНГЕЛЯ

АВТОРЕФЕРАТ МАГИСТЕРСКОЙ РАБОТЫ

Студентки 2 курса 248 группы
направления 09.04.03 — Прикладная информатика
механико-математического факультета
Бабаянц Терезы Артуровны

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

Заведующий кафедрой
д. ф.-м. н., доцент _____ С. П. Сидоров

Саратов 2022

# ВВЕДЕНИЕ

Кривые Энгеля описывают, как связаны расходы населения на определенные товары и услуги с их доходом или расходом. Кривые представляют собой график и получили свое название в честь немецкого статистика и экономиста Эрнста Энгеля, которые был первым, который исследовал систематические отношения между доходами и расходами в статье, опубликованной в 1857 году.

С тех пор кривые Энгеля стали важной частью эмпирического анализа спроса и используются во многих областях экономики, таких как анализ структурных изменений, теория роста, исследования международной торговли, а также при измерении инфляции.

В своей статье Энгель, собрав статистические данные ряда стран за некоторый период времени, отметил постепенное уменьшение относительной доли расходов на питание в бюджете по мере его роста. Таким образом, он вывел некий закон: чем меньше доход, тем большая часть его тратится на питание, которое также ухудшается, а также чем меньше доход, тем большая часть его приходится на физическое содержание и меньше остается для духовного развития. Способ оценки отношения расходов к доходу является непараметрическим, поскольку перед оценкой не указывается функциональная форма. По существу, он представляет собой яркий пример эконометрического метода, который основан на предпосылке, что исследователи могут обнаруживать теоретические взаимосвязи непосредственно из данных, и не ограничивается оценкой гипотез, основанных исключительно на существующих теориях.

Актуальность данной работы состоит в том, что она может дать практический навык построения и дальнейшего исследования кривых Энгеля для изучения изменений расходов домохозяйств на различные аспекты(питание, рестораны и кафе, досуг и развлечения, налоги, одежда и иные продукты) от общих расходов, услуги, а так же помочь в изучении различий потребительского выбора. Эти знания и навыки могут помочь при анализе социально-экономичекой ситуации в России.

**Целью магистерской работы** является изучение непараметрических

методов оценивания регрессионных моделей.

**Объект исследования** - российские домохозяйства, расходующие свои средства на определенные товары и услуги.

**Предмет исследования** - расходы на различные аспекты для комфортного проживания, в число которых входят расходы на питание, услуги, развлечения и досуг, одежду и иные товары.

Ниже представлены **задачи** настоящей работы:

— изучение методов оценивания непараметрических регрессионных моделей;

— изучение языка программирования R;

— проведение численного эксперимента по непараметрическому сглаживанию данных средствами языка R.

**Практическая значимость** состоит в том, что форма кривых Энгеля позволяет классифицировать изучаемые товары на нормальные (с выделением в этой группе предметов первой необходимости и предметов роскоши) и на относительно худшие. Если полученная кривая имеет положительный наклон, то товар относится к группе нормальных.

Работа прошла апробацию на различных конференциях, в частности, на ежегодной студенческой конференции "Актуальные проблемы математики и механики которую проводил механико-математический факультет СГУ в апреле 2022 года, в секции "Анализ данных в X Международной научно-практической конференции «Математическое и компьютерное моделирование в экономике, страховании и управлении рисками», ноябрь 2021 года.

**Содержание работы.** Работа состоит из введения, трех разделов, заключения, списка использованных источников, а также приложения, содержащего программный код и графики регрессионных моделей.

## Основное содержание работы

**Введение** содержит в себе определение понятий параметрической и непараметрической регрессионных моделей, а также понятия сглаживания данных. Кроме того, формулируются цели и задачи работы.

**Первый** раздел описывает исследование различных методов построения непараметрических моделей регрессии. Были изучены такие методы как:

ядерное сглаживание, оценка $k$-ближайших сосоедей, сглаживание сплайнами. Для данной работы был выбран метод ядерного сглаживания.

Регрессионный анализ позволяет проводить оценку зависимости между переменными посредством измерений, возмущенных случайными ошибками. Регрессионная кривая показывает общую взаимосвязь между объясняющей(независимой) переменной $X$ и переменной отклика(зависимой) $Y$. $X$ среднее значение переменной $Y$ задается функцией регрессии. Вид функции регрессии может дать представление, для каких значений $X$ следует ожидать наибольшие значения наблюдений $Y$ или обнаружен ли специальный вид зависимости между двумя переменными. Фокусирование на особо важных деталях средней зависимости переменной $Y$ от $X$ при ее интерпретации происходит благодаря сокращению ошибок наблюдения. Данная процедура получила определение - сглаживание.

Параметрический подход к анализу регрессионной зависимости получил свое название, так как предполагает, что вид функции полностью описывается конечным набором параметров. В качестве примера параметрической модели можно рассматривать полиномиальное уравнение регрессии, когда параметрами являются коэффициенты при неизвестных. Нужно отметить, что данный подход предполагает, что кривая может быть представлена в терминах параметрической модели, или, по крайней мере, имеется уверенность в том, что ошибка аппроксимации для наилучшего параметрического приближения пренебрежимо мала.

Предварительное задание параметрической модели может оказаться слишком ограничительным или чересчур малой размерности для аппроксимации непредвиденных характеристик, в то время как непараметрическое сглаживание предоставляет гибкие средства анализа неизвестных регрессионных зависимостей.

**Сглаживание данных** $(X_i, Y_i)_{i=1}^n$ включает в себя аппроксимацию кривой среднего значения отклика $m$ в соотношении регрессии (1).

При наличии $n$ пар данных $(X_i, Y_i)_{i=1}^n$ регрессионное соотношение может моделироваться следующим образом:

$$Y_i = m(X_i) + \varepsilon, i = 1, ..., n, \tag{1}$$

где $m$ - неизвестная функция регрессии, $\varepsilon$ - ошибки наблюдения. Целью регрессионного анализа является проведение разумной аппроксимации неизвестной зависимой функции $m$ (метод приближения, т.е. замены переменных для достижения более точного результата).

Особое внимание может привлекать кривая регрессии, а также ее некоторые производные или функции от производных. К примеру, экстремумы или точки перегиба. Существуют различные способы для представления набора данных. Если имеются повторные наблюдения в фиксированной точке $X = x$, оценивание $m(x)$ может быть выполнено только за счет использования среднего соответствующих значений $Y$. Однако получать повторные отклики для данного x, как правило, невозможно. В большинстве случаев соотношение регрессии в (1) содержит только одно значение переменной отклика $y$ и одно значение предикторной переменной $X_1$ которая может быть вектором в $\mathbb{R}^d$.

Сглаживание данных $(X_i, Y_i)_{i=1}^n$ включает в себя аппроксимацию кривой среднего значения отклика $m$ в соотношении регрессии (1).Особое внимание может привлекать кривая регрессии, а также ее некоторые производные или функции от производных. К примеру, экстремумы или точки перегиба. Существуют различные способы для представления набора данных. Если имеются повторные наблюдения в фиксированной точке $X = x$, оценивание $m(x)$ может быть выполнено только за счет использования среднего соответствующих значений $Y$. Однако получать повторные отклики для данного x, как правило, невозможно. В большинстве случаев соотношение регрессии в (1) содержит только одно значение переменной отклика $y$ и одно значение предикторной переменной $X_1$ которая может быть вектором в $\mathbb{R}^d$.

**Ядерное сглаживание** Идейно простой подход к представлению последовательности весов $\{W_{ni}(x)\}_{i=1}^n$ состоит в описании формы весовой функции $W_{ni}(x)$ посредством функции плотности со скалярным параметром, который регулирует размер и форму весов около $x$. Эту функцию формы принято называть ядром. Ядро — это непрерывная ограниченная симметричная ве-

щественная функция  с единичным интегралом.

$$\int K(u)du = 1. \tag{2}$$

Последовательность весов для ядерных оценок (для одномерного $x$) определяется как

$$W_{ni}(x) = K_{h_n}(x - X_i)/\hat{f}_{h_n}(x), \tag{3}$$

где

$$\hat{f}_{h_n}(x) = n^{-1}\sum_{i=1}^{n} K_{h_n}(x - X_i), \tag{4}$$

а

$$K_{h_n}(x - X_i) = h_n^{-1}K(u/h_n) \tag{5}$$

представляет собой ядро с параметром масштаба $h_n$. Подчеркнув зависимость $h = h_n$ от объема выборки $n$, условимся сокращенно обозначать последовательность весов (4) через $\{W_{hi}(x)\}_{i=1}^{n}$. Функция $\hat{f}_h n(\cdot)$ является ядерной оценкой плотности Розенблата и Парзена для (маргинальной) плотности переменной $X$. Вид (4) ядерных весов $\{W_{ni}(x)\}_{i=1}^{n}$ был предложен в работах Надарая и Ватсона, как следствие,

$$\hat{m}_h(x) = \frac{n^{-1}\sum_{i=1}^{n} K_h(x - X_i)Y_i}{n^{-1}\sum_{i=1}^{n} K_h(x - X_i)} \tag{6}$$

часто называют оценкой Надарая — Ватсона. Форма ядерных весов определяется ядром  в то время как размер весов параметризируется посредством переменной $h$, называемой шириной окна. Нормализация весов $\hat{f}_h(x)$ позволяет адаптироваться к локальной интенсивности переменной $X$ и, кроме того, гарантирует, что сумма весов равна единице. Вообще говоря, можно брать различные ядерные функции, но как практика, так и теория ограничивают выбор. Так, например, ядерные функции, принимающие очень малые зна-

чения, могут приводить к машинному нулю компьютера, поэтому разумно рассматривать такие ядерные функции, которые равны нулю вне некоторого фиксированного интервала. Обычно используется ядерная функция, обладающая некоторыми свойствами оптимальности. Это функция параболического типа:

$$K(u) = 0.75(1 - u^2)I(|u| \leq 1) \tag{7}$$

**Второй** раздел содержит описание языка программирования R и его возможности. R – это язык программирования и среда для статистических вычислений и графического анализа, сходный с языком S, первоначально разработанным в Белл Лабораториз (Bell Labs). Это программа для анализа данных с открытым кодом, которая поддерживается большим и активным исследовательским сообществом по всему миру. R – это чувствительный к регистру клавиатуры интерпретируемый язык программирования. Возможно либо вводить команды по одной в ответ на приглашение на ввод команды (>), либо запускать набор команд из исходного файла. Типы данных очень разнообразны: векторы, матрицы, таблицы данных и списки (совокупность нескольких объектов).

**Программная реализация метода.**

С помощью описанных в предыдущих главах средств языка R был написан программный код, по которому строятся графики параметрической регрессии, а также стало возможным осуществить аппроксимацию данных с помощью метода ядерного сглаживания. Программа задействует несколько видов ядер: прямоугольное ядро, гауссовское, а также биквадратное ядро.

Для исследования была взята сводка данных о российских домохозяйствах. Она представляет собой статистические данные о российских семьях. Исследование проводилось в 2020 году и содержало статистику расходов семей на продукты питания, походы в рестораны, затраты на алкогольную продукцию, развлечения и досуг, затраты на налоги, различные покупки для комфортного проживания. А также сводка содержит информацию и количестве детей в семье и сколько всего человек в семье.

В магистерской работе в соответствии с п**риложениями А, Б и В** представлены программный код и графики аппроксимации, построенные с использованием различных ядер и отражающий зависимость общих расходов

от расходов на определенные сферы, а также с помощью линейной регрессии.

Характер тренда во всех случаях отображает закономерность увеличения расходов на определенные сферы с повышением общих расходов (что исходит из увеличения доходов в целом). Также была рассмотрена параметрическая регрессия. Характер тренда в целом соответствует результатом непараметрической регрессии. Параметрическая регрессия ожидаемо дает более узкую доверительную полосу.

Для примера ниже представлены несколько графиков построения регрессионных моделей методом ядерного сглаживания с использованием различных ядер(ядро Гаусса, прямоугольное ядро и биквадратное ядро). Графики 1 - 3 отображают построение зависимости между расходами на походы в ресторн и кафе с использованием ядер Гаусса, биквадратного и прямоугольного соответственно. График 4 отображает линейную регрессию.

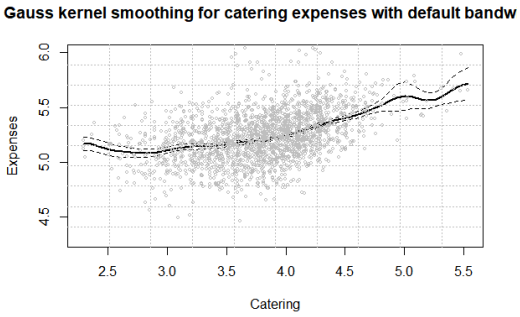


Рисунок 1 – График аппроксимации с помощью ядра Гаусса

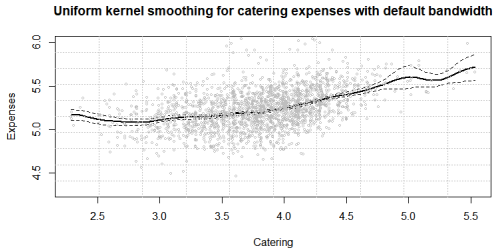


Рисунок 2 – График аппроксимации с помощью прямоугольного ядра

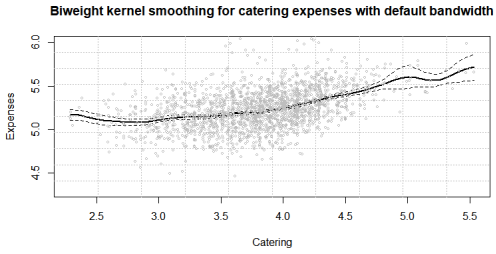


Рисунок 3 – График аппроксимации с помощью биквадратного ядра

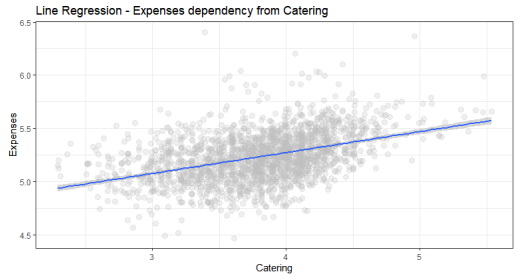


Рисунок 4 – График аппроксимации с помощью линейной регрессии

В соответствии с таблицей 1 представлены результаты применения функций MSE, MAPE - среднеквадратичные ошибки прогноза. Данные отображают процент ошибок, согласно которым появляется возможность выбрать наилучшую модель. Функция была применена к непараметрическим регрессиям с различным параметром ширины окна. Исходя из полученных результатов, можно сделать вывод, что оптимальной шириной окна для зависимости общих расходов от расходов на питание, услуги, одежду и иные расходы является 0,01. В остальных случаях (зависимость затрат общих от затрат на походы в ресторан, покупку алкоголя, налоги и досуг) - оптимальная ширина окна равна 0,1.

Таблица 1 – Вычисление параметров MSE и MAPE при аппроксимации данных с помощью непараметрической регрессией

| Expenses type | Bandwidth($h$) | MSE(%) | MAPE(%) |
|---|---|---|---|
| Food | 0,035 | 2,43 | 2,92 |
| Food | 0,1 | 2,43 | 2,92 |
| Food | 0,05 | 2,44 | 2,93 |
| Food | 0,01 | 2,41 | 2,84 |
| Catering | 0,03 | 3,022 | 4,12 |
| Catering | 0,1 | 2,43 | 2,92 |
| Catering | 0,05 | 3,01 | 4,09 |
| Catering | 0,01 | 2,92 | 3,88 |
| Alcohol | 0,035 | 3,03 | 4,3 |
| Alcohol | 0,1 | 2,43 | 2,92 |
| Alcohol | 0,05 | 3,02 | 4,27 |
| Alcohol | 0,01 | 2,96 | 4,1 |
| Services | 0,035 | 2,26 | 2,45 |
| Services | 0,1 | 2,43 | 2,92 |
| Services | 0,05 | 2,25 | 2,24 |
| Services | 0,01 | 2,22 | 2,32 |
| Tax | 0,035 | 2,79 | 3,7 |
| Tax | 0,1 | 2,43 | 2,92 |
| Tax | 0,05 | 2,78 | 3,67 |
| Tax | 0,01 | 2,71 | 3,49 |
| Other Products | 0,035 | 1,72 | 1,35 |
| Other Products | 0,1 | 2,43 | 2,92 |
| Other Products | 0,05 | 1,72 | 1,35 |
| Other Products | 0,01 | 1,68 | 1,3 |
| Leisure | 0,035 | 3,35 | 5,05 |
| Leisure | 0,1 | 2,43 | 2,92 |
| Leisure | 0,05 | 3,34 | 5,02 |
| Leisure | 0,01 | 3,28 | 4,85 |

## Основные результаты

1. Во всех случаях подтверждает общий закон о снижении доли затрат на питание по мере возрастания общих затрат. Также существенно уменьшается доля расходов на следующие аспекты: питание, алкоголь, досуг, налоги;
2. В меньшей степени уменьшаются затраты на походы в ресторан и в кафе;
3. Также возрастают затраты на обслуживание и различные услуги;
4. Затраты на категорию покупок, в которых входят повседневные покупки, а также одежда, сильно возрастают;
5. Также существенно возрастают общие затраты в зависимости от количества детей и членов семьи в целом;
6. Существенных различий в использовании разных типов ядер не было обнаружено. Однако, более узкую доверительную полосу можно получить, используя биквадратное ядро, а более точную аппроксимацию можно построить, применяя процедуру с ядром Гаусса.
7. Основываясь на результатах применения функций MSE, MAPE, можно сделать вывод, что для многих типов расходов наиболее оптимальной является аппроксимация с шириной окна, равной 0,1.

По результатам численного эксперимента можно сделать следующие выводы.

— С увеличением общих затрат, у некоторых групп существенно возрастают затраты на питание, судя по графику рассеивания данных и функции аппроксимации, которая возрастает;
— В меньшей степени возрастают затраты на походы в ресторан и в кафе;
— Также возрастают затраты на обслуживание и различные услуги;
— Затраты на категорию покупок, в которых входят повседневные покупки, а также одежда, в целом возрастают. Однако, судя по графику рассеивания, несущественно;
— Также несущественно возрастают общие затраты в зависимости от количества детей и членов семьи в целом.
— Наиболее точную аппроксимацию можно получить с помощью ядра Гаусса. Ширина окна влияет на степень гладкости функции. Чем она

выше, тем более гладкой получается функция аппроксимации.

— Основываясь на результатах применения функций MSE, MAPE, можно сделать вывод, что для некоторых типов расходов наиболее оптимальной является аппроксимация с шириной окна, равной 0,01. В других случаях - 0,1