Федеральное государственное бюджетное учреждение науки
ИНСТИТУТ МАТЕМАТИКИ им. С. Л. СОБОЛЕВА
Сибирского отделения Российской академии наук

На правах рукописи

Шаталин Евгений Викторович

Эмпирический мост и задачи тестирования адекватности регрессионных моделей анализа данных

05.13.17 - Теоретические основы информатики

Диссертация на соискание ученой степени
кандидата физико-математических наук

Научный руководитель
д. ф.-м. н., профессор
С. Г. Фосс

Новосибирск 2017

Содержание

Введение

В диссертационной работе строится и обосновывается алгоритм анализа адекватности линейных регрессионных моделей на порядковые статистики с двумя параметрами, а также рассматриваются аспекты практического применения построенного алгоритма к задачам анализа данных. В основе предлагаемого алгоритма лежит кусочно-линейная случайная ломаная, так называемый эмпирический мост, построенный по остаткам линейной регрессионной модели. Конструкция эмпирического моста является удобным механизмом анализа адекватности (соответствия) той или иной регрессионной модели наблюдаемому процессу.

Наиболее простая (однопараметрическая) модель линейной регрессии имеет вид

$$Y_{ni} = \theta X_{ni} + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

где X_{ni} (значения регрессора) обычно предполагаются фиксированными (неслучайными) величинами, $\theta \in \mathbf{R}$ — неизвестный (подлежащий оценке) параметр регрессионной модели, $\varepsilon_1, \ldots, \varepsilon_n$ (регрессионные ошибки) — независимые, одинаково распределенные случайные величины с нулевым математическим ожиданием и конечной ненулевой дисперсией σ^2. Y_{ni} — значения зависимой переменной или, как их часто еще называют, значения отклика. Далее двойные индексы будем опускать там, где это не вызывает недоразумений.

Неизвестный параметр регрессионной модели θ обычно оценивают с помощью метода наименьших квадратов (МНК-оценка параметра), получая оценку $\widehat{\theta} = \overline{XY}/\overline{X^2}$, где через $\overline{V} = \frac{1}{n}\sum_{i=1}^{n} V_i$ обозначено выборочное среднее случайных величин $V_1, \ldots, V_n$. На основании регрессионной модели строятся прогнозные значения $\widehat{Y}_{ni} = \widehat{\theta} X_{ni}$. Остатками регрессии называют случайные величины $\widehat{\varepsilon}_{ni} = Y_{ni} - \widehat{Y}_{ni}$.

Приведем определение основного объекта исследования диссертации — эмпирического моста, а также некоторые связанные с ним полезные факты. Эм-

пирический мост — это кусочно-линейная случайная ломаная $\widehat{Z}_n = \{\widehat{Z}_n(t),\ 0 \leqslant t \leqslant 1\}$ с узлами в точках

$$\left(\frac{k}{n},\ \frac{\widehat{\Delta}_{nk} - \frac{k}{n}\widehat{\Delta}_{nn}}{\sqrt{\widehat{\sigma}^2 n}} \right),$$

где $\widehat{\Delta}_{nk} = \widehat{\varepsilon}_{n1} + \ldots + \widehat{\varepsilon}_{nk},\ k = 0, 1, \ldots, n,\ \widehat{\Delta}_{n0} = 0,\ \widehat{\sigma^2} = \overline{\widehat{\varepsilon}^2} - (\overline{\widehat{\varepsilon}})^2$. Далее в двойных индексах будем опускать индекс n там, где это не вызывает недоразумений.

Сформулируем кратко основные шаги разработанного в диссертации алгоритма анализа адекватности регрессионных моделей (более подробное описание приведено в параграфе 1.3 главы 1):

Шаг 1. С помощью МНК оцениваются параметры регрессионной модели.

Шаг 2. Рассчитываются регрессионные остатки модели.

Шаг 3. Оценивается выборочная дисперсия модели $\widehat{\sigma^2}$.

Шаг 4. По регрессионным остаткам строится эмпирический мост.

Шаг 5. Подбирается функционал, предельное распределение которого от эмпирического моста известно или табулировано.

Шаг 6. Рассчитывается значение выбранного функционала от эмпирического моста.

Шаг 7. Если значение функционала превышает свое пороговое значение, то гипотеза об адекватности регрессионной модели отклоняется, в противном случае гипотеза принимается.

Очевидно, что применение построенного алгоритма наталкивается на необходимость отыскания предельного распределения эмпирического моста, что и сделано в настоящей диссертации для ряда регрессионных моделей.

Механизм работы метода эмпирического моста можно наглядно описать следующим образом. Если предложенная регрессионная модель неправильно описывает данные, то значения отклика Y_i будут систематически уклоняться от регрессионной кривой, и это уклонение можно выявить суммированием регрессионных остатков $\hat{\varepsilon}_i$ (разностей между наблюдаемыми и прогнозируемыми значениями). Для изучения значимости этих уклонений необходимо знать

предельное распределение процесса центрированных и нормированных частичных сумм регрессионных остатков. Этот процесс и называется эмпирическим мостом и был определен выше. Нормировка, присутствующая в определении эмпирического моста, как и вообще в разных версиях центральной предельной теоремы, необходима для сходимости процесса сумм остатков к предельному. Эмпирический мост - это процесс самонормированных сумм: вместо неизвестной дисперсии регрессионных ошибок используется выборочная дисперсия регрессионных остатков. Отметим, что в случае равенства суммы регрессионных остатков нулю с вероятностью единица (как в некоторых изучаемых ниже моделях) центрирования не требуется.

Таким образом, метод эмпирического моста является (наряду с описанными в параграфе 1.1 главы 1) одним из способов проверки адекватности регрессионной модели исследуемым данным. Более того, если для описания данных предложено несколько моделей (как в примере с зависимостью массы тела от роста, который будет приведен ниже), то вычисление достигнутых уровней значимости позволяет выбрать модель, наилучшим образом описывающую исследуемые данные.

Таким образом, можно достаточно быстро отсеивать еще на первом этапе исследования неподходящие модели, а также проводить сравнение подходящих моделей между собой. Подход к анализу соответствия данных вероятностным моделям, основанный на функционалах от эмпирического моста, разрабатывался в работах [16], [22] и применялся к анализу текстов в [16], тестированию моделей цен на недвижимость и автомобили в [2], [62], поиску неоднородностей строительных конструкций в [23].

Отметим также прикладную полезность графического изображения эмпирического моста для решения известной задачи о разладке. Задача о разладке состоит в скорейшем обнаружении изменения вероятностных характеристик наблюдаемого процесса. После появления основополагающих работ Ширяева и Зигангирова [39] и [21] соответственно и позднее монографии [40] интерес к

задаче разладки только возрастает. Мы не будем вдаваться в подробное описание задачи о разладке, так как это не является целью нашего исследования. Желающие могут ознакомиться с предметом, изучив полезный обзор [8].

Если говорить непосредственно о регрессиях, то задача о разладке — это, как правило, задача обнаружения изменения параметров регрессионной модели со временем. Для регрессионной модели она была впервые поставлена в [69], [70]. Разладка в авторегрессионных процессах широко изучена в [12]–[10] (см. также ссылки в них).

Анализ графиков же эмпирического моста позволяет выявить моменты разладки регрессии. А именно, если на каком-то из участков мы видим, что график очень быстро растет или наоборот снижается, то резонно говорить о разладке в регрессионной модели в точке экстремального значения эмпирического моста. После этого выборка может быть разбита на ряд кусков (в точках экстремальных значений моста), на каждом из которых строятся своя регрессионная модель и соответствующие свои оценки параметров. Процедура дробления выборки повторяется до тех пор, пока на каждом из участков не будет получено приемлемого приближения, а также отсутствие непропорционального изменения графика эмпирического моста. Данный прием будет продемонстрирован нами на примере в параграфе 2.2 главы 2 настоящей диссертации.

При условии сходимости оценки дисперсии к ее истинному значению, то есть при условии $\widehat{\sigma^2} \xrightarrow{p} \sigma^2$ при $n \to \infty$, слабые пределы в пространстве непрерывных на $[0,1]$ функций $C(0,1)$ эмпирического моста и случайной ломаной, построенной по точкам

$$\left(\frac{k}{n}, \ \frac{\widehat{\Delta}_k - \frac{k}{n}\widehat{\Delta}_n}{\sigma\sqrt{n}} \right),$$

совпадают.

Отметим также, что указанная ломаная получается непрерывным в равномерной метрике на $[0,1]$ преобразованием $x^0(t) = x(t) - tx(1)$, отображающим

x в x^0, случайной ломаной Z_n, построенной по точкам

$$\left(\frac{k}{n},\ \frac{\widehat{\Delta}_k}{\sigma\sqrt{n}}\right).$$

Таким образом, с помощью простого преобразования и вполне естественного предположения о состоятельности выборочной дисперсии мы перешли от эмпирического моста к более простому процессу.

В случае, когда в качестве регрессора используется вектор значений неслучайной гладкой функции в равноотстоящие моменты времени, предельный процесс для Z_n изучен MacNeill в [63]. Это центрированный гауссовский процесс с ковариационной функцией

$$K_f(s,t) = \min(s,t) - \int_0^s \int_0^t g(x,y)\,dx\,dy,\ \ s,t \in [0,1],$$

где функция $g(x,y)$ определяется через регрессор. Позднее Bischoff в [47] улучшил результат MacNeill, обобщив его на случай непрерывной функции, порождающей регрессор.

В работе [44] рассматривается многопараметрическая регрессионная модель, в которой в качестве регрессора выступают неслучайные векторы вида $(1, i/n, \ldots, (i/n)^p)^T$, то есть значения степенных функций в равноотстоящие моменты времени. Для этой модели проверяется нулевая гипотеза, которая состоит в том, что вектор параметров регрессии не зависит от времени и равен β_0. В качестве альтернативы рассматривается гипотеза, состоящая в том, что до некоторого неизвестного момента времени k^* параметр регрессии равен β_0, а в момент времени k^* происходит изменение параметра, и все оставшееся время он равняется $\beta_A \neq \beta_0$. Проверка описанной гипотезы строится на основе критерия, использующего статистику, построенную на основе сумм квадратов регрессионных остатков.

Задача проверки нормальности регрессионных ошибок успешно решена в [31]. Здесь анализ регрессионных остатков приводит к другому предельному гауссовскому процессу. В статье показано, что распределение статистики омега-квадрат от этого предельного процесса вычисляется в явном виде.

Целью же настоящей диссертации является построение решающих правил для анализа адекватности линейных регрессионных моделей на порядковые статистики с двумя параметрами (что эквивалентно отысканию предельных процессов для эмпирического моста, построенного по остаткам указанных моделей).

Порядковые статистики довольно часто встречаются в статистической науке. Порядковым статистикам посвящены целые монографии, например, [35], [20] и обширные ссылки в них, что говорит о высокой значимости этого объекта. Порядковые статистики получаются путем упорядочения элементов выборки по возрастанию. Оказывается, такое упорядочение позволяет каждому члену такого упорядоченного ряда (вариационный ряд) давать важную информацию об истинном распределении. В частности, первая и последняя порядковая статистики дают приближенное представление о коридоре изменения возможных значений исследуемого объекта, их разность говорит о степени разброса его значений. Средний член вариационного ряда, или медиана, характеризует своеобразный центр рассматриваемых данных.

В главе 1 настоящей диссертации будет построен алгоритм анализа адекватности для ряда линейных регрессионных моделей на порядковые статистики, а также сформулированы и доказаны предельные теоремы, обосновывающие построенный алгоритм.

В параграфах 1.1 и 1.2 главы 1 приведены необходимые теоретические и исторические сведения регрессионного анализа и теории случайных процессов соответственно.

В параграфе 1.3 главы 1 будут введены необходимые понятия, сформулирован разрабатываем алгоритм и сформулированы основные теоремы (теоремы

1–3) диссертационной работы.

В параграфе 1.4 главы 1 будет рассмотрена модель (1) со случайными и зависимыми между собой элементами регрессора и будет найдено предельное распределение процесса Z_n, а следовательно, и эмпирического моста (теорема 1). А именно, в качестве регрессора будет использоваться набор $\{\xi_{1:n}, \ldots, \xi_{n:n}\}$ порядковых статистик, построенных по выборке из некоторого (вообще говоря, неизвестного) распределения, то есть $X_i = \xi_{i:n}$. Случайные величины $\xi_1 \ldots \xi_n$ предполагаются независимыми, одинаково распределенными с функцией распределения F и не зависящими от случайных величин $\varepsilon_1 \ldots \varepsilon_n$. Заметим, что независимость $\xi_1, \ldots, \xi_n$ и $\varepsilon_1, \ldots, \varepsilon_n$, вообще говоря, не является сама собой разумеющейся и на практике нуждается в проверке: в предположениях совместной нормальности достаточно проверять гипотезу о равенстве коэффициента корреляции нулю, а в общем случае может быть применен критерий независимости типа хи-квадрат.

Эмпирический мост для модели выборки слабо сходится к стандартному броуновскому мосту, а сходимость эмпирического моста в модели однопараметрической линейной регрессии требует доказательства. При этом, как будет доказано ниже (теорема 1), предельный гауссовский процесс отличается от стандартного броуновского моста.

Предложенная модель возникает всякий раз, когда анализируется двумерная выборка и предполагается линейная зависимость одной ее компоненты (отклика) от другой (регрессора) с точностью до случайной регрессионной ошибки. Упорядочение значений регрессора по возрастанию естественно возникает, например, при графическом изображении регрессионной зависимости. Например, для исследования зависимости массы тела человека W_i от роста H_i в [71] предложена модель пропорциональности массы тела квадрату роста. Отметим, что пропорциональность предполагается для лиц одной возрастной группы и одного пола. В [60] на основании масштабных исследований показано, что эта зависимость является наилучшей в классе степенных зависимостей. Эту про-

порциональность можно проинтерпретировать в виде двух различных регрессионных моделей: $\ln W_i = \ln(\theta H_i^2) + \varepsilon_i$ и $W_i = \theta H_i^2 + \varepsilon_i$. Отметим, что рост и массу тела индивидуума можно считать случайными величинами. При этом, как было отмечено выше, необходимо проверить независимость роста H_i и корректирующего фактора ε_i.

Первая из этих моделей после замены переменных $Y_i = \ln(W_i/H_i^2)$, $\alpha = \ln\theta$ приводит к модели выборки $Y_i = \alpha + \varepsilon_i$, а вторая является моделью однопараметрической линейной регрессии. Для того, чтобы проверить соответствие каждой модели реальным данным, предлагается упорядочить наблюдения по неубыванию величин H_i.

Другой пример — модель радиоактивного распада $C_i = C\exp(-\alpha T_i + \varepsilon_i)$ в радиоуглеродном анализе археологических памятников (см. [43]). Здесь C — начальная концентрация изотопа (известная и одинаковая для всех исследуемых образцов), C_i — концентрация в образце, возраст которого предполагается равным T_i. Логарифмируя, приходим к рассматриваемой в работе модели. Здесь также образцы упорядочиваются по возрасту для проверки гипотезы о том, что возраст T_i в каждом случае определен достоверно.

В параграфе 1.5 главы 1 будет рассмотрена двухпараметрическая регрессионная модель: $\{Y_{ni},\ 1 \leqslant i \leqslant n,\ n \geqslant 1\}$

$$Y_{ni} = a + bX_{ni} + \varepsilon_i,\ i = 1, \ldots, n, \tag{2}$$

где, как и выше, $\{X_{ni},\ 1 \leqslant i \leqslant n,\ n \geqslant 1\}$ порядковые статистики, построенные по выборке $\xi_1, \ldots, \xi_n,\ n \geqslant 1$, с распределением F, не зависящей от регрессионных остатков, то есть $X_{ni} = \xi_{i:n}$.

Результатом параграфа 1.5 главы 1 будет отыскание предельного процесса для эмпирического моста в случае двухпараметрической регрессионной модели на порядковые статистики (2).

В заключении параграфа 1.5 главы 1 мы приведем другое, более „прямое“

и короткое, доказательство теоремы 1, основанное на методах параграфа 1.5 главы 1.

В параграфе 1.6 главы 1 будет рассмотрена регрессионная модель, аналогичная рассматриваемой в параграфе 1.5, с тем отличием, что регрессионные ошибки управляются марковской цепью. В англоязычной литературе для обозначения регрессионных ошибок, управляемых марковской цепью, используют термин „Markov-modulated noise“.

Чтобы определить модель, мы введем три взаимно независимых семейства случайных величин:

1) $\{\varepsilon_i^v, i \geq 1, 1 \leq v \leq M\}$ — семейство независимых случайных величин, где $\{\varepsilon_i^v, i \geq 1\}$ одинаково распределены для каждого v, $\mathbf{E}\varepsilon_1^v = 0$, $\mathbf{Var}\varepsilon_1^v = \sigma_v^2 \geq 0$ и $\sum_{v=1}^M \sigma_v^2 > 0$;

2) $\{\xi_i\}_{i=1}^\infty$ – последовательность независимых одинаково распределенных случайных величин с функцией распределения F и конечной положительной дисперсией $\mathbf{Var}\xi$;

3) $\{V_i\}_{i=1}^\infty$ — неразложимая апериодическая цепь Маркова, заданная на множестве состояний $\{1, \ldots, M\}$, со стационарным распределением $\{\pi_i\}_{i=1}^M$.

Рассмотрим модель:

$$Y_i = a + b\xi_i + \varepsilon_i^{V_i},\ n \geqslant 1, i = 1, \ldots, n.$$

Таким образом, у нас имеется последовательность трехмерных вектор-строк $(Y_i, \xi_i, \varepsilon_i^{V_i})$. Упорядочивая эти векторы (для каждого n) по второй компоненте, мы получим вектор-строки $(Y_{ni}, X_{ni}, \varepsilon_{ni}^{V_i})$. Здесь для каждого $n = 1, 2, \ldots$ $X_{ni} = \xi_{i:n}$ — порядковые статистики, а величины Y_{ni}, $\varepsilon_{ni}^{V_{ni}}$ соответствующие X_{ni} значения Y и ε^V соответственно.

В итоге мы приходим к регрессионной модели:

$$Y_{ni} = a + bX_{ni} + \varepsilon_{ni}^{V_{ni}},\ n \geqslant 1, i = 1, \ldots, n. \tag{3}$$

Результатом параграфа 1.6 главы 1 будет отыскание предельного процесса для эмпирического моста в случае двухпараметрической регрессионной модели, ошибки в которой управляются цепью Маркова (3).

В случае вырожденной цепи Маркова, описанная модель является частным случаем результатов, полученных в [76]. Ковалевский в [62] использовал эту частную модель для анализа зависимости цены автомобиля и года его производства в соответствии с объявлением о продаже. В данной модели имеет место сильная зависимость дисперсии от даты подачи объявления, стандартный тест на гомоскедастичность (равенство дисперсий случайных ошибок) обнаруживает эту особенность. Модель (3) охватывает случай гетероскедастичности (противоположность гомоскедастичности), допуская управления шумом посредством цепи Маркова. Дисперсия в данной модели может варьироваться в широком диапазоне значений, однако асимптотический результат сохраняется и при этом не зависит от распределения цепи Маркова. Индекс n, как и было указано выше, мы будем опускать.

В параграфе 1.7 главы 1 мы, проведем сравнение предлагаемого алгоритма анализа регрессионных моделей, основанного на конструкции эмпирического моста, с широко известным F-тестом (более подробно об F-тесте можно узнать из параграфа 1.1 главы 1). Будет приведен пример регрессионной модели несоответствующей реальным данным, которая принимается на основании F-теста, но при этом решительно отвергается с помощью эмпирического моста.

Глава 2 настоящей диссертации будет посвящена иллюстрации практических приложений построенного в главе 1 алгоритма.

В параграфе 2.1 главы 2 мы рассмотрим некоторые аспекты практического применения результатов, полученных в настоящей диссертации, а именно теорем 1 и 2. Сформулируем необходимые следствия из указанных теорем и подробно разберем варианты приложения результатов на основании двух статистических критериев: хи-квадрат и омега-квадрат.

Параграфы 2.2-2.4 главы 2 будут посвящены непосредственному примене-

нию полученных результатов (будут рассмотрены три практических задачи). В каждой из указанных задач мы проведем анализ адекватности предлагаемых регрессионных моделей эмпирическим данным с помощью критерия на основании конструкции эмпирического моста. В обоих случаях нами будет указана наиболее подходящая модель в смысле применяемого критерия анализа соответствия.

В частности, в параграфе 2.2 мы исследуем зависимость курсов американского доллара и евро. Сравнение будет производиться не напрямую, а через относительные курсы евро и доллара к швейцарскому франку. В параграфе исследуются две гипотезы: евро следует за долларом или наоборот доллар следует за евро. В качестве исходной была выбрана выборка курсов евро/франк и доллар/франк за период с 1 января 2011 года по 1 января 2014 года. В результате исследования с помощью статистического критерия установлена зависимость курсов в трех зонах, приведен реально достигнутый уровень значимости.

Параграф 2.3 главы 2 посвящен выбору регрессионной модели зависимости массы тела человека от его роста на основе конструкции эмпирического моста. В параграфе рассматривается двумерная выборка объема 750 значений роста и веса студенток первого курса Волгоградского медицинского университета. Далее выборка упорядочивается по значениям роста и предлагается для анализа двенадцать регрессионных моделей зависимости массы человеческого тела от его роста. С помощью статистических пакетов R и MatLab проводятся необходимые вычисления и в итоге делается вывод о предпочтительности одной из рассматриваемых моделей. Полученные в этом параграфе графики приведены в приложении к диссертации.

Наконец, параграф 2.4 главы 2 посвящен проверке гипотезы о линейной зависимости длины прыжка человека и его роста. В качестве исходных данных, как и в параграфе 2.3 главы 2, были взяты биометрические данные студентов Волгоградского медицинского университета.

Актуальность темы. Объектом исследования настоящей работы являют-

ся проблемы анализа данных и обработки информации. Предмет исследования – вероятностно – статистические методы анализа данных, а именно методы тестирования адекватности регрессионных моделей. Цель исследования – построение решающих правил (статистических критериев) для анализа соответствия линейных регрессионных моделей с двумя параметрами обрабатываемым данным. Мотивация исследования – отсутствие каких-либо алгоритмов, позволяющих получить не только качественный, но и количественный результат, чувствительных при этом к систематическим уклонениям регрессионных остатков.

В современном мире обилия информации набирают актуальность исследования процессов создания, накопления и обработки информации. Важным методом анализа данных, обнаружения скрытых закономерностей в данных является исследование регрессионных моделей. Для изучаемого массива данных, как правило, строится громадное число регрессионных зависимостей, и важно научиться определять (как можно реже ошибаясь), какие из них являются истинными, а какие ложными. Разработка решающих правил для такого анализа ведет отсчет с обсуждаемой выше работы МакНилла. В своей работе МакНилл изучал временные ряды данных. Однако, помимо временных рядов, огромный практический интерес представляет изучение данных в виде набора пар связанных значений. Такого рода задачи возникают всякий раз, когда необходимо провести анализ пар данных на предмет их взаимозависимости. И, в случае обнаружения зависимости, необходимо подобрать адекватную модель этой зависимости. Такого вида данные и изучаются в настоящей диссертации. Для анализа эти пары упорядочиваются по одной из компонент, что приводит к модели регрессии на порядковые статистики. В качестве разрешающей процедуры предлагается использовать конструкцию эмпирического моста. В диссертации строятся и теоретически обосновываются решающие правил и приводятся алгоритмы и примеры их практического применения.

Цель работы. В качестве целей данной диссертационной работы выступают:

– построение и теоретическое обоснование решающих процедур (критериев) и алгоритмов, основанных на конструкции эмпирического моста, для анализа адекватности линейных регрессионных моделей исследуемым данным, обнаружения скрытых закономерностей и ложных регрессионных зависимостей в данных;

– сравнение алгоритма, основанного на конструкции эмпирического моста, с другими методами анализа адекватности регрессионных моделей;

– исследование практической применимости и результативности использования полученного алгоритма на реальных прикладных задачах и обозначение основных рекомендаций для практического применения построенных решающих правил, основанных на статистических критериях типа хи-квадрат и омега-квадрат;

– отыскание и исследование предельных процессов для эмпирических мостов, построенных по остаткам линейных регрессионных моделей на порядковые статистики.

Методы исследования. В работе используются методы теории случайных процессов, математической статистики, теории меры, регрессионного анализа, статистического анализа, математического анализа, линейной алгебры, методы обработки информации. Все проделанные в работе расчеты проведены с помощью пакета для математических расчетов MatLab и свободно распространяемого пакет обработки данных R[1].

Научная новизна.

Полученные в данной диссертационной работе решающие правила являются новыми, весьма результативными методами анализа данных. Лежащие в их основе предельных теоремы также являются новыми теоретическими результатами.

Как показало сравнение с классическим F-тестом, предлагаемый в диссертации подход не содержит свойственного F-тесту недостатка (сложности при

[1] http://www.r-project.org

сравнении моделей с различным числом параметров). Этот факт открывает новые горизонты анализа регрессионных моделей, что и проиллюстрировано практическими применениями доказанных теорем для получения новых прикладных результатов о зависимостях (а) массы тела от роста человека; (б) длины прыжка от роста человека; (в) курсов валют.

Важным новым и отличительным от других работ моментом диссертационного исследования является рассмотрение регрессионных моделей с порядковыми статистиками в качестве регрессора.

Еще одной отличительной особенностью исследования является отказ от классического предположения регрессионного анализа о гомоскедастичности, которое на практике не всегда выполнено, что также несет в себе научную новизну. Исследование модели, в которой ошибки управляются цепью Маркова, показывает универсальность конструкции эмпирического моста и для случая „неклассической“ регрессии.

Теоретическая ценность и практическая значимость. Результаты диссертационной работы могут быть использованы в различных отраслях науки и техники, в задачах, где необходимо обнаружить зависимость между данными, а также отсечь ложные зависимости. В частности, полученные результаты могут применяться в задачах финансовой математики, медицины, инвестиционного анализа, эконометрики, биометрики и т.д.

Исследование описываемых в диссертации зависимостей сталкивается с принципиальными трудностями, разрешение которых само по себе имеет высокую научную ценность. В частности, возникают постановочные трудности, которые преодолеваются с помощью подбора адекватного аппарата описания моделей и их исследования. Кроме того, исследование регрессионных моделей на порядковые статистики затрудняется наличием зависимости регрессионных величин, что в данной диссертации решается путем замены значений регрессора на их математические ожидания. Последнее основано на применении теоремы Хефдинга.

Полученный алгоритм анализа данных весьма универсален, что открывает большие перспективы его применения. С помощью эмпирического моста можно еще на первом этапе исследования быстро и эффективно отвергать ложные регрессионные модели. Это приводит к существенной экономии вычислительных мощностей, оптимизации времен вычислительных циклов, что является очень важным в современном мире "больших данных".

Кроме того, полученные теоретические результаты могут быть использованы в научных исследованиях, посвященных проблеме анализа данных, распознавания образов и обнаружения зависимостей в данных, а также в спецкурсах для студентов и аспирантов по указанным разделам науки.

Личный вклад. Основные научные результаты, выносимые на защиту, численные расчеты получены автором самостоятельно. Постановки задач предложены научным руководителем. В совместных работах А.П. Ковалевскому принадлежит интерпретация полученных результатов.

На защиту выносится (а) разработанный алгоритм и построенные на его основе решающие правила, обеспечивающие анализ соответствия регрессионных моделей реальным данным и (б) совокупность математических результатов в виде предельных теорем, обосновывающих предлагаемые методы анализа.

Апробация работы. Основные результаты диссертации неоднократно были представлены на заседаниях семинара по теории вероятностей и математической статистики лаборатории теории вероятностей и математической статистики Института математики им. С.Л. Соболева, г. Новосибирск, на заседании семинара „Статистика случайных процессов и ее приложения“ в Томском государственном университете, а также на конференциях:

1) Международная научная студенческая конференция–2011 и Международная научная студенческая конференция–2014 (г. Новосибирск).

2) V International Conference „Limit Theorems in Probability Theory and Their Applications“, 2011 (Novosibirsk).

3) Четырнадцатый всероссийский Симпозиум по прикладной и промышлен-

ной математике, 2013 (Москва).

4) 11th International conference on ordered statistical data, 2014 (Bedlewo, Poland).

Также результаты работы (теоремы 1 и 2) включены в материалы курсов „Прикладной регрессионный анализ“ и „Applied regression analysis“, которые читаются студентам ФГБОУ ВО „Новосибирский государственный технический университет“ и ФГАОУ ВО «Новосибирский национальный исследовательский государственный университет» соответственно.

Публикации. Основные результаты диссертации опубликованы в девяти работах, четыре из которых в журналах из перечня ВАК. В совместных с А.П. Ковалевским автору диссертации принадлежат доказательства теорем и проведение расчетов, его соавтору интерпретация полученных результатов.

Структура и объем диссертации. Диссертация состоит из введения, 2 глав, заключения, списка литературы, а также одного приложения с десятью рисунками. Общий объем диссертации составляет 102 страниц машинописного текста. Библиография содержит 86 наименований, в том числе 9 работ автора по теме диссертации (приведены в конце списка литературы ([78]-[86]).

Глава 1

Предельные теоремы для эмпирического моста, возникающего в линейных регрессионных моделях на порядковые статистики

1.1 Исторический экскурс в регрессионный анализ

Регрессионные модели имеют широкие и далекоидущие практические применения. Так, например, они оказываются весьма полезными при описании временных рядов (см., например, в [1]). Они также используются в инвестиционном анализе (см. [17]), при решении ряда эконометрических, биометрических и других прикладных задач. Поэтому вопрос анализа адекватности регрессионных моделей является весьма актуальным.

Регрессия возникает всякий раз, когда стоит задача исследования и описание влияния одних количественных переменных на другие. Благо, что в современном мире информации и технологического прогресса нет недостатка в измерительных данных. На любом предприятии, будь то промышленный завод, финансово-кредитное учреждение или торговая компания, существует масса измерительных приборов, позволяющих получать огромные массивы эмпирических данных. И для принятия коммерческих, управленческих и других решений оказывается весьма полезным понимание зависимости между наблюдаемыми величинами.

Конечно, в некоторых случаях в качестве зависимости может выступать простая функциональная зависимость, но во многих приложениях, особенно при наблюдении физических процессов, это скорее исключение, чем правило. Функциональная связь может быть очень сложной или вообще не поддаваться выражению в элементарных терминах. В таком случае можно попытаться ввести некоторую аппроксимационную связь, например, линейную, построив линейную регрессионную модель, или относительно простую нелинейную зави-

симость.

Иногда также полезно строить регрессионные связи и между величинами, существование физической или иной связи между которыми априори неочевидно. Это может помочь открыть новый закон или по крайней мере позволит делать хоть какие-то суждения и предсказания о поведении одних величин через другие. Может показаться, что данный подход является неестественным и надуманным, однако это, пожалуй, единственный выход при принятии решений в ситуациях полной неопределенности и отсутствия каких-либо иных релевантных данных.

Термин „регрессия“ был впервые введен английским исследователем Френсисом Гальтоном (1822–1911) в конце XIX века в естественно-научных работах. Изначально Гальтон употреблял термин „реверсия“ в своих работах, что означает обращение, движение вспять. Позднее термин „регрессия“ появился в президентском адресе, прочитанном перед секцией H Британской ассоциации в Абердине в 1885г. и опубликованном в журнале „Nature“ в сентябре 1885г. (см. стр. 24 в [19]), а также в статье „Регрессия к середине в наследовании роста“ (см. [54]).

Гальтон провел в 1886–1889 годах серию измерительных опытов, в том числе им были изучены 205 пар родителей и 930 человек их взрослых детей. В итоге проведенных исследований Гальтон опубликовал ряд статей, в которых им был сформулирован «Закон регрессии к среднему». Исследователь обнаружил, что дети родителей с высоким или низким ростом обычно не наследуют нестандартный рост и назвал этот феномен „регрессия к посредственности“. Из почтения к выдающемуся ученому отметим также, что Гальтоном был введен термин „корреляционный анализ“. И именно Гальтон первым понял, что коэффициент корреляции — это мера зависимости между переменными (см. стр. 25 в [19]). Таким образом, Фрэнсис Гальтон внес вклад, который сложно переоценить, во многие области науки, но регрессия и корреляция — его важнейший вклад в статистику.

Но вернемся к регрессии. Из вышесказанного следует, что термин „регрессия“ использовался первоначально исключительно в естественно-научном смысле. Позднее, после работ английского математика Карла Пирсона (1857–1936), термин „регрессия“ стал все чаще ассоциироваться с математической статистикой. Пирсон, если можно так выразиться, добавил регрессии математичности, усовершенствовав предложенные Гальтоном методы (см. стр. 11 в [36]). В частности, Пирсоном был предложен термин „множественная регрессия“ для описания связи между несколькими независимыми переменными (называемыми также регрессорами, предикторами или факторами) и зависимой переменной (называемой также откликом).

Однако основателем регрессионного анализа по праву все же принято считать выдающегося немецкого математика Карла Фридриха Гаусса (1777–1855), так как именно Карл Гаусс (и независимо от него Адриен Мари Лежандр (1752–1833)) заложил основы важного метода, используемого в регрессионном анализе, МНК (метод наименьших квадратов). Между двумя выдающимися учеными (Гауссом и Лежандром) даже возникла ссора по поводу первооткрытия метода. Аргументы обоих ученых, а также история вопроса, тщательно изучены и обсуждены в работе [68].

Суть МНК состоит в минимизации суммы квадратов отклонений наблюдаемых переменных от регрессионной кривой путем выбора значений параметров регрессионной модели. В результате получаются оценки параметров регрессии, минимизирующие сумму квадратов отклонений и оптимальные в этом смысле. Так как отклонения могут быть в обе стороны, то есть как положительные, так и отрицательные, предлагается перед суммированием возводить их в квадрат.

На сегодняшней день в статистической науке изучается масса различных регрессионных зависимостей: построенные по детерминированному или случайному регрессору, одномерные и многомерные, непараметрические и параметрические (в том числе многопараметрические модели), линейные и нелинейные и другие.

Наиболее простая (однопараметрическая) модель линейной регрессии была нами приведена во введении (1).

Наличие случайного члена в модели обусловлено рядом вполне естественных причин. Во-первых, это невключение в модель каких-то важных и необходимых объясняющих переменных. Это может может быть вызвано, помимо простого упущения по незнанию или недопониманию происходящего процесса, например, невозможностью или трудоемкостью их измерения. Так, в частности, в экономической науке при измерении и прогнозировании различных макроэкономических показателей большую роль играют ожидания (психологические факторы) тех или иных субъектов экономических отношений, измерение которых либо невозможно, либо существенно затруднено. Наличие и влияние этих факторов приводит к отклонению наблюдаемого значения отклика от регрессионной кривой.

Во-вторых, нами может быть выбрана неправильная структура модели. Приведем пример. При проведении денежно-кредитной политики финансовые органы государства (как правило, центральные банки) опираются на данные макроэкономической статистики, но ни для кого не секрет, что любые действия в этой области сопровождаются эффектом запаздывания (так называемые временные лаги). И при построении регрессионной модели правильнее опираться не на показатели текущего периода, а на показатели более ранних периодов, причем „дальность“ этих периодов сильно зависит от экономических настроений, поэтому получить однозначный ответа на вопрос какие данные использовать весьма затруднительно. Таким образом, получаемая модель является смещенной относительно временного периода, что и отражается в наличии случайного члена.

В-третьих, агрегирование переменных. Опять же поясним на примере. В экономике часто изучаются совокупные переменные. Например, совокупный спрос, совокупное предложение, совокупное потребление и т.д. Сами же эти показатели являются результатом агрегирования поведения хозяйствующих субъектов,

которое сильно различается в зависимости от их вида. В частности, потребление отдельных групп домохозяйств зависит от различных их предпочтений и привычек, например, религиозных. Таким образом, совокупное потребление пытается собрать воедино все эти соотношения, которые, очевидно, имеют разные параметры, в результате максимум, на что мы можем рассчитывать — это достаточно точное приближение.

В-четвертых, неправильная спецификация. Истинная зависимость может быть вовсе нелинейной. Конечно, в случае сильной нелинейности лучше вообще отказаться от линейной регрессии. Но если зависимость близка к линейной, то можно получить приемлемые результаты, используя линейную модель, но за это и приходится платить случайным членом.

Наконец, в-пятых, это ошибки измерений. Этому как минимум две причины. К сожалению, в любом измерительном приборе заложена некоторая погрешность измерений, как в плане неточностей в работе (раскалибровка, разбалансировка), так и в плане заложенного предела точности измерений. Вторая причина более тонкая. В силу известного принципа неопределенности, открытого и опубликованного Вернером Гейзенбергом в 1927 г. в [58], случайность в измерениях на уровне квантовых систем не только неизбежна, но и вполне естественна.

Вернемся к регрессионным моделям. Примером немного более сложной модели может являться двухпараметрическая линейная регрессионная модель вида (2), в которой также предполагается пропорциональность регрессора и отклика, но уже с некоторым постоянным сдвигом в виде неизвестного параметра a. При замене параметров на их оценки $\hat{a}$ и $\hat{b}$ модель приобретает вид $\hat{Y} = \hat{a}+\hat{b}X$ и имеет простую интерпретацию. При изменении объясняющей переменной X на единицу отклик Y изменяется на $\hat{b}$ единиц, конечно, с учетом единиц измерения соответствующих переменных. Параметр $\hat{a}$, в свою очередь, дает прогнозируемое значение отклика при значении $X = 0$, хотя в конкретной ситуации это может и не иметь ясного смысла.

Наконец, приведем общий вид линейной регрессионной модели

$$Y_i^{(n)} = \sum_{j=1}^{m} \theta_j X_{ij}^{(n)} + \varepsilon_i^{(n)},\ i = 1, \ldots, n,$$

или в матричной форме записи

$$Y^{(n)} = X^{(n)}\theta + \varepsilon.$$

В настоящей работе применяются и исследуются исключительно линейные регрессионные модели, поэтому приведем только простой пример нелинейной регрессии для полноты изложения

$$Y_{ni} = f(\theta, X_{ni}) + \varepsilon_i, \quad i = 1, \ldots, n,$$

где $f(x, y)$ — некоторая нелинейная функция.

Вернемся к модели (1). Неизвестный параметр регрессионной модели θ обычно оценивают с помощью метода наименьших квадратов (МНК-оценка параметра), получая оценку $\widehat{\theta} = \overline{XY}/\overline{X^2}$, где через $\overline{V} = \frac{1}{n}\sum_{i=1}^{n} V_i$ обозначено выборочное среднее случайных величин $V_1, \ldots, V_n$. На основании регрессионной модели строятся прогнозные значения $\widehat{Y}_{ni} = \widehat{\theta} X_{ni}$. Остатками регрессии называют случайные величины $\widehat{\varepsilon}_{ni} = Y_{ni} - \widehat{Y}_{ni}$. Как было отмечено выше, оценка метода наименьших квадратов строится из соображений минимизации суммы квадратов регрессионных остатков, то есть $\widehat{\theta}$ является решением оптимизационной задачи поиска минимума

$$(\widehat{\varepsilon}_{n1}^2 + \widehat{\varepsilon}_{n2}^2 + \ldots + \widehat{\varepsilon}_{nn}^2) \to \min_{\theta \in \mathbf{R}}.$$

Помимо метода наименьших квадратов существует множество других методов оценки параметров линейной регрессионной модели. Отметим здесь лишь немногие из них: знаковый метод (см. [4]), метод наименьших модулей (см. [32]),

обобщенный метод наименьших модулей (см. [37]). Однако несмотря на многообразие вариантов именно метод наименьших квадратов получил наибольшее распространение и применение. Связано это, конечно, как с его простотой и естественностью, так и со свойствами, которыми обладают получаемые МНК-оценки.

При выполнении известных условий Гаусса—Маркова (см. стр. 80–81 в [18]) оценки метода наименьших квадратов будут состоятельными, несмещенными и эффективными оценками в классе всех линейных несмещённых оценок (в англоязычной литературе употребляют аббревиатуру BLUE (Best Linear Unbiased Estimator). Дополнительно к условиям Гаусса—Маркова часто еще требуют нормальность ошибок регрессии, данное условие оказывается очень полезным при проверке многих гипотез. Отметим также, что в случае выполнения последнего условия (так называемая нормальная регрессия) МНК-оценка совпадает с оценкой максимального правдоподобия.

Отметим попутно, что оценки параметров случае авторегрессионного процесса, в том числе в случае зависимых помех, широко рассмотрены в [25]–[27] (см. также ссылки в них). Авторегрессии широко используются при анализе валютных курсов (об этом мы скажем чуть более подробно в параграфе 2.2 главы 2 настоящей диссертации).

Следующим этапом в изучении регрессионных моделей является проверка статистических гипотез. Задачу проверки гипотез можно рассматривать с двух дополняющих друг друга ракурсов. Во-первых, мы можем сначала выдвинуть определенную гипотезу, а потом проводить эксперимент (или серию экспериментов) с целью подтверждения или опровержения выдвинутой гипотезы (анализ статистической значимости выдвинутой гипотезы). Во-вторых, мы можем по результатам уже проведенного эксперимента анализировать, какие из теоретических гипотез согласуются с результатами проведенного эксперимента. Понятно, что оба подхода на практике применяются в зависимости от конкретной прикладной задачи и потребностей исследователя.

Общая теория проверки статистических гипотез подробно и доступно изложена во многих учебниках по математической статистике (см., например, соответствующие параграфы в [6] и [29]), поэтому мы поговорим лишь об определенных специфических вопросах проверки гипотез, свойственных регрессионному анализу.

При проверке статистических гипотез в регрессионном анализе в качестве нулевой гипотезы обычно выступает простая гипотеза о равенстве коэффициента (вектора коэффициентов) регрессионной модели некоторому значению (некоторому вектору значений). В качестве альтернативы может выступать как сложная гипотеза (например, простое отрицание нулевой гипотезы), так и простая (множество простых) гипотеза. В качестве критического выбирается обычно пятипроцентный или однопроцентный уровень значимости.

Еще одной важной и интересной особенностью проверки гипотез в регрессионном анализе является широкое применение так называемых односторонних статистических критериев. Примером может служить односторонний вариант t-теста, используемого для проверки значимости отличия коэффициентов регрессии от того или иного значения.

Использование односторонних критериев позволяет увеличить мощность критерия относительно двустороннего аналога при любом уровне значимости. Помимо проверки гипотез о тех или иных значениях коэффициентов регрессии отдельный интерес представляет также и проверка значимости самого регрессионного уравнения (по сути наличия зависимости между регрессором и откликом). Для этого часто используется F-тест (см. стр. 109 в [18]), основанный на статистике

$$F = \frac{R^2/k}{(1-R^2)(n-k-1)}, \text{ где } R^2 = \frac{\mathbf{Var}\hat{Y}}{\mathbf{Var}Y} \text{ — коэффициент детерминации.}$$

Коэффициент детерминации R^2 иногда еще называют долей объясненной дисперсии.

Сам же коэффициент детерминации R^2 часто используется для сравнения двух или нескольких регрессионных моделей. Чем больше коэффициент детерминации, тем регрессионная модель лучше. Однако указанное правило работает адекватно только при условии, что функциональные зависимости в регрессионных моделях подобны. Так, например, сравнение линейной регрессионной модели с логарифмической на основании коэффициента детерминации и суммы квадратов регрессионных остатков не будет корректным из-за разности масштабов самой переменной Y и ее логарифма $\ln Y$.

В данном случае на помощь приходит тест, разработанный Полом Зарембкой в [77]. Тест предлагает перемасштабирование наблюдений путем нормирования их средним геометрическим значением. По перемасштабированным значениям Y строятся линейная и логарифмическая модели, которые уже являются сравнимыми как с помощью коэффициента детерминации, так и с помощью суммы квадратов регрессионных остатков.

Отметим также, что тест Зарембки является частным случаем более общей процедуры, известной под названием теста Бокса—Кокса, предложенной соответственно Боксом и Коксом в [48]. Бокс и Кокс заметили, что функции y и $\ln y$ являются частными случаями функции $\frac{y^\lambda - 1}{\lambda}$ (данное преобразование называется преобразованием Бокса-Кокса) при $\lambda = 1$ и $\lambda \to 0$ соответственно. Процедура, предложенная Боксом и Коксом, состоит в подборе оптимального значения параметра λ из соображений минимизации суммы квадратов регрессионных остатков, описанную процедуру еще называют решетчатым поиском.

Также для сравнения различных видов форм нелинейных моделей и обнаружения ошибок в спецификации регрессии предложено несколько тестов, например, RESET тест Рамсея (см. [72]) и тест Хаусмана (см. [57]). Данные (и другие) тесты основаны либо на преобразованиях случайных ошибок, либо на преобразовании масштаба зависимой переменной, и подробно на них мы останавливаться не будем.

На этом краткий обзор разделов регрессионного анализа, относящихся к

материалам настоящей диссертации, можно считать законченным.

1.2 Предварительные сведения теории случайных процессов

Отвлечемся ненадолго от эмпирического моста и поговорим о теории случайных процессов вообще. Теория случайных процессов в силу своей практической и теоретической важности широко изучена и описана в литературе. Желающие могут получить представление о ней, например, в книгах [9], [34], [15], [11]. Указанный список, конечно, не претендует на полноту, а приводится лишь как один из вариантов возможного набора литературы для первого ознакомления с указанной темой. Поэтому, как и в случае с регрессионным анализом, мы ограничимся лишь рассмотрением фрагментов теории случайных процессов, непосредственно относящихся и необходимых для понимания материалов настоящей диссертации.

Большинство задач статистики, таких как проверка гипотез, построение критических областей, основано на предельных соотношениях для выборочных статистик, то есть на приближении распределений вероятностей одних процессов другими. Предельные теоремы об асимптотических распределениях в теории вероятностей и математической статистике во многом опираются на теорию слабой сходимости вероятностных мер в метрических пространствах (см., например, [3]), развитием которой в свое время занимались А. Н. Колмогоров, Дж. Дуб, М. Донскер, Ю.В. Прохоров, А. В. Скороход, А. А. Боровков и другие. В связи с этим возникает острая потребность в методах и приемах отыскания и описания предельных процессов.

Всюду далее мы будем рассматривать случайные процессы, заданные в пространстве непрерывных на отрезке $[0,1]$ функций $C(0,1)$, снабженном равномерной метрикой, а именно в пространстве задано расстояние и норма (равно-

мерная норма или супремум-норма) в следующем виде соответственно

$$\rho(f,g) = \sup_t |f(t) - g(t)|, \ \|h\|_\infty = \sup_t |h(t)|.$$

Попутно отметим важные свойства пространства $C(0,1)$: пространство $C(0,1)$ полно и сепарабельно.

Для дальнейших рассуждений нам понадобится понятие относительной компактности семейства вероятностных мер: семейство вероятностных мер относительно компактно, если любая подпоследовательность его элементов содержит слабо сходящуюся последовательность. Легко показать, что при условии относительной компактности семейства вероятностных мер и сходимости их конечномерных распределений будет иметь место просто слабая сходимость данного семейства вероятностных мер.

Таким образом, мы получили замечательный метод доказательства предельных теорем для случайных процессов, а именно для доказательства сходимости последовательности случайных процессов к некоторому предельному процессу достаточно показать относительную компактность этой последовательности случайных процессов, а также сходимость конечномерных распределений (за более подробными рассуждениями читатель может обратиться к §6 в [3]). В данном случае остается только одна небольшая сложность: проверка относительной компактности семейства мер или последовательности случайных процессов является сама по себе довольно непростой задачей. Поэтому для применения описанного метода доказательства нам необходим эффективный способ проверки (критерий) относительной компактности.

Фундамент для такого критерия, известный под названием теорема Прохорова был предложен нашим соотечественником русским ученым Ю.В. Прохоровым в 1956 году (см. [33]). Для формулирования теоремы Прохорова нам понадобиться еще одно понятие — плотность семейства вероятностных мер.

Будем говорить, что семейство вероятностных мер Π плотно на некотором

метрическом пространстве S, если для каждого положительного ε найдется такое компактное множество K, что $P(K) > 1-\varepsilon$ для любой меры P из семейства П.

Теперь мы можем сформулировать теорему Прохорова, которая гласит: если семейство вероятностных мер плотно, то оно относительно компактно, при этом если метрическое пространство, на котором задано семейство, сепарабельно и полно, то верно и обратное — относительная компактность влечет плотность семейства мер.

Вернемся к пространству $C(0,1)$, как было отмечено выше, оно сепарабельно и полно, а значит, в силу теоремы Прохорова, относительная компактность и плотность семейства вероятностных мер в $C(0,1)$ эквивалентны. А вот для проверки плотности в $C(0,1)$ семейства распределений уже существуют эффективные критерии проверки (желающие могут ознакомиться с ними в §8 в [3]). Одним из таких критериев плотности мы и воспользуемся ниже при доказательстве теорем.

Одним из примеров применения описанной техники является доказательство одного из фундаментальных фактов теории сходимости вероятностных мер и случайных процессов широко известного принципа инвариантности Донскера—Прохорова (или, как еще говорят, функциональная центральная предельная теорема). Принцип состоит в том, что если мы по случайному блужданию $\{S_n\}$ построим случайный процесс путем линейной интерполяции, то при достаточно общих условиях такой процесс будет сходиться к винеровскому процессу (о котором мы поговорим чуть позже). Более подробно: пусть даны случайные величины $\xi_1, \xi_2 \ldots$, заданные на вероятностном пространстве $(\Omega, B, \mathbf{P})$, по которым построено случайное $\{S_n\}$. Зададим при каждом $\omega \in \Omega$ функцию $W_n(t,\omega), t \in [0,1]$ как непрерывную кусочно-линейную ломаную следующим образом:

$$W_n(t,\omega) = \frac{1}{\sigma\sqrt{n}} S_{[nt]}(\omega) + (nt - [nt])\frac{1}{\sigma\sqrt{n}}\xi_{[nt]+1}(\omega).$$

Тогда при некоторых дополнительных предположениях о случайных величинах $\xi_1, \xi_2 \ldots$ будет иметь место слабая сходимость W_n к винеровскому процессу W, то есть $W_n \Longrightarrow W$. Здесь и далее на протяжении всей работы через $\Longrightarrow$ будем обозначать слабую сходимость (сходимость по распределению) в соответствующем пространстве. Так, в частности, здесь через $\Longrightarrow$ обозначена слабая сходимость в пространстве $C(0,1)$, снабженном равномерной метрикой.

Донскер доказал принцип инвариантности в случае независимых, одинаково распределенных случайных величин $\xi_1, \xi_2 \ldots$ с нулевым средним и конечной ненулевой дисперсией. Прохоров же обобщил результат Донскера, ослабив условия и доказав принцип инвариантности в предположении выполнения условия Линдеберга.

Пришло время поговорить о винеровском процессе. Часто в литературе его еще называют процессом броуновского движения. Не приводя строго математического определения винеровского процесса, которое есть в большинстве учебников по случайным процессам, попробуем представить его сущность. Для этого обратимся к исследованиям Р. Броуна, который в 1827 году открыл движение пыльцы в жидкости. Броун установил, что частица пыльцы, помещенная в жидкость, совершает хаотические движения во все стороны. Движение вызвано беспорядочным столкновением частицы с молекулами жидкости, и для описания указанного движения частицы как раз и используется винеровский процесс, отсюда и второе его наименование — броуновское движение.

Теперь мы можем дать простую и наглядную интерпретацию принципа инвариантности: если частица подвергается частым независимым смещениям, то она будет совершать приближенно броуновское движение. Однако более важной является другая интерпретация принципа инвариантности: предельное распределение любого непрерывного преобразования от W_n инвариантно относительно распределения независимых величин $\xi_1, \xi_2 \ldots$, от которых по большому счету требуется лишь выполнение условия Линдеберга.

Броуновское движение является частным случаем класса случайных про-

цессов — гауссовских случайных процессов. Процесс называется гауссовским, если конечномерные его распределения нормальны. Гауссовские процессы образуют класс процессов, которые имеют широкое применение в естествознании и технике, например при описании различных помех и шумов в радиофизике (в частности, широко известен гауссовский процесс — белый шум).

Другим полезным гауссовским процессом является связанное броуновское движение или броуновский мост W^0. Значение броуновского моста в момент времени t выражается через значение винеровского процесса в следующем виде: $W_t^0 = W_t - tW_1$. Слово „мост“ появилось, вероятно, в силу равенства $W_0^0 = W_1^0 = 0$ с вероятностью единица. Отметим, что в термине „эмпирический мост“, введенном выше, слово „мост“ возникает по аналогичным причинам.

1.3 Основные результаты работы

Рассмотрим однопараметрическую модель линейной регрессии на порядковые статистики:

$$Y_{ni} = \theta X_{ni} + \varepsilon_i, \quad i = 1, \ldots, n, \tag{4}$$

где $X_{ni} = \xi_{i:n},\ i = 1, \ldots, n,$ — порядковые статистики, построенные по случайным величинам $\xi_1, \ldots, \xi_n$, $\theta \in \mathbf{R}$, $\varepsilon_1, \ldots, \varepsilon_n$ — независимые, одинаково распределенные случайные величины с нулевым математическим ожиданием и конечной ненулевой дисперсией σ^2. Случайные величины $\xi_1, \ldots, \xi_n$ предполагаются независимыми, одинаково распределенными с функцией распределения F и не зависящими от случайных величин $\varepsilon_1, \ldots, \varepsilon_n$.

Оценим неизвестный параметр регрессии по методу наименьших квадратов $\widehat{\theta} = \overline{XY}/\overline{X^2}$, где через $\overline{V} = \frac{1}{n}\sum_{i=1}^n V_i$ обозначено выборочное среднее случайных величин $V_1, \ldots, V_n$. Подставив оценку параметра в модель, получим прогнозные значения $\widehat{Y}_{ni} = \widehat{\theta} X_{ni}$, а затем и остатки регрессии $\widehat{\varepsilon}_{ni} = Y_{ni} - \widehat{Y}_{ni}$.

Для дальнейших построений и формулирования основных результатов диссертационной работы, нам понадобиться ряд понятий и объектов, а также ряд

связанных с ними фактов.

Пусть $GL_n(t)$ — случайная кусочно-линейная кривая, построенная по точкам $(k/n, S_k/n)$, где $S_k = \sum_{i=1}^{k} \xi_{i:n}$ — частичные суммы порядковых статистик, $k = 1, \ldots, n,\ S_0 = 0$. Эта случайная ломаная называется *эмпирической обобщенной кривой Лоренца* (см. в [55]). Применем к $GL_n(t)$ преобразование, с помощью которого из броуновского движения был получен броуновский мост, и обозначим через $GL_n^0(t) = GL_n(t) - tGL_n(1)$ случайную ломаную, построенную соответственно по точкам

$$\left(\frac{k}{n}, \frac{nS_k - kS_n}{n^2}\right).$$

В работе [56] была показана сходимость эмпирической обобщенной кривой Лоренца к ее теоретическому аналогу $\|GL_n - GL_F\|_\infty \to 0$ п.н., где $GL_F(t) = \int_0^t F^{-1}(s)\, ds$ – *теоретическая обобщенная кривая Лоренца* (см. в [55]), $\|h\|_\infty = \sup_x |h(x)|$ – супремум-норма в пространстве непрерывных на отрезке [0,1] функций $C(0,1)$ заданной функции $h(x)$, $F^{-1}(s) = \sup\{x :\ F(x) < s\}$ — квантильное преобразование (обобщенная обратная функция) функции распределения $F(x)$.

Тогда, в силу непрерывности в равномерной метрике в пространстве $C(0,1)$ преобразования, отображающего x в x^0, где $x^0(t) = x(t) - tx(1)$ (преобразование броуновского моста), получаем (см. §5 в [3]), что $\|GL_n^0 - GL_F^0\|_\infty \to 0$ п.н, где $GL_F^0(t) = GL_F(t) - tGL_F(1)$.

На протяжении всей работы, через $\Longrightarrow$ мы будем обозначать слабую сходимость в соответствующем метрическом пространстве. Так, в частности, в теоремах 1-3 ниже через $\Longrightarrow$ обозначена слабая сходимость в пространстве непрерывных на [0,1] функций $C(0,1)$, снабженном равномерной метрикой (см. [3], стр. 82).

Напомним здесь данное во введение определение эмпирического моста. Эмпирический мост — это кусочно-линейная случайная ломаная $\widehat{Z}_n = \{\widehat{Z}_n(t),\ 0 \leqslant$

$t \leqslant 1\}$ с узлами в точках

$$\left(\frac{k}{n}, \ \frac{\widehat{\Delta}_{nk} - \frac{k}{n}\widehat{\Delta}_{nn}}{\sqrt{\widehat{\sigma^2} n}} \right),$$

где $\widehat{\Delta}_{nk} = \widehat{\varepsilon}_{n1} + \ldots + \widehat{\varepsilon}_{nk},\ k = 0, 1, \ldots, n,\ \widehat{\Delta}_{n0} = 0,\ \widehat{\sigma^2} = \overline{\widehat{\varepsilon}^2} - (\overline{\widehat{\varepsilon}})^2$.

Теперь мы готовы пошагово сформулировать предлагаемый в настоящей диссертации

Алгоритм анализа линейный регрессионных моделей:

Шаг 1. С помощью метода наименьших квадратов оцениваются все входящие в регрессионную модель параметры.

Шаг 2. С помощью полученных на предыдущем шаге оценок рассчитываются регрессионные остатки модели по формуле $\widehat{\varepsilon}_{ni} = \ Y_{ni} - \widehat{Y}_{ni}$.

Шаг 3. По полученным на предыдущем шаге значениям остатков модели оценивается выборочная дисперсия модели $\widehat{\sigma^2}$.

Шаг 4. По полученным на Шаге 2 регрессионным остаткам рассчитываются значения эмпирического моста в узловых точках.

Шаг 5. Подбирается такой функционал, предельное распределение которого от эмпирического моста известно или табулировано.

Шаг 6. Рассчитывается значение выбранного на предыдущем шаге функционала от эмпирического моста.

Шаг 7. Если значение функционала превышает свое пороговое значение (недостаточное значение РДУЗ, эталонное значение которого задается априори из соображений желаемой гарантированной надежности применяемого алгоритма), то гипотеза об адекватности регрессионной модели отклоняется как несогласующаяся с наблюдаемыми данными, в противном случае гипотеза принимается.

Конечно, мы только что привели самое общее описание алгоритма. В конкретных практических задачах в зависимости от их специфики могут быть проведены дополнительные шаги. В частности, если мы хотим использовать для Шагов 6-7 статистики типа омега-квадрат, необходимо провести проверку

регрессора на нормальность. В главе 2 на реальных задачах будет продемонстрировано как и какие шаги дополнительно необходимо совершать.

Как отмечалось во введении, для применения указанного алгоритма необходимо знать предельное поведение эмпирического моста, что эквивалентно в указанном во введении смысле отысканию предельного процесса для кривой Z_n, построенной по точкам

$$\left(\frac{k}{n},\ \frac{\widehat{\Delta}_k}{\sigma\sqrt{n}}\right).$$

Итак мы готовы сформулировать предельную теорему для эмпирического моста, позволяющего реализовать построенный алгоритм в случае однопараметрической модели линейной регрессии на порядковые статистики. Доказательство теоремы будет проведено в параграфе 1.4 настоящей главы.

Теорема 1 Если $0 < \mathbf{E}\xi_1^2 < \infty$, то справедливы следующие утверждения:

1) $Z_n \Longrightarrow Z_F$, где Z_F — центрированный гауссовский процесс с ковариационной функцией

$$K_F(t,s) = \min\{t,s\} - \frac{GL_F(s)GL_F(t)}{\mathbf{E}\xi_1^2},\ s,t \in [0,1];$$

2) $\widehat{Z}_n \Longrightarrow\ Z_F^0$, где Z_F^0 — центрированный гауссовский процесс с ковариационной функцией

$$K_F^0(t,s) = \min\{t,s\} - ts - \frac{GL_F^0(s)GL_F^0(t)}{\mathbf{E}\xi_1^2},\ s,t \in [0,1].$$

Рассмотрим теперь двухпараметрическую линейную регрессионную модель: $\{Y_{ni},\ 1 \leqslant i \leqslant n,\ n \geqslant 1\}$

$$Y_{ni} = a + bX_{ni} + \varepsilon_i,\ i = 1,\ldots,n, \tag{5}$$

где, как и выше, $\{X_{ni},\ 1 \leqslant i \leqslant n,\ n \geqslant 1\}$ порядковые статистики, построенные по выборке $\xi_1,\ldots,\xi_n,\ n \geqslant 1$, с распределением F, независящей от регрессион-

ных остатков, то есть $X_{ni} = \xi_{i:n}$.

Для неизвестных параметров a и b также обычно используются оценки наименьших квадратов (или, как еще их называют, оценки Гаусса-Маркова):

$$\widehat{b}_n = \frac{\overline{XY} - \overline{X}\,\overline{Y}}{\overline{X^2} - \overline{X}^2},\ \widehat{a}_n = \overline{Y} - \widehat{b}_n\,\overline{X}.$$

Аналогично вводятся массивы прогнозных значений $\{\widehat{Y}_{ni}\}$, регрессионных остатков $\{\widehat{\varepsilon}_{ni}\}$ и их частичных сумм $\{\widehat{\Delta}_{ni}\},\ 1 \leqslant i \leqslant n,\ n \geqslant 1$:

$$\widehat{Y}_{ni} = \widehat{a}_n + \widehat{b}_n X_{ni},\ \widehat{\varepsilon}_{ni} = Y_{ni} - \widehat{Y}_{ni},\ \widehat{\Delta}_{ni} = \widehat{\varepsilon}_{n1} + \ \ldots + \ \widehat{\varepsilon}_{ni}.$$

Заметим, что, в отличии от ранее рассмотренной модели (4), в модели (5) имеет место равенство $\overline{\widehat{\varepsilon}}_n = 0$. В дальнейшем, как и при рассмотрении первой модели, в двойных индексах мы будем опускать индекс n.

Эмпирический мост $\widehat{Z}_n$ и кривая Z_n строятся аналогично по $\widehat{\Delta}_i$. Оценка дисперсии в данном случае принимает вид $\widehat{\sigma^2} = \overline{\widehat{\varepsilon}^2} - (\overline{\widehat{\varepsilon}})^2 = \overline{\widehat{\varepsilon}^2}$.

Как и в случае однопараметрической регрессионной модели, для двухпараметрической регрессионной модели справедлива предельная теорема для эмпирического моста

Теорема 2 Если $0 < \mathbf{Var}\xi_1 < \infty$, тогда $Z_n \Longrightarrow \widetilde{Z}^0_F$, $\widehat{Z}_n \Longrightarrow \widetilde{Z}^0_F$, где $\widetilde{Z}^0_F$ центрированный — центрированный гауссовский процесс с ковариационной функцией

$$\widetilde{K}^0_F(t,s) = \min\{t,s\} - ts - \frac{GL^0_F(t)GL^0_F(s)}{\mathbf{Var}\xi_1},\ t,s \in [0,1].$$

Доказательство теоремы 2 будет проведено в параграфе 1.5 настоящей главы. Кроме того, в заключении параграфа 1.5 главы 1 мы приведем другое, более „прямое“ и короткое, доказательство теоремы 1, основанное на методах параграфа 1.5 главы 1.

Внесем теперь некоторые видоизменения в модель (5), а именно, пусть регрессионные ошибки управляются марковской цепью. В англоязычной литературе для обозначения регрессионных ошибок, управляемых марковской цепью, используют термин „Markov-modulated noise“.

Чтобы определить модель, мы введем три взаимно независимых семейства случайных величин:

1) $\{\varepsilon_i^v, i \geq 1, 1 \leq v \leq M\}$ — семейство независимых случайных величин, где $\{\varepsilon_i^v, i \geq 1\}$ одинаково распределены для каждого v, $\mathbf{E}\varepsilon_1^v = 0$, $\mathbf{Var}\varepsilon_1^v = \sigma_v^2 \geq 0$ и $\sum_{v=1}^M \sigma_v^2 > 0$;

2) $\{\xi_i\}_{i=1}^\infty$ – последовательность независимых одинаково распределенных случайных величин с функцией распределения F и конечной положительной дисперсией $\mathbf{Var}\xi$;

3) $\{V_i\}_{i=1}^\infty$ — неразложимая апериодическая цепь Маркова, заданная на множестве состояний $\{1, \ldots, M\}$, со стационарным распределением $\{\pi_i\}_{i=1}^M$.

Рассмотрим модель:

$$Y_i = a + b\xi_i + \varepsilon_i^{V_i},\ n \geqslant 1, i = 1, \ldots, n.$$

Таким образом, у нас есть последовательность трехмерных вектор-строк $(Y_i, \xi_i, \varepsilon_i^{V_i})$. Упорядочивая эти векторы (для каждого n) по второй компоненте, мы получим векторы $(Y_{ni}, X_{ni}, \varepsilon_{ni}^{V_i})$ — аналоги порядковых статистик в многомерном случае. Здесь для каждого $n = 1, 2, \ldots$ $X_{ni} = \xi_{i:n}$, а величины Y_{ni}, $\varepsilon_{ni}^{V_{ni}}$ — соответствующие X_{ni} значения Y и ε^V соответственно.

В итоге мы приходим к модели:

$$Y_{ni} = a + bX_{ni} + \varepsilon_{ni}^{V_{ni}},\ n \geqslant 1, i = 1, \ldots, n. \tag{6}$$

Как и в случае модели (5), в качестве оценок неизвестных параметров вы-

ступают

$$\widehat{b}_n = \frac{\overline{XY} - \overline{X}\,\overline{Y}}{\overline{X^2} - \overline{X}^2},\ \widehat{a}_n = \overline{Y} - \widehat{b}_n\,\overline{X}$$

классические оценки Гаусса-Маркова.

Отметим, что отсутствие гомоскедастичности, которая является необходимым условием приемлемости оценок МНК, в данном случае не портит оценки, полученные с помощью метода наименьших квадратов. Это происходит в силу эргодичности цепи Маркова, которая обеспечивает гомоскедастичность в пределе.

Определения прогнозных значений, регрессионных остатков, их частичных сумм и эмпирического моста для модели (6) аналогичны введенным ранее для моделей (4) и (5). В (6) оцениваемая в определении эмпирического моста дисперсия дается формулой $\sigma^2 = \sum_{v=1}^{M} \sigma_v^2 \pi_v$.

Оказывается, что описанное обобщение модели (5) также позволяет получить предельную теорему для эмпирического моста (доказательство которой будет дано в параграфе 1.6 главы 1).

Теорема 3 Случайная ломаная Z_n и эмпирический мост $\widehat{Z}_n$ слабо сходятся при $n \to \infty$ к центрированному Гауссовскому процессу $\widetilde{Z}_F^0$ с ковариационной функцией

$$\widetilde{K}_F^0(t,s) = \min\{t,s\} - ts - \frac{GL_F^0(t)GL_F^0(s)}{\mathbf{Var}\xi_1},\ t,s \in [0,1].$$

1.4 Модель однопараметрической линейной регрессии на порядковые статистики (доказательство теоремы 1)

Доказательство теоремы 1 будет основано на ряде вспомогательных утверждений (леммы 1-5). Доказательства же вспомогательных лемм, в свою очередь, опираются на известные утверждения об асимптотическом поведении порядковых статистик и равномерной интегрируемости случайных величин. Приведем эти утверждения в качестве предложений без подробных доказательств в силу их общеизвестности.

Введем следующие обозначения $\xi^+ = \max(0, \xi)$ и $\xi^- = \min(0, \xi)$. Тогда $\xi = \xi^+ + \xi^-$.

Предложение 1 Справедливы следующие утверждения:

1) (см. задачу 21.16 в [28]) Пусть $\xi_n \Longrightarrow \xi$ при $n \to \infty$. Тогда следующие утверждения эквивалентны:

а) последовательность $\{\xi_n\}$ равномерно интегрируема.

б) $\mathbf{E}|\xi_n| < \infty$, $\mathbf{E}|\xi| < \infty$ и при $n \to \infty$ $\mathbf{E}\xi_n^+ \to \mathbf{E}\xi^+, \mathbf{E}\xi_n^- \to \mathbf{E}\xi^-$;

2) (см. теорему 5 на стр. 235 в [41]) Пусть $0 \leqslant \xi_n \to \xi$ п.н. и $\mathbf{E}\xi_n < \infty$. Тогда $\mathbf{E}\xi_n \to \mathbf{E}\xi < \infty$ тогда и только тогда, когда семейство случайных величин $\{\xi_n\}_{n \geqslant 1}$ равномерно интегрируемо.

Предложение 2 Пусть $\xi_1, \xi_2, \ldots$ – независимые и одинаково распределенные случайные величины, $S_n = \xi_1 + \cdots + \xi_n$. Тогда из сходимости $\frac{S_n}{n} \to a$ п.н. следует, что $\mathbf{E}|\xi_1| < \infty$, а последовательность $\{\frac{S_n}{n}\}$ равномерно интегрируема.

Предложение 2 следует из УЗБЧ Колмогорова, предложения 1 и соотношений $S_n = \xi_1^+ + \cdots + \xi_n^+ - (-\xi_1^- - \cdots - \xi_n^-)$, $\frac{\xi_1^+ + \cdots + \xi_n^+}{n} \to c = \mathbf{E}\xi^+ < \infty$, $\mathbf{E}\left(\frac{\xi_1^+ + \cdots + \xi_n^+}{n}\right) = \mathbf{E}\xi^+$.

Предложение 3 Пусть $\{\xi_{i:n}\}$ – порядковые статистики, построенные по выборке $\{\xi_1, \ldots \xi_n\}$ с конечным математическим ожиданием, $0 \leqslant c_1 < c_2 \leqslant 1$. Тогда последовательности $\{\frac{\xi_{n:n}}{n}\}$ и $\left\{\frac{\sum_{i=[nc_1]+1}^{[nc_2]} \xi_{i:n}}{n}\right\}$ равномерно интегрируемы.

Предложение 3 следует из неравенств $|\xi_{n:n}| \leq \sum_{i=1}^n |\xi_i|$, $\sum_{i=[nc_1]+1}^{[nc_2]} |\xi_{i:n}| \leq$

$\sum_{i=1}^{n} |\xi_i|$ и предложения 2.

Предложение 4 Если $\mathbf{E}|\xi_1| < \infty$, то имеют место следующие сходимости $\frac{\xi_{n:n}}{n} \to 0$ п.н., $\frac{\mathbf{E}\xi_{n:n}}{n} \to 0$.

Предложение 4 является широко известным фактом, который легко доказывается с помощью леммы Бореля—Кантелли (см. лемму 1 на стр. 327 в [41]) и предложения 3.

Подставляя $\xi_i = |\eta_i|^\alpha$ в предложение 4, получаем

Предложение 5 Если $\mathbf{E}|\eta_1|^\alpha < \infty,\ \alpha > 0$, то $\frac{|\eta_{n:n}|^\alpha}{n} \to 0$ п.н., $\frac{\mathbf{E}|\eta_{n:n}|^\alpha}{n} \to 0$, $\frac{\eta_{n:n}}{n^{1/\alpha}} \to 0$ п.н., $\frac{\eta_{i:n}}{n^{1/\alpha}} \to 0$ п.н. равномерно по $1 \leqslant i \leqslant n$.

Применяя неравенство Йенсена к последнему утверждению, получаем

Предложение 6 Если $\mathbf{E}|\eta_1|^\alpha < \infty,\ \alpha \geqslant 1$, то $\mathbf{E}\eta_{i:n} = o(n^{1/\alpha})$ равномерно по $1 \leqslant\ i \leqslant n$.

Предложение 7 (см. теорему 1 в [59]) Пусть $\xi_1, \xi_2, ...$ – независимые и одинаково распределенные случайные величины, $\mathbf{E}|\xi_1| < \infty$ и $g(x)$ – вещественнозначная, непрерывная функция такая, что $|g(x)| \leqslant h(x)$, где функция $h(x)$ выпуклая и $\mathbf{E}h(\xi_1) < \infty$. Тогда

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} g(\mathbf{E}\xi_{i:n}) = \mathbf{E}g(\xi_1).$$

Выбирая в предложении 7 в качестве функции $g(x) = x^2$ и вспоминая, что $\frac{1}{n}\sum_{i=1}^{n} \mathbf{E}\xi_{i:n}^2 = \mathbf{E}\xi_1^2$, получаем

Предложение 8 Пусть $\xi_1, \xi_2, ...$ – независимые и одинаково распределенные случайные величины, $\mathbf{E}\xi_1^2 < \infty$, тогда

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} \mathbf{Var}\xi_{i:n} = 0.$$

Теперь сформулируем и докажем леммы 1-5.

Лемма 1 Пусть $0 \leqslant s < t \leqslant 1$. Если $\mathbf{E}\xi_1^2 < \infty$, то

$$\frac{1}{\sqrt{n}} \left(\sum_{i=[ns]+1}^{[nt]} \xi_{i:n}\varepsilon_i - \sum_{i=[ns]+1}^{[nt]} \varepsilon_i \mathbf{E}\xi_{i:n} \right) \xrightarrow{p} 0.$$

Доказательство.

Для начала заметим, что, в силу существования математического ожидания $\mathbf{E}\xi_1$, моменты $\mathbf{E}\xi_{i:n}$ также существуют (см. [20], стр. 40). Далее, в силу неравенства Чебышева, достаточно доказать, что $\frac{1}{n}\mathbf{Var}\left(\sum\limits_{i=[ns]}^{[nt]} (\xi_{i:n} - \mathbf{E}\xi_{i:n})\varepsilon_i \right) \to 0$ при $n \to \infty$.

В силу того, что случайные величины $\{\varepsilon_i\}$ независимы и не зависят от $\{\xi_{i:n}\}$, а $\mathbf{E}\varepsilon_1 = 0$, дисперсия суммы $\mathbf{Var}\left(\sum\limits_{i=[ns]}^{[nt]} (\xi_{i:n} - \mathbf{E}\xi_{i:n})\varepsilon_i \right)$ равна сумме дисперсий, и

$$\frac{1}{n}\mathbf{Var}\left(\sum_{i=[ns]}^{[nt]} (\xi_{i:n} - \mathbf{E}\xi_{i:n})\varepsilon_i \right) = \frac{1}{n} \sum_{i=[ns]}^{[nt]} \mathbf{Var}(\xi_{i:n} - \mathbf{E}\xi_{i:n})\mathbf{Var}\varepsilon_i$$

$$= \frac{\sigma^2}{n} \sum_{i=[ns]}^{[nt]} \mathbf{Var}\xi_{i:n} \leqslant \frac{\sigma^2}{n} \sum_{i=1}^{n} \mathbf{Var}\xi_{i:n}.$$

Последнее отношение стремится к нулю с ростом n в силу предложения 8.

Таким образом, лемма 1 доказана. Лемма 1 позволила нам избавиться от зависимости в суммах, заменив порядковые статистики, порождающие зависимость, на их средние значения.

Сходимость конечномерных распределений процесса $\widehat{Z_n}$ будет доказана с помощью следующей леммы.

Лемма 2 Если $0 < \mathbf{E}\xi_1^2 < \infty$, то для любых $0 = t_0 < t_1 < \ldots < t_k = 1$, $k \geqslant 1$, случайный вектор

$$\frac{1}{\sigma\sqrt{n}} \left(\sum_{i=[nt_{j-1}]+1}^{[nt_j]} \varepsilon_i,\ j = 1, \ldots, k,\ \sum_{i=1}^{n} \varepsilon_i \xi_{i:n} \right) \tag{7}$$

слабо сходится к $(k+1)$-мерному нормальному вектору $(\zeta_1, \ldots, \zeta_{k+1})$ с нулевым математическим ожиданием и ковариациями

$$\mathbf{E}\zeta_i\zeta_j = \begin{cases} 0, & i \neq j,\ i < k+1,\ j < k+1; \\ t_i - t_{i-1}, & i = j < k+1; \\ GL_F(t_i) - GL_F(t_{i-1}), & i < k+1, j = k+1; \\ \mathbf{E}\xi_1^2, & i = j = k+1. \end{cases}$$

Доказательство.

Заменяя в (7) случайные величины $\xi_{i:n}$ на их математические ожидания $a_{in} = \mathbf{E}\xi_{i:n}$, получаем случайный вектор

$$\frac{1}{\sigma\sqrt{n}} \left(\sum_{i=[nt_{j-1}]+1}^{[nt_j]} \varepsilon_i,\ j = 1, \ldots, k,\ \sum_{i=1}^{n} \varepsilon_i a_{in} \right). \tag{8}$$

В силу леммы 1 разность между векторами (7) и (8) сходится по вероятности к нулю. Следовательно, достаточно показать, что вектор (8) сходится по распределению к вектору ζ.

Обозначим через $\mathbf{i}$ мнимую единицу. Характеристическая функция случайного вектора (8) равна

$$\varphi(u_1, \ldots, u_k, u_{k+1}) = \prod_{j=1}^{k} \mathbf{E} \exp\left(\frac{\mathbf{i}}{\sigma\sqrt{n}} \left(u_j \sum_{i=[nt_{j-1}]+1}^{[nt_j]} \varepsilon_i + u_{k+1} \sum_{i=1}^{n} \varepsilon_i a_{in} \right) \right)$$

$$= \prod_{j=1}^{k} \prod_{i=[nt_{j-1}]+1}^{[nt_j]} \mathbf{E} \exp\left(\frac{\mathbf{i}}{\sigma\sqrt{n}} \left((u_j + u_{k+1}a_{in})\varepsilon_i \right) \right).$$

Справедливо следующее разложение характеристической функции в ряд Тейлора

$$\mathbf{E} \exp \frac{\mathbf{i}t\varepsilon_i}{\sigma\sqrt{n}} = 1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right).$$

Обозначим через $\alpha_{in} = \frac{(u_j + u_{k+1}a_{in})^2}{2n}$. Так как, в силу предложения 6, $a_{in} =$

$o(\sqrt{n})$ равномерно по $1 \leqslant i \leqslant n$, то $\alpha_{in} = o(1)$ равномерно по $1 \leqslant i \leqslant n$, и

$$\varphi(u_1, \ldots, u_{k+1}) \to \prod_{j=1}^{k} \lim_{n\to\infty} \prod_{i=[nt_{j-1}]+1}^{[nt_j]} (1 - \alpha_{in})$$

$$= \prod_{j=1}^{k} \lim_{n\to\infty} \exp\left\{ \sum_{i=[nt_{j-1}]+1}^{[nt_j]} \ln(1 - \alpha_{in}) \right\}.$$

Так как $-x(1 + o(1)) \leqslant \ln(1 - x) \leqslant -x$ при $x \to 0$, то для любого $\varepsilon > 0$ найдется n_0 такое, что для всех $n \geqslant n_0$ выполнено неравенство $-\alpha_{in}(1 + \varepsilon) \leqslant \ln(1 - \alpha_{in}) \leqslant -\alpha_{in}$. В силу последнего соотношения $\lim\limits_{n\to\infty} \sum\limits_{i=[nt_{j-1}]+1}^{[nt_j]} \ln(1 - \alpha_{in}) = \lim\limits_{n\to\infty} \sum\limits_{i=[nt_{j-1}]+1}^{[nt_j]} \alpha_{in}$. Покажем существование последнего предела.

$$\lim_{n\to\infty} \sum_{i=[nt_{j-1}]+1}^{[nt_j]} \alpha_{in} = \frac{1}{2} u_{k+1}^2 \lim_{n\to\infty} \sum_{i=1}^{n} \frac{a_{in}^2}{n} + \frac{1}{2}(t_j - t_{j-1}) u_j^2$$

$$+ u_j u_{k+1} \lim_{n\to\infty} \sum_{i=[nt_{j-1}]+1}^{[nt_j]} \frac{a_{in}}{n}.$$

Так как $\frac{1}{n} \sum_{i=1}^{n} \mathbf{E}\xi_{i:n}^2 = \mathbf{E}\xi_1^2$, то

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} a_{in}^2 = \mathbf{E}\xi_1^2 - \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbf{Var}\xi_{i:n} = \mathbf{E}\xi_1^2$$

в силу предложения 8.

Так как (см. в [56])

$$\eta_n = \frac{1}{n} \sum_{i=[nt_{j-1}]+1}^{[nt_j]} \xi_{i:n} \to GL_F(t_j) - GL_F(t_{j-1}) \text{ п.н.},$$

то для сходимости

$$\frac{1}{n}\sum_{i=[nt_{j-1}]+1}^{[nt_j]} a_{in} \to GL_F(t_j) - GL_F(t_{j-1})$$

достаточно равномерной интегрируемости последовательности случайных величин $\{\eta_n\}$, которая имеет место в силу предложения 3.

Таким образом, существование $\lim\limits_{n\to\infty}\sum\limits_{i=[nt_{j-1}]+1}^{[nt_j]} \alpha_{in}$ доказано.

Итак,

$$\varphi(u_1,\ldots,u_k,u_{k+1}) \to \exp\left(-\frac{1}{2}u_{k+1}^2\mathbf{E}\xi_1^2\right)$$

$$\times\prod_{j=1}^{k}\exp\left(-\frac{1}{2}(t_j-t_{j-1})u_j^2 - u_ju_{k+1}(GL_F(t_j)-GL_F(t_{j-1}))\right).$$

Следовательно, вектор (8) сходится по распределению к вектору ζ. Лемма доказана.

Докажем теперь сходимость конечномерных распределений случайного процесса Z_n к соответствующим конечномерным распределениям предельного процесса.

Лемма 3 Если $0 < \mathbf{E}\xi_1^2 < \infty$, то конечномерные распределения случайного процесса Z_n сходятся к конечномерным распределениям процесса Z_F.

Доказательство.

Остатки регрессии $\hat{\varepsilon}_i$ допускают следующее представление:

$$\hat{\varepsilon}_i = Y_i - \hat{\theta}\xi_{i:n} = \theta\xi_{i:n} + \varepsilon_i - \frac{\overline{X(\theta X+\varepsilon)}}{\overline{X^2}}X_i = \varepsilon_i - \frac{\xi_{i:n}}{\sqrt{n}}\frac{\sum_{i=1}^n \xi_{i:n}\varepsilon_i}{\sqrt{n}}\frac{1}{\overline{\xi^2}}. \qquad (9)$$

Вследствие предыдущей леммы сумма $\frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_{i:n}\varepsilon_i$ сходится по распределению к нормальнораспределенной случайной величине с нулевым математическим ожиданием и дисперсией $\sigma^2\mathbf{E}\xi_1^2$, а сходимость $\overline{\xi^2} \to \mathbf{E}\xi_1^2$ п.н. имеет место в силу УЗБЧ.

Вследствие предложения 5 случайные величины $(\hat{\varepsilon}_i - \varepsilon_i)$ сходятся по веро-

ятности к нулю равномерно по $1 \leqslant i \leqslant n$.

Так как

$$Z_n(t) = \frac{\widehat{\Delta}_{[nt]}}{\sigma\sqrt{n}} + \left(t - \frac{[nt]}{n}\right)\frac{\widehat{\varepsilon}_{[nt]+1}}{\sigma\sqrt{n}},$$

то

$$\left|Z_n(t) - \frac{\widehat{\Delta}_{[nt]}}{\sigma\sqrt{n}}\right| \leqslant \left|\frac{\widehat{\varepsilon}_{[nt]+1}}{\sigma\sqrt{n}}\right| \leqslant \left|\frac{\varepsilon_{[nt]+1}}{\sigma\sqrt{n}}\right| + \left|\frac{\widehat{\varepsilon}_{[nt]+1} - \varepsilon_{[nt]+1}}{\sigma\sqrt{n}}\right| \xrightarrow{p} 0.$$

Следовательно, для любых $0 = t_0 < t_1 < \ldots < t_{k-1} < t_k = 1$ слабая сходимость вектора $(Z_n(t_1), \ldots, Z_n(t_k))$ эквивалентна слабой сходимости вектора

$$\frac{1}{\sigma\sqrt{n}}\left(\widehat{\Delta}_{[nt_1]}, \ldots, \widehat{\Delta}_{[nt_k]}\right).$$

В силу (9), имеет место представление

$$\frac{\widehat{\Delta}_{[nt]}}{\sigma\sqrt{n}} = \frac{\Delta_{[nt]}}{\sigma\sqrt{n}} - \frac{\sum_{i=1}^{n}\varepsilon_i\xi_{i:n}}{\sigma\sqrt{n}}\frac{S_{[nt]}}{n\overline{\xi^2}}. \tag{10}$$

Так как $\frac{S_{[nt]}}{n} \to GL_F(t)$ п.н., $\overline{\xi^2} \to \mathbf{E}\xi_1^2$ п.н., то согласно лемме 2

$$\frac{1}{\sigma\sqrt{n}}\left(\widehat{\Delta}_{[nt_1]}, \ldots, \widehat{\Delta}_{[nt_k]}\right) \Longrightarrow \left(\sum_{i=1}^{j}\zeta_i - \zeta_{k+1}\frac{GL_F(t_j)}{\mathbf{E}\xi_1^2},\ j = 1, \ldots, k\right).$$

Итак, конечномерные распределения слабо сходятся к гауссовскому случайному вектору с нулевым вектором математических ожиданий и ковариациями ($t_1 < t_2$):

$$\mathbf{cov}(Z_F(t_1), Z_F(t_2)) = \mathbf{E}\left(\zeta_1 - \zeta_{k+1}\frac{GL_F(t_1)}{\mathbf{E}\xi_1^2}\right)\left(\zeta_1 + \zeta_2 - \zeta_{k+1}\frac{GL_F(t_2)}{\mathbf{E}\xi_1^2}\right)$$

$$= \mathbf{E}\zeta_1^2 - \frac{1}{\mathbf{E}\xi_1^2}\Big(GL_F(t_1)\mathbf{E}(\zeta_1 + \zeta_2)\zeta_{k+1} + GL_F(t_2)\mathbf{E}\zeta_1\zeta_{k+1}\Big)$$

$$+\frac{GL_F(t_1)GL_F(t_2)}{(\mathbf{E}\xi_1^2)^2}\mathbf{E}\zeta_{k+1}^2 = t_1 - \frac{GL_F(t_1)GL_F(t_2)}{\mathbf{E}\xi_1^2}.$$

Лемма доказана.

Докажем теперь относительную компактность семейства распределений, используя теорему Прохорова.

Лемма 4 Пусть $0 < \mathbf{E}\xi_1^2 < \infty$, тогда семейство распределений $\{Z_n(t), 0 \leqslant t \leqslant 1\}$ относительно компактно.

Доказательство.

Заметим, что, так как пространство $C(0,1)$ сепарабельно и полно, то, в силу теоремы Прохорова (см. гл. 1 §6 в [3]), относительная компактность и плотность семейства распределений в нашем случае эквивалентны. Таким образом, для доказательства утверждения леммы достаточно показать плотность семейства распределений, порожденного последовательностью (10). Обозначим $k = [nt]$. На самом деле достаточно показать только плотность семейства распределений, порожденного вторым слагаемым, то есть

$$\frac{\sqrt{n}(\widehat{\theta} - \theta)}{\sigma} \frac{S_k}{n},$$

так как плотность семейства распределений, порожденного последовательностью $\frac{\Delta_k}{n^{1/2}\sigma}$, была показана в [3] при доказательстве принципа инвариантности Донскера—Прохорова.

Для доказательства плотности семейства, порожденного (10), достаточно показать (см. теорему 8.3 в [3]), что для произвольных $\varepsilon > 0, \eta > 0$ найдутся $0 < \delta < 1,\ n_0 \in \mathbb{N}$ такие, что

$$\frac{1}{\delta}\mathbb{P}\left\{\sup_{t \leqslant s \leqslant t+\delta} \left| \frac{\sqrt{n}(\widehat{\theta} - \theta)}{\sigma} \frac{S_{[ns]} - S_{[nt]}}{n} \right| \geqslant \varepsilon \right\} \leqslant \eta \tag{11}$$

для всех $n > n_0,\ 0 \leqslant t \leqslant 1$.

Отметим, что только что сформулированное утверждение является одним из эффективных критерием проверки плотности семейства в пространстве $C(0,1)$, о которых мы говорили во введении.

Докажем соотношение (11). Заметим, что в силу леммы 2

$$\frac{\sqrt{n}(\widehat{\theta}-\theta)}{\sigma} \Longrightarrow \frac{\zeta}{\sqrt{\mathbf{E}\xi_1^2}},$$

и (см. [56])

$$\sup_{t\leqslant s\leqslant t+\delta}\left|\frac{S_{[ns]}-S_{[nt]}}{n}\right| \to \sup_{t\leqslant s\leqslant t+\delta}|GL_F(s)-GL_F(t)| \text{ п.н.},$$

где ζ — случайная величина, имеющая стандартное нормальное распределение, $GL_F(x)$, как и выше, – обобщенная кривая Лоренца.

Обозначим $A(t,\delta)=\sup\limits_{t\leqslant s\leqslant t+\delta}|GL_F(s)-GL_F(t)|$. Тогда достаточно показать, что для произвольных $\varepsilon>0,\ \eta>0$ существует $0<\delta<1$ такое, что

$$\mathbb{P}\left\{\frac{|\zeta|}{\sqrt{\mathbf{E}\xi_1^2}}A(t,\delta)\geqslant\varepsilon\right\}\leqslant\frac{\eta\delta}{2},\ 0\leqslant t\leqslant 1. \tag{12}$$

В силу леммы 2 в [38] на с. 192, левая часть в (12) не превосходит величины $2\frac{e^{-x^2/2}}{x\sqrt{2\pi}}$, где $x=\frac{\varepsilon\sqrt{\mathbf{E}\xi_1^2}}{A(t,\delta)}$. Согласно неравенству Коши—Буняковского,

$$A(t,\delta)=\sup_{t\leqslant s\leqslant t+\delta}\left|\int_t^s F^{-1}(x)dx\right|\leqslant\sup_{t\leqslant s\leqslant t+\delta}\int_t^s|F^{-1}(x)|dx\leqslant\sqrt{\delta\mathbf{E}\xi_1^2}.$$

Ясно, что требуемое δ всегда найдется. Лемма доказана.

Нам осталось доказать лишь состоятельность оценки дисперсии.

Лемма 5 Пусть $\widehat{\sigma^2}=\overline{\widehat{\varepsilon}^2}-(\overline{\widehat{\varepsilon}})^2$. Если $\mathbf{E}\xi_1^2<\infty$, то $\widehat{\sigma^2}\xrightarrow{\mathbf{p}}\sigma^2$.

Доказательство.

Так как $(\widehat{\theta}-\theta)\xrightarrow{\mathbf{p}}0$ вследствие леммы 2, то $\overline{\widehat{\varepsilon}}=(\theta-\widehat{\theta})\overline{X}+\overline{\varepsilon}\xrightarrow{\mathbf{p}}0$, $\overline{\widehat{\varepsilon}^2}=\overline{\varepsilon^2}-\frac{\left(\overline{X\varepsilon}\right)^2}{\overline{X^2}}\xrightarrow{\mathbf{p}}\sigma^2$.

Лемма доказана.

Итак, мы сформулировали и доказали все необходимые вспопмогательные утверждения. Теперь доказательство теоремы 1 будет простым следствием из

них.

Заключение пункта 1 теоремы 1 является следствием теоремы 8.1 в [3] на стр. 82, леммы 3 (сходимость конечномерных распределений) и леммы 4 (относительная компактность).

В силу непрерывности в равномерной метрике в $C(0,1)$ преобразования, отображающего x в x^0, где $x^0(t) = x(t) - tx(1)$, и леммы 5, из пункта 1 теоремы 1 вытекает пункт 2.

Теорема 1 доказана.

1.5 Модель двухпараметрической линейной регрессии на порядковые статистики (доказательство теоремы 2)

Для простоты восприятия, мы разобьем доказательство теоремы 2 на пять последовательных шагов.

На первом шаге доказательства мы покажем, что сумму $\sum\limits_{i=1}^{n} \frac{\varepsilon_i^0 X_i^0}{\sqrt{n}}$ можно заменить суммой $\sum\limits_{i=1}^{n} \frac{\varepsilon_i^0 \mathbf{E} X_i^0}{\sqrt{n}}$. На втором шаге мы докажем слабую сходимость нормированного вектора с координатами $(\widehat{\Delta}_{k_1}, \ldots, \widehat{\Delta}_{k_m})$ к нормированному вектору с координатами $(\Delta_{k_1}, \ldots, \Delta_{k_m})$, где Δ_{k_i} будут определены позднее через центрированные статистики и ошибки. Третий шаг состоит в доказательстве слабой сходимости соответствующих конечномерных распределений. На четвертом шаге будет показана относительная компактность семейства вероятностных распределений, порожденных семейством $\{Z_n(t), 0 \leqslant t \leqslant 1\}$. Последний, пятый шаг, состоит из доказательства сходимости по вероятности выборочной дисперсии $\widehat{\sigma^2}$ к истинной дисперсии σ^2.

Шаг 1 Обозначим $X_i^0 = X_i - \overline{X}$, $\varepsilon_i^0 = \varepsilon_i - \overline{\varepsilon}$. Тогда

$$\widehat{\Delta}_k = \sum_{i=1}^{k} \left(\varepsilon_i^0 - \frac{\overline{X^0 \varepsilon^0}}{\overline{(X^0)^2}} X_i^0 \right). \tag{13}$$

Покажем, что

$$\frac{1}{\sqrt{n}} \left(\sum_{i=1}^{n} \varepsilon_i^0 X_i^0 - \sum_{i=1}^{n} \varepsilon_i^0 \mathbf{E} X_i^0 \right) \xrightarrow{\mathbf{P}} 0. \tag{14}$$

Так как $\mathbf{E}\varepsilon_1 = 0$, случайные величины $\{\varepsilon_i\}$ не зависимы между собой и не зависят от $\{\xi_{i:n}\}$, имеем

$$\mathbb{P}\left\{ \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i^0 (X_i^0 - \mathbf{E} X_i^0) \right| \geqslant \delta \right\} \leqslant \frac{\mathbf{Var} \sum\limits_{i=1}^{n} \varepsilon_i^0 (X_i^0 - \mathbf{E} X_i^0)}{n\delta^2} = \frac{\sigma^2 \sum\limits_{i=1}^{n} \mathbf{Var} X_i^0}{n\delta^2}.$$

Заметим, что

$$\sum_{i=1}^{n} 2|\mathbf{cov}(X_i, \overline{X})| \leqslant \sum_{i=1}^{n} 2\sqrt{\mathbf{Var}X_i \mathbf{Var}\overline{X}} \leqslant \sum_{i=1}^{n} 2\left(\frac{1+\mathbf{Var}X_i}{2}\right)\sqrt{\frac{\mathbf{Var}X_i}{n}}$$

$$= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\sqrt{\mathbf{Var}X_i}(1+\mathbf{Var}X_i)\right)$$

$$\leqslant n^{-1/2} \sum_{i=1}^{n} \frac{1+\mathbf{Var}X_i}{2} + n^{-1/2}\left(\sum_{i=1}^{n} \mathbf{Var}X_i\right)^{3/2}.$$

В силу предложения 8 из параграфа 1.2 главы 1, $\frac{1}{n}\sum\limits_{i=1}^{n} \mathbf{Var}X_i \to 0$, при $n \to \infty$. Поэтому и $\frac{1}{n}\sum\limits_{i=1}^{n} \mathbf{Var}X_i^0 \to 0$, и, тем самым, (14) доказано.

Шаг 2 Пусть $[t]$ обозначает целую часть числа t. Для каждого m и каждого $0 \leqslant s_1 < \cdots < s_m \leqslant 1$, $k_i = [ns_i]$, мы покажем слабую сходимость, при $n \to \infty$, вектора $\vec{\eta} = \frac{1}{\sigma\sqrt{n}}(\widehat{\Delta}_{k_1}, \ldots, \widehat{\Delta}_{k_m})$ к вектору $\vec{Z_F^0} = (Z_F^0(s_1), \ldots, Z_F^0(s_m))$.

Из (13), (14) и сходимостей $\overline{(X^0)^2} \to \mathbf{Var}\xi_1$ п.н., $\frac{1}{n}\sum\limits_{i=1}^{k_i} X_i^0 \to GL_F^0(s_i)$ п.н. (см. [56]), следует, что вектор $\vec{\eta}$ может быть заменен вектором $\vec{\zeta} = \frac{1}{\sigma\sqrt{n}}(\Delta_{k_1}, \ldots, \Delta_{k_m})$, где

$$\Delta_{k_j} = \sum_{i=1}^{k_j} \varepsilon_i^0 - \frac{GL_F^0(s_j)}{\mathbf{Var}\xi_1} \sum_{i=1}^{n} \varepsilon_i^0 \mathbf{E}X_i^0 = \sum_{i=1}^{k_j} \varepsilon_i^0 - \frac{GL_F^0(s_j)}{\mathbf{Var}\xi_1} \sum_{i=1}^{n} \varepsilon_i \mathbf{E}X_i^0.$$

Шаг 3 Мы покажем слабую сходимость $\vec{\zeta} \Longrightarrow \vec{Z_F^0}$, используя метод характеристических функций. Заметим что

$$\sum_{j=1}^{m} t_j \left(\sum_{i=1}^{k_j} (\varepsilon_i - \overline{\varepsilon}) - \frac{GL_F^0(s_j)}{\mathbf{Var}\xi_1} \sum_{i=1}^{n} \varepsilon_i \mathbf{E}X_i^0\right)$$

$$= \sum_{i=1}^{n} \varepsilon_i \sum_{j=1}^{m} t_j \left(\mathbf{I}\{i \leqslant k_j\} - \frac{k_j}{n} - \frac{GL_F^0(s_j)}{\mathbf{Var}\xi_1}\mathbf{E}X_i^0\right).$$

Применяя предложение 4 из параграфа 1.2 главы 1 и используя неравенство

Гельдера, имеем $\mathbf{E}X_i^0 = o(\sqrt{n})$ равномерно по всем $1 \leqslant i \leqslant n$.

Определим $\beta_i = \sum\limits_{j=1}^{m} t_j \left(\mathbf{I}\{i \leqslant k_j\} - \frac{k_j}{n} - \frac{GL_F^0(s_j)}{\mathbf{Var}\xi_1} \mathbf{E}X_i^0 \right)$. Тогда характеристическая функция $\varphi_{\vec{\zeta}}(\vec{t}\,)$ может быть представлена как

$$\varphi_{\vec{\zeta}}(\vec{t}\,) = \prod_{i=1}^{n} \mathbf{E} \exp\left(\frac{\mathbf{i}}{\sigma\sqrt{n}} \varepsilon_i \beta_i \right) = \prod_{i=1}^{n} \left(1 - \frac{\beta_i^2}{2n}(1 + o(1)) \right)$$

$$= \exp\left(\sum_{i=1}^{n} \ln\left(1 - \frac{\beta_i^2}{2n}(1 + o(1)) \right) \right),$$

где $o(1)$ равномерно по $1 \leqslant i \leqslant n$.

Действительно,

$$\sum_{i=1}^{n} \ln\left(1 - \frac{\beta_i^2}{2n}(1 + o(1)) \right) = -(1 + o(1)) \sum_{i=1}^{n} \frac{\beta_i^2}{2n}.$$

Тогда

$$\sum_{i=1}^{n} \frac{\beta_i^2}{n} = \sum_{i=1}^{n} \frac{1}{n} \left(\sum_{j=1}^{m} t_j \left(\mathbf{I}\{i \leqslant k_j\} - \frac{k_j}{n} - \frac{GL_F^0(s_j)}{\mathbf{Var}\xi_1} \mathbf{E}X_i^0 \right) \right)^2$$

$$= \sum_{j_1=1}^{m} \sum_{j_2=1}^{m} \sum_{i=1}^{n} \frac{1}{n} t_{j_1} t_{j_2} \left(\mathbf{I}\{i \leqslant k_{j_1}\} - \frac{k_{j_1}}{n} - \frac{GL_F^0(s_{j_1})}{\mathbf{Var}\xi_1} \mathbf{E}X_i^0 \right)$$

$$\times \left(\mathbf{I}\{i \leqslant k_{j_2}\} - \frac{k_{j_2}}{n} - \frac{GL_F^0(s_{j_2})}{\mathbf{Var}\xi_1} \mathbf{E}X_i^0 \right)$$

$$\to \sum_{j_1=1}^{m} \sum_{j_2=1}^{m} t_{j_1} t_{j_2} \left(\min(s_{j_1}, s_{j_2}) - s_{j_1} s_{j_2} - \frac{GL_F^0(s_{j_1}) GL_F^0(s_{j_2})}{\mathbf{Var}\xi_1} \right).$$

Таким образом,

$$\sum_{i=1}^{n} \frac{\beta_i^2}{n} \to \sum_{j_1=1}^{m} \sum_{j_2=1}^{m} t_{j_1} t_{j_2} K_F^0(s_{j_1}, s_{j_2}) := C_F,$$

и

$$\sum_{i=1}^{n} \ln\left(1 - \frac{\beta_i^2}{2n}(1 + o(1)) \right) \to -\frac{C_F}{2}.$$

В итоге мы получили, что $\varphi_{\vec{\zeta}}(\vec{t}\,) \to \exp(-C_F/2)$. Таким образом, сходимость конечномерных распределений полностью доказана.

Шаг 4 Покажем, что семейство распределений, порожденное $\{Z_n(t), 0 \leqslant t \leqslant 1\}$, относительно компактно.

Обозначим $S_k = \sum\limits_{i=1}^{k} \xi_{i:n},\ k = 1, \ldots, n,\ S_0 = 0$.

В соответствии с теоремой Прохорова (см. гл. 1 §6 в [3]) достаточно показать, что семейство распределений случайных процессов последовательности $\left\{\frac{\widehat{\Delta}_{[nt]}}{\sigma\sqrt{n}}, 0 \leqslant t \leqslant 1\right\}$ плотно. Обозначим $k = [nt]$.

Пусть

$$\widehat{\Delta}_k^0 = \sum_{i=1}^{k} \left(\varepsilon_i - \frac{\overline{X^0\varepsilon^0}}{\overline{(X^0)^2}} X_i \right).$$

Тогда

$$\widehat{\Delta}_k = \widehat{\Delta}_k^0 - \frac{k}{n}\widehat{\Delta}_n^0,$$

Плотность семейства $\left\{\frac{\sum_{i=1}^{k}\varepsilon_i}{\sigma\sqrt{n}}, 0 \leqslant t \leqslant 1\right\}$ была показана в [3] при доказательстве принципа инвариантности Донскера-Прохорова.

Поэтому, мы докажем только плотность семейства

$$\left\{ \frac{\overline{X^0\varepsilon^0}\sqrt{n}}{\sigma\overline{(X^0)^2}} \frac{S_k}{n}, 0 \leqslant t \leqslant 1 \right\}.$$

В силу теоремы 8.3 в [3] достаточно доказать, что для любых $\varepsilon > 0,\ \alpha > 0$ найдутся $0 < \delta < 1,\ n_0 \in \mathbb{N}$ такие, что

$$\frac{1}{\delta}\mathbb{P}\left\{ \sup_{t \leqslant s \leqslant t+\delta} \left| \frac{\overline{X^0\varepsilon^0}\sqrt{n}}{\sigma\overline{(X^0)^2}} \frac{S_{[ns]} - S_{[nt]}}{n} \right| \geqslant \varepsilon \right\} \leqslant \alpha,$$

для любых $n > n_0,\ 0 \leqslant t \leqslant 1$.

Заметим, что

$$\frac{\overline{X^0\varepsilon^0}\sqrt{n}}{\sigma\overline{(X^0)^2}} \Longrightarrow \frac{\zeta}{\sqrt{\mathbf{Var}\xi_1}},$$

и (см. [56])

$$\sup_{t\leqslant s\leqslant t+\delta}\left|\frac{S_{[ns]}-S_{[nt]}}{n}\right|\to\sup_{t\leqslant s\leqslant t+\delta}|GL_F(s)-GL_F(t)|\text{ п.н..}$$

Здесь ζ — случайная величина, имеющая стандартное нормальное распределение, $GL_F(x)$ — обобщенная кривая Лоренца, которая была определена выше в параграфе 1.1.

Обозначим $A(t,\delta)=\sup\limits_{t\leqslant s\leqslant t+\delta}|GL_F(s)-GL_F(t)|$. Покажем, что для любых $\varepsilon>0,\ \alpha>0$ найдется $0<\delta<1$ такое, что

$$\mathbb{P}\left\{\frac{|\zeta|}{\sqrt{\mathbf{Var}\xi_1}}A(t,\delta)\geqslant\varepsilon\right\}\leqslant\frac{\alpha\delta}{2},\ 0\leqslant t\leqslant 1.$$

Левая часть последнего неравенства не превосходит $2\frac{e^{-x^2/2}}{x\sqrt{2\pi}}$, где $x=\frac{\varepsilon\sqrt{\mathbf{Var}\xi_1}}{A(t,\delta)}$, в силу леммы 2 (см. [38], стр. 175). С помощью неравенства Коши-Буняковского имеем

$$A(t,\delta)=\sup_{t\leqslant s\leqslant t+\delta}\left|\int_t^s F^{-1}(x)dx\right|\leqslant\sup_{t\leqslant s\leqslant t+\delta}\int_t^s|F^{-1}(x)|dx\leqslant\sqrt{\delta\mathbf{E}\xi_1^2}.$$

Очевидно, что требуемое δ всегда найдется.

Шаг 5 Осталось доказать, что $\widehat{\sigma^2}\xrightarrow{\mathbf{p}}\sigma^2$. В самом деле,

$$\overline{\widehat{\varepsilon^2}}=\frac{1}{n}\sum_{i=1}^n\left(\varepsilon_i-\overline{\varepsilon}-\frac{\overline{X^0\varepsilon^0}}{\overline{(X^0)^2}}(X_i-\overline{X})\right)^2=\overline{(\varepsilon^0)^2}-\frac{(\overline{X^0\varepsilon^0})^2}{\overline{(X^0)^2}}\xrightarrow{\mathbf{p}}\sigma^2.$$

Это завершает доказательство теоремы 2.

Теперь, как и было обещано выше, приведем другое доказательство теоремы 1. Оно основано на приемах, использованных при доказательстве теоремы 2, поэтому мы дадим лишь схему, которая также состоит из пяти последовательных шагов.

Шаг 1 Имеем

$$\widehat{\Delta}_k = \sum_{i=1}^{k} \left(\varepsilon_i - \frac{\overline{XY}}{\overline{X^2}} X_i \right). \tag{15}$$

Используя аналогичные рассуждения, что и при доказательстве теоремы 2, получаем сходимость

$$\frac{1}{\sqrt{n}} \left(\sum_{i=1}^{n} X_i \varepsilon_i - \sum_{i=1}^{n} \varepsilon_i \mathbf{E} X_i \right) \xrightarrow{p} 0. \tag{16}$$

Шаг 2 Для каждого m и любых $0 \leqslant s_1 < \cdots < s_m \leqslant 1$ мы докажем слабую сходимость случайного вектора $\vec{\eta} = \frac{1}{\sigma\sqrt{n}}(\widehat{\Delta}_{k_1}, \ldots, \widehat{\Delta}_{k_m})$, где $k_i = [ns_i]$, к вектору $\vec{Z_F} = (Z_F(s_1), \ldots, Z_F(s_m))$.

Из соотношений (15), (16) и сходимостей $\overline{X^2} \to \mathbf{E}\xi_1^2$ п.н., $\frac{1}{n}\sum\limits_{i=1}^{k_i} X_i \to GL_F(s_i)$ п.н. следует, что случайный вектор $\vec{\eta}$ может быть заменен на случайный вектор $\vec{\zeta} = \frac{1}{\sigma\sqrt{n}}(\widetilde{\Delta}_{k_1}, \ldots, \widetilde{\Delta}_{k_m})$, где

$$\widetilde{\Delta}_{k_j} = \sum_{i=1}^{k_j} \varepsilon_i - \frac{GL_F(s_j)}{\mathbf{E}\xi_1^2} \sum_{i=1}^{n} \varepsilon_i \mathbf{E} X_i.$$

Шаг 3 Как и при доказательстве теоремы 2 обозначим $\beta_i = \sum\limits_{j=1}^{m} t_j \left(\mathbf{I}\{i \leqslant k_j\} - \frac{GL_F(s_j)}{\mathbf{E}\xi_1^2} \mathbf{E} X_i \right)$, $C_F = \sum\limits_{j_1=1}^{m} \sum\limits_{j_2=1}^{m} t_{j_1} t_{j_2} K_F(s_{j_1}, s_{j_2})$ и получим требуемую сходимость $\vec{\zeta} \Longrightarrow \vec{Z_F}$.

Шаги 4 и **5** очень близки к соответствующим шагам доказательства теоремы 2.

Что и завершает другое доказательство теоремы 1.

1.6 Модель двухпараметрической линейной регрессии на порядковые статистики, в которой ошибки управляются цепью Маркова (доказательство теоремы 3)

Доказательство теоремы, как и при доказательстве теоремы 2 в предыдущем параграфе, мы разобьем на пять шагов.

Пусть $X_i^0 = X_i - \overline{X}$, $\varepsilon_i^0 = \varepsilon_i^{V_i} - \overline{\varepsilon}$, где $\overline{\varepsilon} = \sum_{i=1}^n \varepsilon_i^{V_i}$.

На первом шаге мы покажем, что сумму $\sum\limits_{i=1}^{n} \frac{\varepsilon_i^0 X_i^0}{\sqrt{n}}$ можно заменить суммой $\sum\limits_{i=1}^{n} \frac{\varepsilon_i^0 \mathbf{E} X_i^0}{\sqrt{n}}$. На втором шаге, мы докажем слабую сходимость нормированного вектора с координатами $(\widehat{\Delta}_{k_1}, \ldots, \widehat{\Delta}_{k_m})$ к нормированному вектору с координатами $(\Delta_{k_1}, \ldots, \Delta_{k_m})$, где Δ_{k_i} будут определены ниже. Затем (третий шаг) мы докажем слабую сходимость конечномерных распределений. Четвертый шаг содержит доказательство относительной компактности семейства вероятностных распределений $\{Z_n(t), 0 \leq \ t \leq 1\}$. На заключительном пятом шаге мы завершим доказательство теоремы, показав сходимость по вероятности выборочной дисперсии $\widehat{\sigma^2}$ к дисперсии σ^2.

Перейдем непосредственно к доказательству

Шаг 1

Заметим, что

$$\widehat{\Delta}_k = \sum_{i=1}^{k} \left(\varepsilon_i^0 - \frac{\overline{X^0 \varepsilon^0}}{\overline{(X^0)^2}} X_i^0 \right). \tag{17}$$

Покажем, что имеет место сходимость

$$\frac{1}{\sqrt{n}} \left(\sum_{i=1}^{n} \varepsilon_i^0 X_i^0 - \sum_{i=1}^{n} \varepsilon_i^0 \mathbf{E} X_i^0 \right) \xrightarrow{\mathbf{P}} 0. \tag{18}$$

В силу предложения 8 из параграфа 1.2 главы 1, имеем сходимость $\frac{1}{n} \sum\limits_{i=1}^{n} \mathbf{Var} X_i \to 0$, при $n \to \infty$.

Заметим что $\mathbf{Var}\overline{X} = \mathbf{Var}X_1/n$,

$$\frac{1}{n}\sum_{i,j=1}^{n}\mathbf{cov}(X_i, X_j) = \frac{1}{n}\mathbf{Var}\sum_{i=1}^{n}X_i = \mathbf{Var}X_1.$$

Так как $\sum\limits_{i=1}^{n}(X_i^0 - \mathbf{E}X_i^0) = 0$, имеем

$$\sum_{i=1}^{n}\varepsilon_{ni}^0(X_i^0 - \mathbf{E}X_i^0) = \sum_{i=1}^{n}\varepsilon_i^{V_i}(X_i^0 - \mathbf{E}X_i^0).$$

В силу неравенства Чебышева,

$$\mathbf{P}\left\{\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i^{V_i}(X_i^0 - \mathbf{E}X_i^0)\right| \geqslant \delta\right\} \leqslant \frac{\mathbf{Var}\sum\limits_{i=1}^{n}\varepsilon_i^{V_i}(X_i^0 - \mathbf{E}X_i^0)}{n\delta^2}.$$

Так как $\{\varepsilon_i^{V_i}\}$ независимы при фиксированном $\{V_i\}$ и независимы от величин $\{X_i\}$, имеем

$$\mathbf{Var}\sum_{i=1}^{n}\varepsilon_i^{V_i}(X_i^0 - \mathbf{E}X_i^0) = \sum_{i=1}^{n}\mathbf{Var}\varepsilon_i^{V_i}\mathbf{Var}\left(X_i^0 - \mathbf{E}X_i^0\right) = \sum_{i=1}^{n}\mathbf{Var}\varepsilon_i^{V_i}\mathbf{Var}X_i^0.$$

Величины $\mathbf{Var}\varepsilon_i^{V_i}$ ограничены сверху и

$$\sum_{i=1}^{n}\mathbf{Var}X_i^0 = \sum_{i=1}^{n}\mathbf{Var}X_i - 2\sum_{i=1}^{n}\mathbf{cov}(X_i, \overline{X}) + n\mathbf{Var}\overline{X}$$

$$= \sum_{i=1}^{n}\mathbf{Var}X_i - \frac{2}{n}\sum_{i,j=1}^{n}\mathbf{cov}(X_i, X_j) + \mathbf{Var}X_1$$

$$= \sum_{i=1}^{n}\mathbf{Var}X_i - \mathbf{Var}X_1 = o(n).$$

В итоге получаем

$$\frac{1}{n}\mathbf{Var}\sum_{i=1}^{n}\varepsilon_i^{V_i}(X_i^0-\mathbf{E}X_i^0)\to 0.$$

Таким образом, (18) доказано.

Шаг 2 Пусть $[t]$ обозначает целую часть t. Для каждых фиксированного m и для $0\le s_1<\cdots<s_m\le 1$, $k_i=[ns_i]$, мы покажем слабую сходимость при $n\to\infty$ случайного вектора $\vec{\eta}=\frac{1}{\sigma\sqrt{n}}(\widehat{\Delta}_{k_1},\dots,\widehat{\Delta}_{k_m})$ к вектору $\vec{\tilde{Z}}_F^0=(\tilde{Z}_F^0(s_1),\dots,\tilde{Z}_F^0(s_m))$.

Из (17), (18) и сходимостей $\overline{(X^0)^2}\to\mathbf{Var}\xi_1$ п.н., $\frac{1}{n}\sum\limits_{i=1}^{k_i}X_i^0\to GL_F^0(s_i)$ п.н. (см. в [56]), достаточно доказать слабую сходимость $\vec{\zeta}\Longrightarrow\vec{\tilde{Z}}_F^0$, где $\vec{\zeta}=\frac{1}{\sigma\sqrt{n}}(\Delta_{k_1},\dots,\Delta_{k_m})$,

$$\Delta_{k_j}=\sum_{i=1}^{k_j}\varepsilon_i^0-\frac{GL_F^0(s_j)}{\mathbf{Var}\xi_1}\sum_{i=1}^{n}\varepsilon_i^0\mathbf{E}X_i^0=\sum_{i=1}^{k_j}\varepsilon_i^0-\frac{GL_F^0(s_j)}{\mathbf{Var}\xi_1}\sum_{i=1}^{n}\varepsilon_i^{V_i}\mathbf{E}X_i^0.$$

Шаг 3 Мы докажем слабую сходимость $\vec{\zeta}\Longrightarrow\vec{\tilde{Z}}_F^0$, используя метод характеристических функций. Характеристическая функция в данном случае имеет вид

$$\varphi_{\vec{\zeta}}(\vec{t}\,)=\mathbf{E}\prod_{j=1}^{m}\exp\left(\mathbf{i}\frac{t_j\Delta_{nk_j}}{\sigma\sqrt{n}}\right).$$

Заметим, что имеет место следующее равенство

$$\sum_{j=1}^{m}t_j\left(\sum_{i=1}^{k_j}(\varepsilon_i^{V_i}-\overline{\varepsilon})-\frac{GL_F^0(s_j)}{\mathbf{Var}\xi_1}\sum_{i=1}^{n}\varepsilon_i^{V_i}\mathbf{E}X_i^0\right)$$

$$=\sum_{i=1}^{n}\varepsilon_i^{V_i}\sum_{j=1}^{m}t_j\left(\mathbf{I}\{i\le k_j\}-\frac{k_j}{n}-\frac{GL_F^0(s_j)}{\mathbf{Var}\xi_1}\mathbf{E}X_i^0\right).$$

Хорошо известно, что конечность момента $\mathbf{E}\psi_1$ влечет за собой сходимость

$\frac{\psi_{n:n}}{n} \to 0$ п.н. и в среднем для последовательности независимых одинаково распределенных случайных величин $\psi_1, \ldots, \psi_n, \ldots$. На самом деле, более общо, данная сходимость имеет место для стационарной эргодической последовательности, что является следствием субаддитивной эргодической теоремы Кингмана (см. в [61]).

Применяя этот факт и используя неравенство Гельдера, имеем $\mathbf{E}X_i^0 = o(\sqrt{n})$ равномерно по $1 \le i \le n$.

Пусть $\beta_i = \sum\limits_{j=1}^{m} t_j \left(\mathbf{I}\{i \le k_j\} - \frac{k_j}{n} - \frac{GL_F^0(s_j)}{\mathbf{Var}\xi_1} \mathbf{E}X_i^0 \right)$. Тогда $\beta_{ni}/\sqrt{n} \to 0$,

$$\sum_{i=1}^{n} \frac{\beta_{ni}^2}{n} \to C_F := \sum_{j_1=1}^{m} \sum_{j_2=1}^{m} t_{j_1} t_{j_2} K_F(s_{j_1}, s_{j_2}).$$

Так как для каждого $1 \leqslant v \leqslant M$, $t \to 0$

$$\mathbf{E}e^{\mathbf{i}t\varepsilon_{ni}^v} = \exp\left(-\frac{1}{2} t^2 \mathbf{Var}\varepsilon_{ni}^v \right) (1 + o(1)),$$

и $\mathbf{Var}\varepsilon_{ni}^{V_{ni}} \to \sigma^2$ при $i, n \to \infty$, имеем (рассуждая аналогично Шагу 1)

$$\varphi_{\vec{\zeta}}(\vec{t}\,) = \mathbf{E} \prod_{i=1}^{n} \exp\left(\mathbf{i} \frac{\varepsilon_{ni}^{V_{ni}} \beta_{ni}}{\sigma\sqrt{n}} \right)$$

$$= \exp\left(-\frac{1}{2} \sum_{i=1}^{n} \frac{\beta_{ni}^2 \mathbf{Var}\varepsilon_{ni}^{V_{ni}}}{n\sigma^2} \right) (1 + o(1)) \to \exp(-C_F/2).$$

Таким образом, сходимость конечномерных распределений доказана.

Шаг 4. Покажем, что

семейство распределений $\{Z_n(t), 0 \le t \le 1\}$ относительно компактно. (19)

Пусть $S_k = \sum\limits_{i=1}^{k} \xi_{i:n},\ k = 1, \ldots, n,\ S_0 = 0.$

В силу теоремы Прохорова в [3] достаточно показать, что семейство распре-

делений случайных процессов $\left\{\frac{\widehat{\Delta}_{[nt]}}{\sigma\sqrt{n}}, 0 \le t \le 1\right\}, n = 1, 2, \ldots$ плотно. Обозначим $k = [nt]$ и пусть

$$\widehat{\Delta}_k^0 = \sum_{i=1}^{k} \left(\varepsilon_i^{V_i} - \frac{\overline{X^0\varepsilon^0}}{\overline{(X^0)^2}} X_i \right).$$

Далее, $\widehat{\Delta}_k = \widehat{\Delta}_k^0 - \frac{k}{n}\widehat{\Delta}_n^0$.

Так как $\{\varepsilon_{ni}^v\}$ независимы и одинаково распределены для каждого v, из принципа инвариантности (см. в [7]) следует плотность семейства $\left\{\frac{\sum_{i=1}^{k}\varepsilon_i^v}{\sigma\sqrt{n}}, 0 \le t \le 1\right\}$ для каждого $v \in \{1, \ldots, M\}$. Таким образом, $\left\{\frac{\sum_{i=1}^{\lfloor nt \rfloor}\varepsilon_{ni}^{V_{ni}}}{\sigma\sqrt{n}}, 0 \leqslant t \leqslant 1\right\}$ плотно. Принцип инвариантности для этой Марковмодулированной последовательности следует из Следствия 3.9 в [64].

Поэтому, чтобы доказать (19) достаточно показать плотность семейства вероятностных распределений

$$\left\{ \frac{\overline{X^0\varepsilon^0}\sqrt{n}}{\sigma\overline{(X^0)^2}} \frac{S_k}{n}, 0 \le t \le 1 \right\}.$$

В свою очередь, в силу Теоремы 8.3 в [3], достаточно доказать, что для каждых $\varepsilon > 0,\ \alpha > 0$ найдутся $0 < \delta < 1,\ n_0 \in \mathbf{N}$ такие, что имеет место следующее неравенство

$$\frac{1}{\delta}\mathbf{P}\left\{ \sup_{t \le s \le t+\delta} \left| \frac{\overline{X^0\varepsilon^0}\sqrt{n}}{\sigma\overline{(X^0)^2}} \frac{S_{[ns]} - S_{[nt]}}{n} \right| \ge \varepsilon \right\} \le \alpha,$$

для всех $n > n_0,\ 0 \le t \le 1$.

Заметим, что $\frac{\overline{X^0\varepsilon^0}\sqrt{n}}{\sigma\overline{(X^0)^2}} \Longrightarrow \frac{\zeta}{\sqrt{\mathbf{Var}\xi_1}}$, и

$$\sup_{t \le s \le t+\delta} \left| \frac{S_{[ns]} - S_{[nt]}}{n} \right| \to \sup_{t \le s \le t+\delta} |GL_F(s) - GL_F(t)| \text{ п.н. (см. в [56])}.$$

Здесь ζ — случайная величина, имеющая стандартное нормальное распределение, и $GL_F(x)$ — обобщенная кривая Лоренца.

В силу неравенства Коши-Буняковского,

$$\sup_{t\le s\le t+\delta}|GL_F(s)-GL_F(t)|\le\sup_{t\le s\le t+\delta}\int_t^s|F^{-1}(x)|dx\le\sqrt{\delta\mathbf{E}\xi_1^2}.$$

Очевидно, что подходящее положительное δ всегда найдется.

Шаг 5. Осталось доказать сходимость выборочной дисперсии к истинной $\widehat{\sigma^2}\xrightarrow{\mathbf{P}}\sigma^2$. В самом деле, $\overline{\varepsilon}\xrightarrow{\mathbf{P}}0$, $\overline{X\varepsilon}\xrightarrow{\mathbf{P}}0$, $\overline{\varepsilon^2}\xrightarrow{\mathbf{P}}\sigma^2$, и

$$\overline{\widehat{\varepsilon}^2}=\frac{1}{n}\sum_{i=1}^n\left(\varepsilon_i^{V_i}-\overline{\varepsilon}-\frac{\overline{X^0\varepsilon^0}}{\overline{(X^0)^2}}(X_i-\overline{X})\right)^2=\overline{(\varepsilon^0)^2}-\frac{(\overline{X^0\varepsilon^0})^2}{\overline{(X^0)^2}}\xrightarrow{\mathbf{P}}\sigma^2.$$

Что завершает доказательство Теоремы 3.

1.7 Сравнение подхода с использованием эмпирического моста с F-тестом проверки гипотез

В данном параграфе, мы рассмотрим отличие подхода с использованием эмпирического моста от широко известного F-теста проверки гипотезы о регрессионной модели (см. [18], стр. 109). F-тест использует коэффициент детерминации R^2 — долю объясненной выборочной дисперсии: если эта доля достаточно велика, то модель принимается, в противном случае модель отвергается. F-тест основан на статистике

$$F = \frac{R^2/k}{(1-R^2)(n-k-1)}, \text{ где } R^2 = \frac{\mathbf{Var}\hat{Y}}{\mathbf{Var}Y} \text{ — коэффициент детерминации.}$$

Анализ же модели на основе эмпирического моста позволяет отвергнуть модель на основании анализа последовательных сумм регрессионных остатков: модель может отвергаться при сколь угодно близком к единице значении R^2. В этом смысле, анализ регрессионной модели с использованием конструкции эмпирического моста является более привлекательным чем F-тест. Для лучшего понимания приведем здесь поясняющий

Пример 1 Пусть

$$h(t) = \begin{cases} \theta t - c, & t \in [0, 1/2]; \\ \theta t + c, & t \in (1/2, 1], \end{cases}$$

и $Y_i = h(\xi_{i:n}) + \varepsilon_i$, $i = 1, \ldots, n$, где $\{\xi_{i:n}\}$ — порядковые статистики, построенные по последовательности случайных величин $\{\xi_i\}$. Последовательность $\{\xi_i\}$ состоит из независимых одинаково распределенных случайных величин, которые, в свою очередь, не зависят от последовательности регрессионных ошибок $\{\varepsilon_i\}$. Случайные величины $\{\varepsilon_i\}$, в свою очередь, независимы, одинаково распределены с нулевым математическим ожиданием и конечной ненулевой дисперсией σ^2. Пусть также $\mathbf{Var}\xi_1 > 0$. По-прежнему $\widehat{Y}_i = \widehat{\theta}\xi_{i:n}$.

Имеет место сходимость $\overline{Y} \to \theta \mathbf{E}\xi_1$ п.н. Отметим, что $GL_F(1) = \mathbf{E}\xi_1$ (функции $GL_F(t)$ и $GL_F^0(t)$ были введены в параграфе 1.1 главы 1). Выборочная дисперсия

$$\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2$$

в силу УЗБЧ сходится п. н. к величине

$$c^2 + \sigma^2 + \theta^2 \mathbf{Var}\xi_1 - 2c\theta(GL_F(1/2) - \mathbf{E}\xi_1/2) + 2c\theta(GL_F(1) - GL_F(1/2) - \mathbf{E}\xi_1/2)$$

$$= c^2 + \sigma^2 + \theta^2 \mathbf{Var}\xi_1 - 4c\theta GL_F^0(1/2).$$

Необъясненная выборочная дисперсия

$$\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2 \xrightarrow{p} c^2 + \sigma^2$$

(доказательство следует из леммы 5 параграфа 1.4 главы 1).

В силу этого коэффициент детерминации

$$R^2 \xrightarrow{p} 1 - \frac{c^2 + \sigma^2}{c^2 + \sigma^2 + \theta^2 \mathbf{Var}\xi_1 - 4c\theta GL_F^0(1/2)}$$

при $n \to \infty$, и может быть сделан сколь угодно близким к единице выбором соответствующих параметров. В то же время последовательные суммы остатков растут линейно, и $q_n \widetilde{A_n}^{-1} q_n^T \xrightarrow{p} +\infty$ (см. следствие в параграфе 1.4 главы 1) для любого положительного d.

Одним из существенных недостатков коэффициента детерминации является невозможность его использования для сравнения регрессионных моделей с разным числом входящих в них параметров. С ростом числа параметров (введением новых факторов в модель) коэффициент детерминации возрастает (по крайней мере не убывает), что делает его неинформативным при сравнении таких моделей. Предлагаемый нами подход (в отличие от использования коэф-

фициента детерминации) позволяет сравнивать модели с разным числом параметров.

Пример 1 показывает, что модель линейной регрессии может объяснять сколь угодно большую долю выборочной дисперсии, но не удовлетворять строгим требованиям на суммы остатков регрессии, предъявляемых критерием эмпирического моста. А количество параметров, а также характер зависимости от них влияют только на распределение предельного процесса, на основании которого вычисляется реально достигаемый уровень значимости. Таким образом, в частности как показывает пример 1, F-тест может принимать неправильную модель, а критерий эмпирического моста отвергать ее.

Глава 2

Сравнение и анализ прикладных линейных регрессионных моделей

2.1 Некоторые аспекты практического применения основных результатов работы

В настоящем параграфе, как и было заявлено во введении, мы поговорим о возможности и дадим некоторые замечания и рекомендации по практическому применению построенного в диссертации алгоритма (теоремы 1 и 2).

Для того, чтобы применять теорему 1 к анализу соответствия регрессионной модели исследуемым данным, необходим алгоритм оценивания неизвестной ковариационной функции и построенная на его основе статистика, распределение которой при выполнении основной гипотезы сходится к известному распределению.

Мы построим статистику, слабо сходящуюся к распределению хи-квадрат с произвольным наперед заданным числом степеней свободы d. Однако критерий, построенный на ее основе, не является состоятельным при достаточно широком классе альтернатив. Для построения состоятельного критерия будем строить критерий типа омега-квадрат, предельное распределение для которого удается вычислить в ряде частных случаев.

Формулы для ковариационной функции в формулировке теоремы 1 включают неизвестные функции — кривые Лоренца $GL_F(t)$ и $GL_F^0(t)$. При практическом применении их необходимо заменить на их эмпирические аналоги $GL_n(t)$ и $GL_n^0(t)$.

В качестве оценки ковариационной функции выберем

$$\widehat{K_F^0}(t,s) = \min\{t,s\} - ts - GL_n^0(s)GL_n^0(t)/\overline{X^2},\ s,t \in [0,1].$$

Тогда равномерно по $s,t \in [0,1]$ (см. [55]) имеет место следующая сходимость с вероятностью 1

$$|\widehat{K_F^0}(t,s) - K_F^0(t,s)| \to 0.$$

Пусть $d > 0$ — целое число. Обозначим

$$\vec{q_n} = \left(\widehat{Z}_n\left(\frac{1}{d+1}\right), \ldots, \widehat{Z}_n\left(\frac{d}{d+1}\right)\right),$$

тогда

$$\vec{q_n} \Longrightarrow \vec{q_F} = \left(Z_F^0\left(\frac{1}{d+1}\right), \ldots, Z_F^0\left(\frac{d}{d+1}\right)\right).$$

Обозначим через A ковариационную матрицу вектора $\vec{q_F}$. Если матрица A невырождена ($\det A \neq 0$), то, как известно, имеет место слабая сходимость $q_n A^{-1} q_n^T \Longrightarrow \chi_d^2$ при $n \to \infty$, где χ_d^2 — распределение хи-квадрат с d степенями свободы.

Мы можем отметить, что аналогичная сходимость будет иметь место и при замене матрицы A на ее эмпирический аналог — матрицу $\widehat{A_n} = (\widehat{a}_{ij})_{i,j=1}^d$, элементы которой получены заменой элементов матрицы A на их оценки:

$$\widehat{a}_{ij} = \min\left(\frac{i}{d+1}, \frac{j}{d+1}\right) - \frac{ij}{(d+1)^2} - \frac{GL_n^0(\frac{i}{d+1})GL_n^0(\frac{j}{d+1})}{\overline{X^2}}.$$

В самом деле, так как $\vec{q_n} \Longrightarrow \vec{q_F}$, а $\widehat{A_n} \to A$ п.н., то $q_n \widehat{A_n}^{-1} q_n^T \Longrightarrow \chi_d^2$. Таким образом, мы получаем следствие из теоремы 1 (аналогичное следствие может быть получено и из теоремы 2 с той лишь поправкой, что вместо оценки второго момента в знаменателе последнего вычитаемого будет стоять оценка дисперсии).

Следствие 1 Если $0 < \mathbf{E}\xi_1^2 < \infty$, $\det A \neq 0$, то $q_n \widehat{A_n}^{-1} q_n^T$ сходится слабо к случайной величине, имеющей распределение хи-квадрат с d степенями свободы.

Замечание Отметим, что сходимость построенной статистики к распределению хи-квадрат доказана без каких-либо дополнительных предположений о распределении F по сравнению с основной теоремой. Более того, для доказательства следствия используется только часть теоремы 1, касающаяся сходимости конечномерных распределений, показанной в лемме 3. Но критерий, основанный на предлагаемой статистике, использует лишь значения эмпирического моста в конечном числе точек и поэтому, как было указано выше, является несостоятельным. Построение состоятельных критериев наталкивается на проблему нахождения предельного распределения при справедливости основной гипотезы. Наиболее разработан вопрос о предельном распределении для критериев типа омега-квадрат.

Итак, будем использовать критерий типа омега-квадрат, основанный на статистике

$$\omega_n^2 = \int_0^1 (\widehat{Z}_n(x))^2 \, dx.$$

Эта статистика слабо сходится к $\omega_F^2 = \int_0^1 (Z_F^0(x))^2 \, dx$, где гауссовский процесс Z_F^0 определен в формулировке теоремы 1.

Будем предполагать, что функция распределения F известна с точностью до параметров сдвига и масштаба, то есть имеет вид $F(x) = G((x-a)/\sigma)$, где $a \in \mathbf{R}$, $\sigma > 0$, G — известная функция распределения с дисперсией, равной 1. В этом случае ковариационная функция гауссовского процесса полностью известна, так как

$$GL_F^0(x) = \int_0^x (a + \sigma G^{-1}(t)) \, dt - ax - \sigma x GL_G(1) = \sigma GL_G^0(x),$$

$$K_F^0(s,t) = \min(s,t) - st - GL_G^0(s) GL_G^0(t).$$

В теореме 3.2 работы [52] доказано, что если функция $A(s,t) = 2(K_F^0(s,t))^2$ имеет ограниченные частные производные 4-го порядка на единичном квадрате

за исключением, быть может, диагонали, то распределение ω_F^2 в этом случае можно вычислить приближенно, заменяя интеграл суммой

$$\omega_F^{2,m} = \frac{1}{m}\sum_{i=1}^{m}\left(Z_F^0((i-1/2)/m)\right)^2,$$

при этом дисперсия погрешности аппроксимации эквивалентна cm^{-2} при $m \to \infty$, где константа c отыскивается в явном виде. В параграфе 3.3 указанной работы предложен эффективный алгоритм вычисления распределения случайной суммы $\omega_F^{2,m}$.

Другой, классический, подход к вычислению распределения ω_F^2 состоит в нахождении собственных чисел интегрального оператора с ядром $K_F^0(s,t)$ и использования формулы Смирнова. Вычисления осуществляются для конкретных распределений. Ряд результатов такого рода для предельного распределения статистики хи-квадрат при проверке сложных гипотез получен Г. В. Мартыновым в книге [30]. Как мы увидим, эти результаты применимы к отысканию распределения ω_F^2. Для иллюстрации практического применения теоремы 1 рассмотрим два случая: нормального распределения и сдвинутого показательного распределения.

Пример 2 Пусть ξ_1 имеет нормальное распределение с параметрами a и σ^2. Тогда функция распределения ξ_1 F может быть представлена в виде $F(x) = \Phi\left(\frac{x-a}{\sigma}\right)$, где $\Phi(x)$ — функция распределения стандартного нормального закона. Обозначим $\varphi(x) = \Phi'(x)$ — плотность стандартного нормального распределения, тогда

$$GL_F^0(x) = -\varphi(\Phi^{-1}(x)),\ K_F^0(s,t) = \min(s,t) - st - \varphi(\Phi^{-1}(s))\varphi(\Phi^{-1}(t)).$$

Таким образом, мы получили ковариационную функцию вида K_1, где $\varkappa_1(t) = \varphi(\Phi^{-1}(t))$ (см. [30], стр. 34). На стр. 37 в [30] приведена формула для определителя Фредгольма $D_1(\lambda)$, а на стр. 56 формула Смирнова, которая име-

ет место при некоторых дополнительных ограничениях. Однако теорема 2 (см. [30], стр. 57) дает достаточные условия для ее справедливости, которые выполняются, в частности, для рассматриваемой корреляционной функции. Наконец, распределение функционала ω_F^2 в этом случае приведено в таблице 3 (см. [30], стр. 65).

Пример 3 Пусть ξ_1 имеет сдвинутое показательное распределение с плотностью

$$f_{\xi_1}(x) = \begin{cases} \alpha e^{-\alpha(x-\beta)}, & \text{если } x \geqslant \beta, \\ 0, & \text{если } x < \beta. \end{cases}$$

Тогда

$$GL_F^0(x) = \frac{1}{\alpha}(1-x)\ln(1-x),$$

$$K_F^0(s,t) = \min(s,t) - st - (1-s)\ln(1-s)(1-t)\ln(1-t).$$

Распределение функционала ω_F^2 в этом случае получено в [53] (см. также табл. 12 в [30], стр. 71).

Поговорим теперь о применении теоремы 2.

В [16] доказана следующая

Лемма Если $Y_i = \theta + \varepsilon_i$, то $\widehat{Z}_n \Longrightarrow W^0$, где W^0 — стандартный броуновский мост с ковариационной функцией

$$K(t,s) = \min\{t,s\} - ts, \ t,s \in [0,1].$$

Эта лемма имеет следующее

Следствие 2 Если $Y_i = \theta + \varepsilon_i$, то

$$\int_0^1 (\widehat{Z}_n(t))^2\, dt \Longrightarrow \eta,$$

где η имеет распределение ω^2, которое представлено в табл. 6.4а на с. 348 в [5] и в табл. 1 на с. 63-64 в [30].

Аналогичное следствие имеет место и из теоремы 2

Следствие 3 Если $Y_i = a + bX_i + \varepsilon_i$, X_i — порядковые статистики, построенные по выборке из нормального распределения, то

$$\int_0^1 (\widehat{Z}_n(t))^2\,dt \Longrightarrow \widehat{\eta} = \int_0^1 Z_\Phi^2(t)\,dt,$$

где $\widehat{\eta}$ имеет распределение $\widehat{\omega}^2$, которое представлено в табл. 3 на с. 65 в [30], Z_Φ — центрированный гауссовский процесс с ковариационной функцией

$$K_\Phi(t,s) = \min\{t,s\} - ts - \varphi(\Phi^{-1}(t))\varphi(\Phi^{-1}(s)),$$

где φ, Φ^{-1} — плотность и квантильная функция стандартного нормального распределения соответственно.

В [30] получены следующие выражения:

$$F_{\widehat{\eta}}(x) = 1 + \frac{1}{\pi}\sum_{k=1}^{\infty}\int_{\lambda_{2k-1}}^{\lambda_{2k}} \frac{e^{-x\lambda/2}}{(-D(\lambda))^{1/2}}\,\frac{d\lambda}{\lambda}$$

(с. 56), где λ_k — собственные числа ядра $\widehat{K}$, $\lambda_1 \ldots \lambda_{10}$ приведены на с. 37 в [30], для остальных λ_k справедлива эквивалентность $\lambda_k \sim\ ((2k-\ 1)\pi)^2,\ k >\ 10$;

$$D(\lambda) = \frac{2}{\sqrt{\lambda}}\ A(\sqrt{\lambda})\sin\frac{\sqrt{\lambda}}{2};$$

$$A(\mu) = \left(1 + \frac{\mu^2\sqrt{3}}{2\pi}\right)\cos\frac{\mu}{2} + 4\mu^3 I_1(\mu);$$

$$I_1(\mu) = \int_0^{1/2} \varkappa_1(x)\sin\mu x\,dx \int_x^{1/2} \varkappa_1(x)\cos\mu(1/2-t)\,dt;$$

$$\varkappa_1(x) = \varphi(\Phi^{-1}(x)).$$

(с. 34–37 в [30]). С помощью данных формул можно вычислить $F_{\widehat{\eta}}(x)$, результат приведен в табл. 3 в [30].

В двух следующих параграфах мы применим полученные здесь следствия для анализа конкретных регрессионных моделей, возникающих в реальных прикладных задачах.

2.2 Исследование линейных регрессионных моделей зависимости курсов американского доллара и евро с помощью конструкции эмпирического моста

В экономической теории задачи прогнозирования различных экономических показателей и переменных играют очень важную роль. В частности, важным является вопрос прогнозирования валютных курсов, в особенности после отмены системы „жесткой“ привязки (например, во времена золотого или золотовалютного стандартов) валютных пар, а также в периоды экономической нестабильности (кризисов). Для прогнозирования курсов существует множество математических моделей. Например, модель искусственных нейронных сетей (artificial neural networks - ANN см. [66], [46]), модели авторегрессий (AR модели, в частности ARCH и GARCH см. [73], [67], [42]), модели скользящего среднего, а также смешанные модели авторегрессии и скользящего среднего (ARMA, ARIMA см. [65]).

В настоящем параграфе мы проанализируем зависимость курса единой европейской валюты (евро) и американского доллара. Курсы будем сравнивать не напрямую, а через их относительные курсы к швейцарскому франку. Таким образом, мы исследуем зависимость относительного курса евро к швейцарскому франку и относительного курса американского доллара к швейцарскому франку, переходя к их логарифмам. Швейцарский франк выбирается в качестве своеобразной точки отсчета, то есть колебания его (франка) покупательной способности по отношению к чему-либо не учитываются. Специалисты говорят, что эти колебания и в самом деле невелики. Швейцарский франк традиционно относится к валютам налоговых гаваней или оффшорных зон, с нулевым уровнем инфляции.

Доллар и евро, в свою очередь, колеблются довольно сильно, и мы рассмотрим две гипотезы, касающихся взаимной зависимости этих курсовых колебаний. Первая: евро следует за долларом, то есть колебания логарифма евро повторя-

ют колебания логарифма доллара с точностью до «шума», не зависящего от колебаний доллара. Вторая: все наоборот, колебания доллара следуют за колебаниями евро.

Если в обоих случаях есть значимая корреляция между шумом и регрессором, то обе предложенные модели не годятся, в таком случае необходимо рассматривать комбинированную модель с устранением корреляционной зависимости. Если же в какой-то (или в обоих сразу) из этих моделей корреляция будет достаточно близка к нулю, то дальнейшую проверку будем производить с помощью критерия типа хи-квадрат, упорядочивая значения регрессора по неубыванию.

Смысл упорядочения состоит в следующем: проверяется предположение о линейной зависимости отклика от регрессора по всем значениям регрессора от наименьших до наибольших. Если есть отклонения от линейности, то интересно, в чем они состоят (может быть, есть тяготение к некоторому коридору, то есть наибольшие отклонения от линейности отклика будут для крайних значений регрессора). Однако, нельзя не отметить, что при упорядочении не по времени, мы теряем часть информации о курсах.

Отметим, что в отличие от следующего параграфа применение критерия типа ω^2 здесь не представляется возможным, так как, как известно, курсы валют не подчиняются нормальному закону. Поэтому единственным доступным вариантом исследования является критерий хи-квадрат, несмотря на все его недостатки, описанные в параграфе 2.1 главы 2.

В качестве исходных данных были взяты цены закрытия по курсу евро к швейцарскому франку (A_i) и курсу закрытия американского доллара к швейцарскому франку (B_i) за период с 1 января 2011 года по 1 января 2014 года (1092 наблюдения каждой пары). Данные получены с помощью соответствующего онлайн сервиса на интернет сайте российской медиагруппы РосБизнесКонсалтинг по адресу \http://export.rbc.ru.

Теперь перейдем непосредственно к описанию предлагаемых моделей зави-

симости курсов евро и доллара, для этого введем обозначения

$$\xi_i = \ln \frac{A_i}{A_{i-1}},\ \eta_i = \ln \frac{B_i}{B_{i-1}},$$

и рассмотрим две однопараметрические линейные регрессионные модели.

Первая модель:

$$Y_i^{(1)} = \theta X_i^{(1)} + \varepsilon_i, \tag{20}$$

где $X_i^{(1)} = \xi_{i:n}$ — порядковые статистики, а $Y_i^{(1)}$ — соответствующие значения η_i (индуцированные порядковые статистики, конкомитанты (см. в [51] – [45] и [49] соответственно)).

Вторая модель:

$$Y_i^{(2)} = \gamma X_i^{(2)} + \delta_i, \tag{21}$$

где $X_i^{(2)} = \eta_{i:n}$ — порядковые статистики, а $Y_i^{(2)}$ — соответствующие значения величин ξ_i.

Здесь $\theta, \gamma \in \mathbf{R}$ — неизвестные параметры регрессии, $\varepsilon_1, \ldots, \varepsilon_n$ (регрессионные ошибки первой модели) — независимые одинаково распределенные случайные величины с нулевым математическим ожиданием и конечной ненулевой дисперсией σ_1^2, $\delta_1, \ldots, \delta_n$ (регрессионные ошибки второй модели) — независимые одинаково распределенные случайные величины с нулевым математическим ожиданием и конечной ненулевой дисперсией σ_2^2.

Неизвестные параметры регрессии, как было указано ранее, оценим по методу наименьших квадратов, получая оценки $\widehat{\theta}, \widehat{\gamma}$. На основании регрессионных моделей строятся прогнозные значения $\widehat{Y}_i^{(1)} = \widehat{\theta} X_i^{(1)}$, $\widehat{Y}_i^{(2)} = \widehat{\gamma} X_i^{(2)}$, а также остатки регрессии $\widehat{\varepsilon_i} = Y_i^{(1)} - \widehat{Y}_i^{(1)}$ и $\widehat{\delta_i} = Y_i^{(2)} - \widehat{Y}_i^{(2)}$.

Основной задачей настоящего параграфа является анализ адекватности предложенных моделей зависимости исследуемым данным на основании теоремы 1 из параграфа 1.1 главы 1.

С помощью пакета MatLab для каждой из предложенных регрессионных

моделей были рассчитаны выборочные корреляции регрессора и регрессионных остатков. Корреляция в модели (20) составила $-2,1820 \cdot 10^{-5}$, а в модели (21) $8,2216 \cdot 10^{-6}$. В обоих случаях корреляция достаточно близка к нулю, что позволяет нам рассматривать далее обе модели без внесения в них каких-либо корректировок.

Теперь применим теорему 1. С помощью пакета MatLab мы оценили входящие в модели (20) и (21) параметры θ и γ, вычислили регрессионные остатки и значения эмпирического моста в узловых точках. После чего были построены графики эмпирических мостов (приведены на рис. а и рис. б).

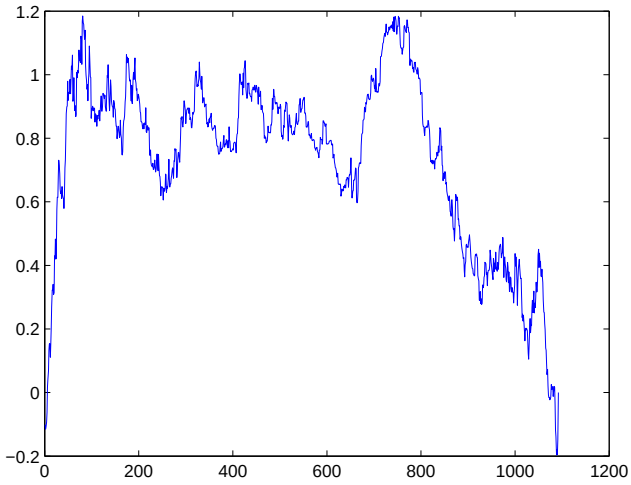


Рис. а) $Y_i^{(1)} = \theta X_i^{(1)} + \varepsilon_i$

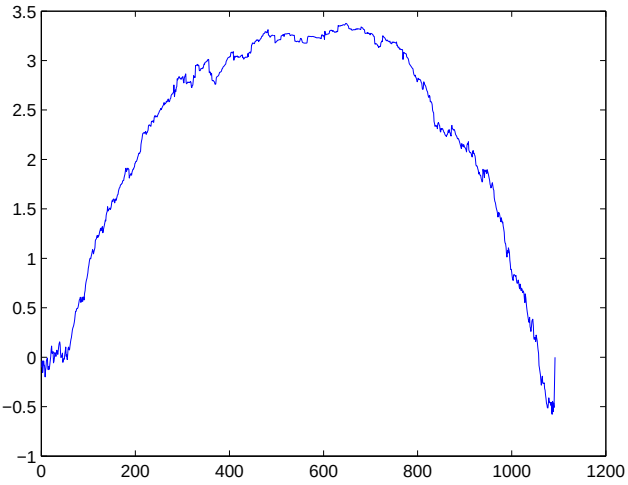


Рис. б) $Y_i^{(2)} = \gamma X_i^{(2)} + \delta_i$

Для сравнения моделей применим критерий хи-квадрат, будем использовать обозначения, введенные в параграфе 2.1.

В качестве d (число степеней свободы критерия) возьмем $d = [n]^{1/3} + 1$ и, с помощью пакета MatLab, вычислим значения статистики $q_n \widetilde{A_n}^{-1} q_n^T$ для каждой из моделей.

Для первой модели значение статистики хи-квадрат получилось равным 35,28, для второй же модели оно составило 199,6. Вторая модель решительно отвергается данными: соответствующий реально достигаемый уровень значимости чрезвычайно мал. Возможным объяснением указанного эффекта может быть тот факт, что в исследуемый период курс франка был привязан к курсу ев-

ро с коэффициентом 1,2, и зависимость курса евро от курса доллара не годится в качестве модели, далее мы ее рассматривать не будем. Напомним, что 6 сентября 2011 года на фоне „отчетливого и продолжающегося ослабления“ евро к франку Национальный банк Швейцарии под руководством Филиппа Хильдербранда ввёл так называемый “минимальный курс“ швейцарского франка к евро на уровне 1,2.

Значение статистики для первой модели также весьма велико, но анализ эмпирического моста явно показывает наличие разладки (см. рис 3) в самом начале эмпирического моста (мост очень быстро растет), поэтому мы продолжим исследование первой модели, внеся в нее некоторые корректировки. Для этого найдем момент разладки — номер наблюдения, дающего наибольшее отклонение эмпирического моста от горизонтальной оси. В данном случае разладка произошла на 81 наблюдении. В данный момент произошел большой отрицательный скачок курса евро, после чего зависимость изменилась.

Зная момент разладки, теперь мы можем разбить нашу изначальную выборку на две части и строить линейные модели на каждом участке выборки, отличающиеся входящими в них параметрами. То есть мы рассматриваем наблюдения с номерами 1-81 и 81-1092 раздельно и для каждого проделываем всю описанную выше процедуру.

На участке 1-81 значение статистики хи-квадрат получилось равным 6,23, а на участке 81-1092 хи-квадрат равно 25,36. И снова анализ эмпирического моста на участке 81-1092 позволяет выделить два подинтервала. В итоге получаем разбиение исходной выборки на три интервала 1-81, 81-771, 771-1092.

Результаты вычислений, а именно число степеней свободы, значения статистики хи-квадрат, значение оценки параметра, а также реально достигнутый уровень значимости (приблизительно), приведены в таблице ниже (табл.1). Значения реально достигнутого уровня значимости (РДУЗ) получены с помощью [5].

Таблица 1. Результаты вычислений.

Интервал	$\hat{\theta}$	d	χ^2	РДУЗ
1-81	0,64	5	6,23	0,3
81-771	1,00	9	5,2	0,8
771-1092	0,86	7	11,87	0,15

Полученный эффект имеет достаточно простую интерпретацию: курсу евро свойственны большие как отрицательные, так и положительные скачки. Первые происходят с вероятностью, которая оценивается как 81/1092, вторые же с вероятностью (1092-770)/1092=322/1092. Для этих скачков обнаруживается другая взаимосвязь с изменением курса американского доллара.

В результате исследования установлено, что вторая модель, утверждающая линейную регрессионную зависимость логарифма курса евро от логарифма курса американского доллара по отношению к швейцарскому франку, отвергается построенным статистическим критерием, а принимается исправленная первая модель: линейная регрессионная зависимость логарифма курса доллара от логарифма курса евро имеет место в трех зонах значений: в центральной зоне значений логарифмов курса евро, упорядоченных по неубыванию, имеет место пропорциональность значений с коэффициентом, близким к 1, и случайной погрешностью; в крайних зонах (выше или ниже некоторого уровня) коэффициент пропорциональности значимо меньше 1 (0,64 для низких и 0,86 для высоких значений). Построенная модель зависимости курсов позволяет учитывать риски на рынке валют для построения оптимальной стратегии хеджирования валютных рисков.

2.3 Выбор линейной регрессионной модели зависимости массы человеческого тела от его роста с помощью конструкции эмпирического моста

Во введении нами были сделаны некоторые замечания по поводу исследования регрессионных моделей зависимости человеческого роста от массы его тела. Здесь мы более подробно остановимся на известных моделях соотношения роста и массы тела человека.

Для начала отметим широко известный индекс массы тела (англ. body mass index (BMI)), который был разработан бельгийским социологом и статистиком Адольфом Кетле в 1869 году. Индекс массы тела - расчетный показатель, с помощью которого можно оценить степень соответствия массы тела человека и его роста. На основе значения индекса можно говорить о недостаточности, нормальности или избыточности массы тела, что имеет определенное значение при определении показаний к проведению лечения. Индекс массы тела рассчитывается как отношение массы, выраженной в килограммах, к квадрату роста, выраженного в метрах. В соответствии с рекомендациями Всемирной организации здравоохранения нормальное значения индекса составляет 18,5-24,99, значение менее 16 указывает на выраженный дефицит, а более 40 свидетельствует об ожирении третьей степени. Согласно исследованию, проведенному иерусалимской больницей «Адаса» совместно с Американским институтом здравоохранения идеальным для мужчин является индекс массы тела в 25—27, средняя продолжительность жизни при котором была максимальной.

Помимо индекса массы тела существуют и другие индексы соотношения массы и роста. Например, также широко известен индекс Брока, который используется при росте 155—170 см, нормальным считается значение массы, лежащее в коридоре (рост в сантиметрах минус 100) $\pm 10\%$. Менее распространенными являются индексы Брейтмана (норма массы=рост(см)*0,7-50кг.), Ноордена (норма массы=рост(см)*420/1000, Татоня (норма массы=рост-(100+(рост-

100)/20)). Однако в клинической практике они не получили широкого распространения и наиболее часто используется индекс массы тела.

Теперь применем теорему 2, лемму и следствия 2 и 3 из параграфа 2.1 главы 2 для анализа регрессионных зависимостей массы тела от роста.

В качестве исходных данных были взяты сведения о росте (в сантиметрах) и массе тела (в килограммах) студенток первого курса лечебного факультета ГБОУ ВПО «Волгоградский государственный медицинский университет» (двумерная выборка объема 750). Данные находятся в открытом доступе на интернет сайте университета по адресу \http://www.volgmed.ru/ru/

Для наглядности изобразим выборку графически (рис. 1).

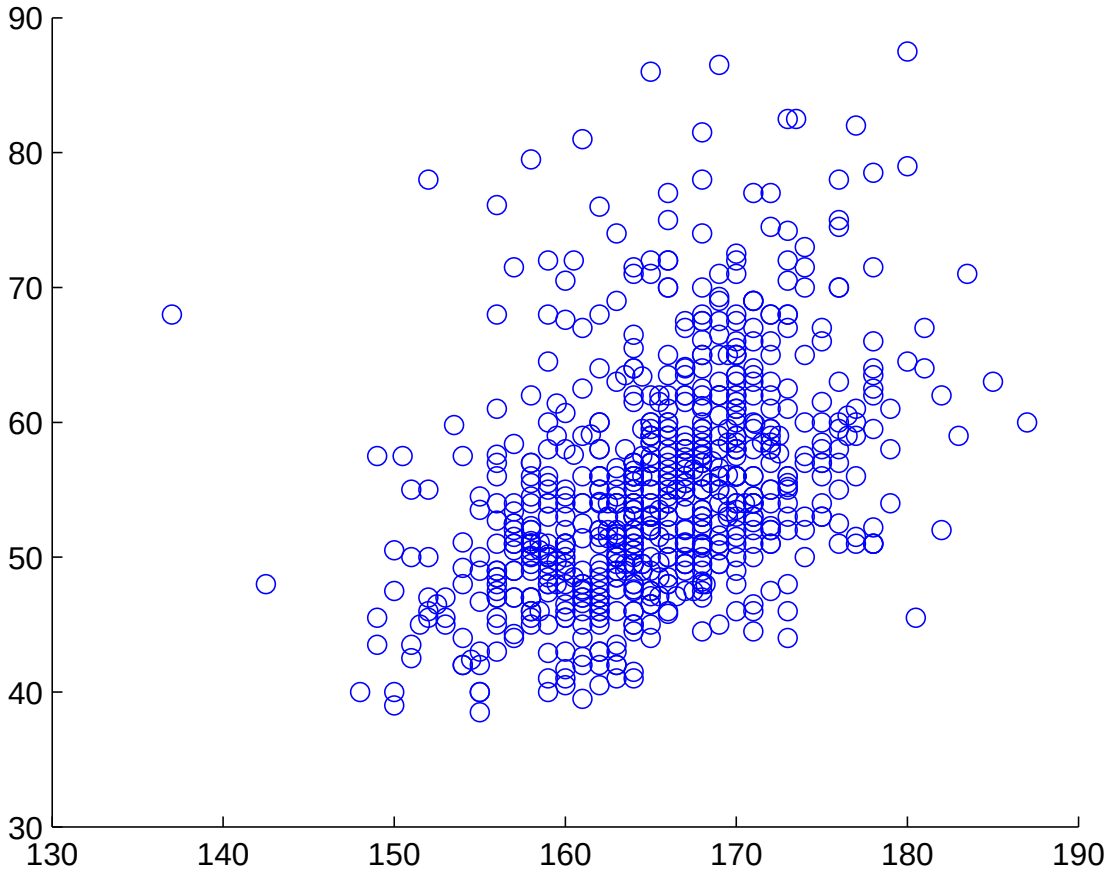


Рис. 1: Зависимость массы тела (в кг) от роста (в см)

Предполагается, что имеется некоторый статистический ансамбль наблюдений (парные наблюдения), иными словами, двумерное распределение индивидуумов по росту и весу. Из этого двумерного распределения осуществляется выборка объема n (в данном случае $n = 750$). Далее эта двумерная выборка упорядочивается по первой компоненте (росту), в результате чего получается последовательность пар (H_i, W_i), где H_i — порядковые статистики, а W_i — индуцированные порядковые статистики (или, как их еще называют в [49], конкомитанты).

Мы рассмотрим следующие регрессионные модели:

$$W_i = \theta + H_i + \varepsilon_i \tag{22}$$

$$\ln W_i = a + \ln H_i + \varepsilon_i \tag{23}$$

$$\ln W_i = a + 1.5 \ln H_i + \varepsilon_i \tag{24}$$

$$\ln W_i = a + 2 \ln H_i + \varepsilon_i \tag{25}$$

$$\ln W_i = a + 2.5 \ln H_i + \varepsilon_i \tag{26}$$

$$\ln W_i = a + 3 \ln H_i + \varepsilon_i \tag{27}$$

$$\ln W_i = a + b \ln H_i + \varepsilon_i \tag{28}$$

$$W_i = a + bH_i + \varepsilon_i \tag{29}$$

$$W_i = a + bH_i^{1.5} + \varepsilon_i \tag{30}$$

$$W_i = a + bH_i^2 + \varepsilon_i \tag{31}$$

$$W_i = a + bH_i^{2.5} + \varepsilon_i \tag{32}$$

$$W_i = a + bH_i^3 + \varepsilon_i \tag{33}$$

Модели (22) – (27) являются однопараметрическими, а модели (28) – (33) двухпараметрическими.

С помощью свободно распространяемого статистического пакета обработки данных и одноименного языка программирования R (пакет создан и распространяется в рамках проекта GNU - проект по разработке свободного программного обеспечения; подробнее о R можно узнать на интернет сайте http://www.r-project.org) на основании критерия Шапиро—Уилка (см. в [74], [75]) была проверена гипотеза нормальности регрессоров, входящих в модели (22) – (33). Реально достигнутые уровни значимости (РДУЗ) для H_i, $\ln H_i$, $H_i^{1.5}$, H_i^2, $H_i^{2.5}$, H_i^3 составили соответственно 0,2124; 0,07253; 0,1705; 0,08175; 0,02305; 0,04053. Поэтому далее регрессионные модели (32) и (33) мы рассматривать не будем (уровень значимости ниже порогового уровня 0,05).

Статистический критерий показывает, что нормальное приближение хорошо работает и для распределения роста , и для распределения его логарифма и степеней 1.5 и 2. Это происходит из-за того, что нормальность сохраняется при линейном преобразовании, а рассматриваемые преобразования ведут себя подобно линейным относительно рассматриваемых случайных величин в том смысле, что среднеквадратическое отклонение преобразованной случайной величины значительно меньше ее математического ожидания, и поэтому близость к нормальному закону сохраняется. Для степеней 2.5 и 3 это не так, и эти модели не проходят тест на нормальность и поэтому далее мы их рассматривать не будем.

С помощью пакета прикладных программ для решения технических задач и вычислений MatLab для каждой из моделей были оценены входящие в них параметры регрессии и вычислены значения эмпирического моста в узловых точках. После чего были построены графики эмпирических мостов (приведены в приложении).

Для сравнения моделей вычисляются выборочные дисперсии остатков, ста-

тистики $\omega_n^2 = \int\limits_0^1 (\widehat{Z}_n(t))^2\, dt$ по формуле

$$\omega_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{3}\left(Z_n^0\left(\frac{i}{n}\right) - Z_n^0\left(\frac{i-1}{n}\right)\right)^2 + Z_n^0\left(\frac{i}{n}\right) Z_n^0\left(\frac{i-1}{n}\right)\right),$$

которая напрямую следует из определения эмпирического моста, и реально достигаемые уровни значимости (по табл. 1 из [30] для однопараметрических моделей, по табл. 3 из [30] для двухпараметрических). Результаты приведены в табл. 2.

Таблица 2. Характеристики регрессионных моделей зависимости массы тела от роста.

Номер модели	Дисперсия $\widehat{\sigma^2}$	Статистика ω_n^2	РДУЗ
22	66,82	7,96	$< 10^{-5}$
23	0,0176	3,316	$< 10^{-5}$
24	0,0171	0,4503	0,053
25	0,0174	0,5928	0,024
26	0,0185	3,591	$< 10^{-5}$
27	0,0203	8,643	$< 10^{-5}$
28	0,0171	0,2697	0,0052
29	57,22	0,2273	0,013
30	57,17	0,2172	0,016
31	57,14	0,2101	0,018

Таблица 2 позволяет провести сравнение исследуемых моделей между собой. В частности, можно сделать вывод о том, что наилучшей из рассматриваемых является модель (24) (модель с наибольшим значением РДУЗ).

Отдельно отметим интересный эффект. Модель (24) лучше с точки зрения критерия согласия ω^2 нежели модель (28), в которой параметр модели b точно оценен. Получается, что в данном случае эффективнее угадать параметр мо-

дели, чем его оценивать. Конечно, ничего необычного в данном случае нет и данный эффект связан с тем, что для одно- и двухпараметрических моделей предельные распределения статистик критерия ω^2 существенно различаются: оценивание второго параметра теоретически (при выполнении предположений соответствующей модели) должно привести к значительно меньшим отклонениям от горизонтальной оси, чего на практике не происходит по причинам указанным выше.

Как показывает табл. 2, ни одна из рассмотренных моделей не демонстрирует высоких реально достигаемых уровней значимости, то есть хорошего соответствия с исследуемыми данными. Поэтому на следующем этапе исследования мы проанализируем выбросы исходных данных относительно предлагаемых моделей и их влияние на изучаемые характеристики. Для исследования выбросов обратимся к графическому изображению данных (рис 1).

На графике явно видны выбросы (аномально большие отклонения от любой из предлагаемых регрессионных зависимостей), которые могут привести к существенному искажению результатов исследования. Для устранения данного недостатка мы многократно провели процедуру очистки выборки (удаление из выборки аномально больших отклонений) с помощью известного правила «трех сигм». Каждый раз, когда несколько значений удалялось, оценки параметров и дисперсии остатков пересчитывались, после чего снова проверялась нормальность регрессора и процедура повторялась до тех пор, пока на очередном шаге ни одно значение не было удалено. В результате для каждой модели была получена новая двумерная выборка, для которой повторно были проведены все вычисления.

Результаты вычислений приведены в табл. 3 (модель (31) исключена, так как на очередном шаге очистки регрессор не прошел проверку на нормальность).

Таблица 3. Характеристики моделей после удаления выбросов.

Модель	Итераций	К-во удаленных	$\widehat{\sigma^2}$	ω_n^2	РДУЗ
22	4	14	53,07	6,87	$< 10^{-5}$
23	2	9	0,0158	4,35	$< 10^{-5}$
24	2	8	0,0158	0,8133	0,0068
25	4	10	0,0149	0,2718	0,165
26	2	9	0,0164	2,84	$< 10^{-5}$
27	2	8	0,0177	7,64	$< 10^{-5}$
28	3	9	0,0151	0,1741	0,0412
29	1	11	47,14	0,1605	0,0563
30	4	20	42,27	0,1674	0,047

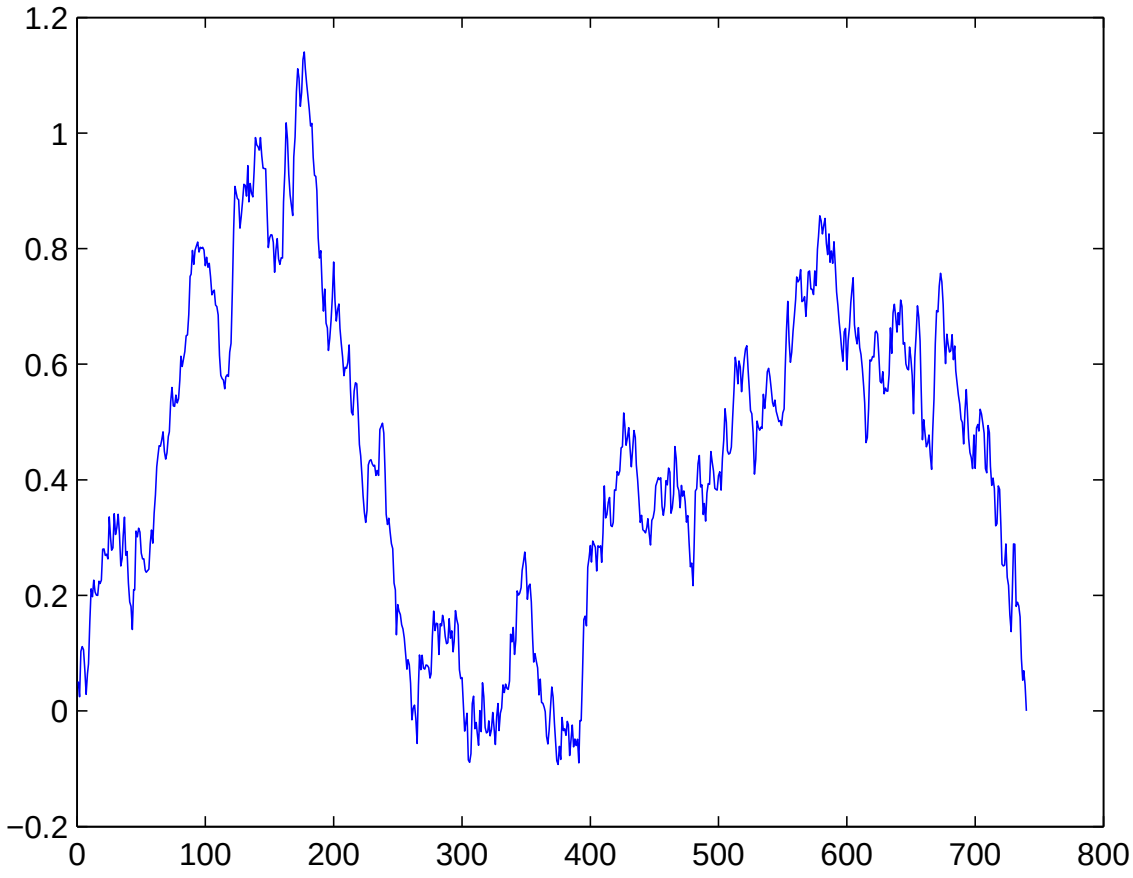


Рис. 2: Эмпирический мост для $\ln W_i = a + 2\ln H_i + \varepsilon_i$ (после окончательной очистки выборки)

Наилучший результат после очистки показала модель (25), эмпирический мост регрессионных остатков для нее приведен на рис. 2. Эту модель и следует использовать для анализа отклонений массы тела от нормы.

Оценка параметра a равна $\widehat{a} = -6,2171\ldots$. Таким образом, проведенное нами исследование позволяет определять значимость отклонений массы тела от нормы на основании логнормального закона с параметрами $\mu = -6,2171 + 2\ln H$, $\sigma^2 = 0,0149$, где H — рост студентки первого курса в сантиметрах.

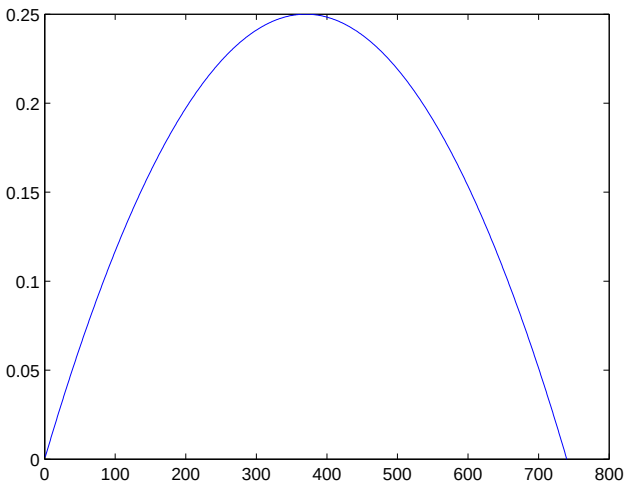
В частности, исключенные 10 наблюдений являются примерами таких отклонений, значимых на критическом уровне 0,0027 (согласно правилу «трех сигм»).

В заключение покажем значимость последнего вычитаемого в $\widetilde{K}_F^0$ (см. теорему 2). Для этого изобразим графически части выражения

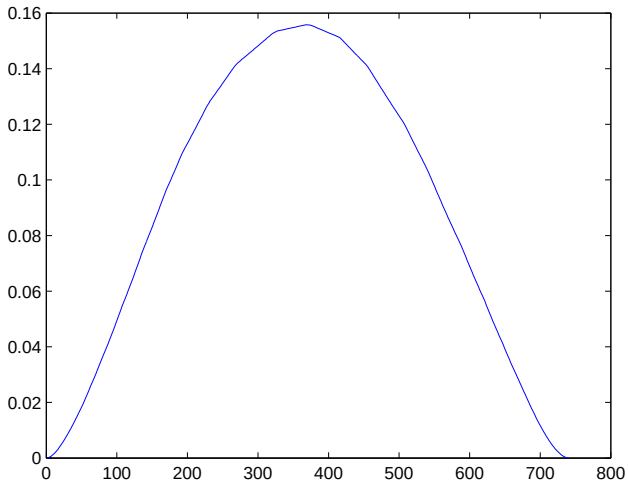
$$\mathbf{Var}\widetilde{Z}_F^0(t) = t(1-t) - \frac{(GL_F^0(t))^2}{\mathbf{Var}\xi_1},\ t \in [0,1],$$

используя сходимость с вероятностью единица

$$\frac{(GL_n^0(t))^2}{\overline{X^2} - \overline{X}^2} \to \frac{(GL_F^0(t))^2}{\mathbf{Var}\xi_1}$$



$t(1-t)$



$\frac{(GL_n^0(t))^2}{\overline{X^2} - \overline{X}^2}$

Отметим, что значения на оси абсцисс имеют характер нумерации точек, а не их значений. Конечно, оба графика определены и построены на отрезке [0,1].

Таким образом, видно, что значения на графиках имеют одинаковый порядок, поэтому пренебречь одной из частей $\widetilde{K}_F^0$ не представляется возможным.

2.4 Проверка гипотезы о линейной зависимости длины прыжка человека от его роста с помощью конструкции эмпирического моста

В данном параграфе мы проверим гипотезу о линейной зависимости длины прыжка человека от его роста. Сам факт наличия зависимости достаточно очевиден. Логично предположить, что высокорослые респонденты должны показывать лучшие результаты прыжков нежели их низкорослые коллеги. Но вот наличие именно линейной зависимости вызывает вопросы.

В качестве исходных данных, как и в предыдущем параграфе, были взяты сведения о росте (в сантиметрах) и длине прыжка с места (в сантиметрах) студенток первого курса лечебного факультета ГБОУ ВПО «Волгоградский государственный медицинский университет» (двумерная выборка объема 743, меньшая размерность выборки нежели в предыдущем параграфе обусловлена отсутствием данных о длине прыжка некоторых студенток в связи с освобождением от занятий по физической культуре). Для наглядности графическое представление выборки приведено на (рис. 3).

Перейдем непосредственно к описанию математической стороны вопроса. Предлагается проверить наличие линейной зависимости длины прыжка от роста. Предварительно, как и на протяжении всей работы, мы провели упорядочение выборочных данных по росту. Для проверки линейности нам необходимо провести анализ адекватности регрессионной модели

$$L_i = a + bH_i + \varepsilon_i, \tag{34}$$

где H_i и L_i – значения роста и длины прыжка соответственно. Отметим попутно, что проводить проверку нормальности значений рост нет необходимости, так как это уже было сделано в предыдущем параграфе.

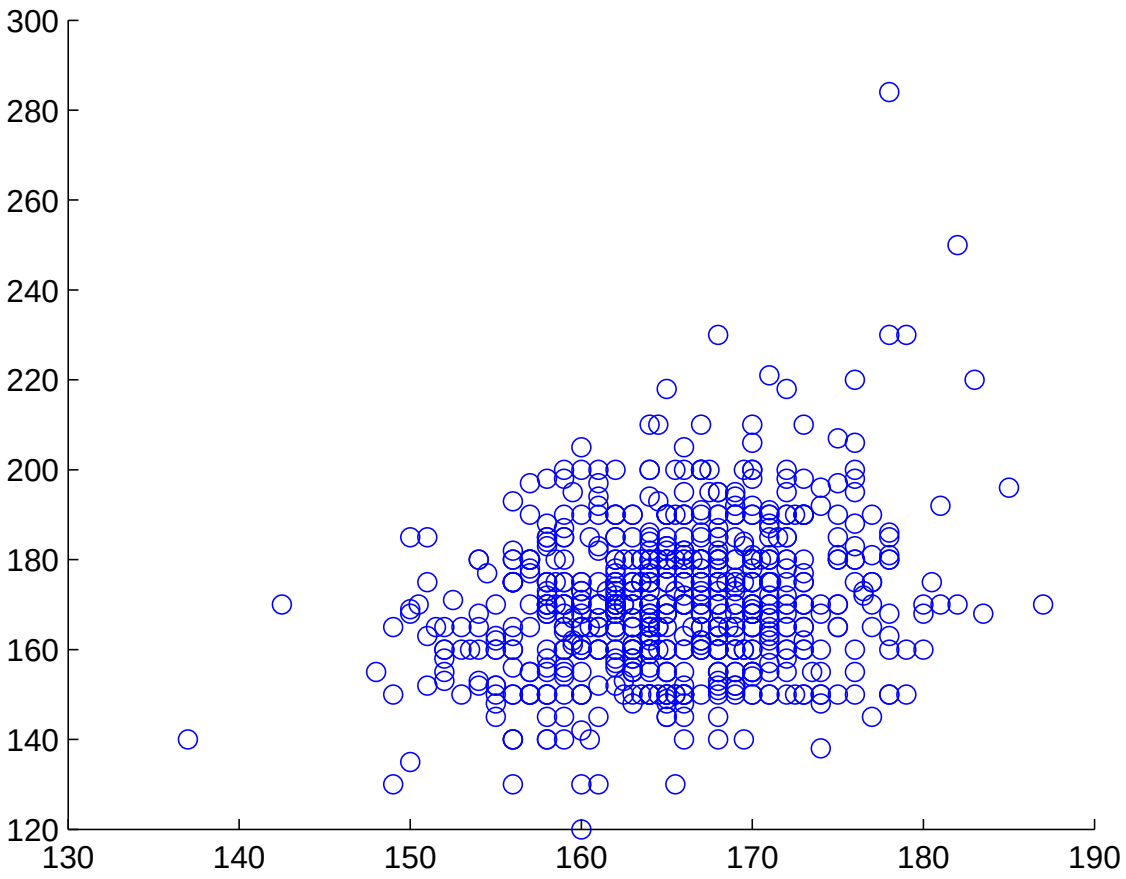


Рис. 3: Зависимость длины прыжка с места (в см) от роста (в см)

Как и в предыдущем параграфе с помощью пакета MatLab были оценены входящие в модель параметры, вычислены значения выборочной дисперсии остатков, эмпирического моста и статистики омега-квадрат. Кроме того, по графику выборки видны аномальные выбросы значений. С помощью правила трех сигм была проведена трехкратная фильтрация выборки. После каждой итерации все параметры были пересчитаны, результаты приведены в таблице 4.

Из таблицы 4 видно, что как для исходной, так и для всех фильтрованных выборок, РДУЗ имеет достаточно высокое значение, что говорит о принятии гипотезы о линейной зависимости. Также для наглядности на рис.4 приведен итоговый график эмпирического моста.

Таким образом, проведенное нами исследование позволяет определять значимость отклонений длины прыжка с места от нормы на основании нормально-

го закона с параметрами $91,24 + 0,48H$, $\sigma^2 = 211,81$, где H — рост студентки первого курса в сантиметрах. В частности, исключенные 10 наблюдений являются примерами таких отклонений, значимых на критическом уровне 0,0027 (согласно правилу «трех сигм»).

Таблица 4. Характеристики модели линейной зависимости длины прыжка от роста по первоначальной и фильтрованным выборкам.

К-во наблюдений	$\widehat{a}$	$\widehat{b}$	$\widehat{\sigma^2}$	ω_n^2	РДУЗ
743	64,12	0,64	256,99	0,04	$0,7168$
738	84,62	0,52	225,04	0,048	$0,6769$
735	87,25	0,5	216,64	0,05	$0,65$
733	91,24	0,48	211,81	0,067	0,4651

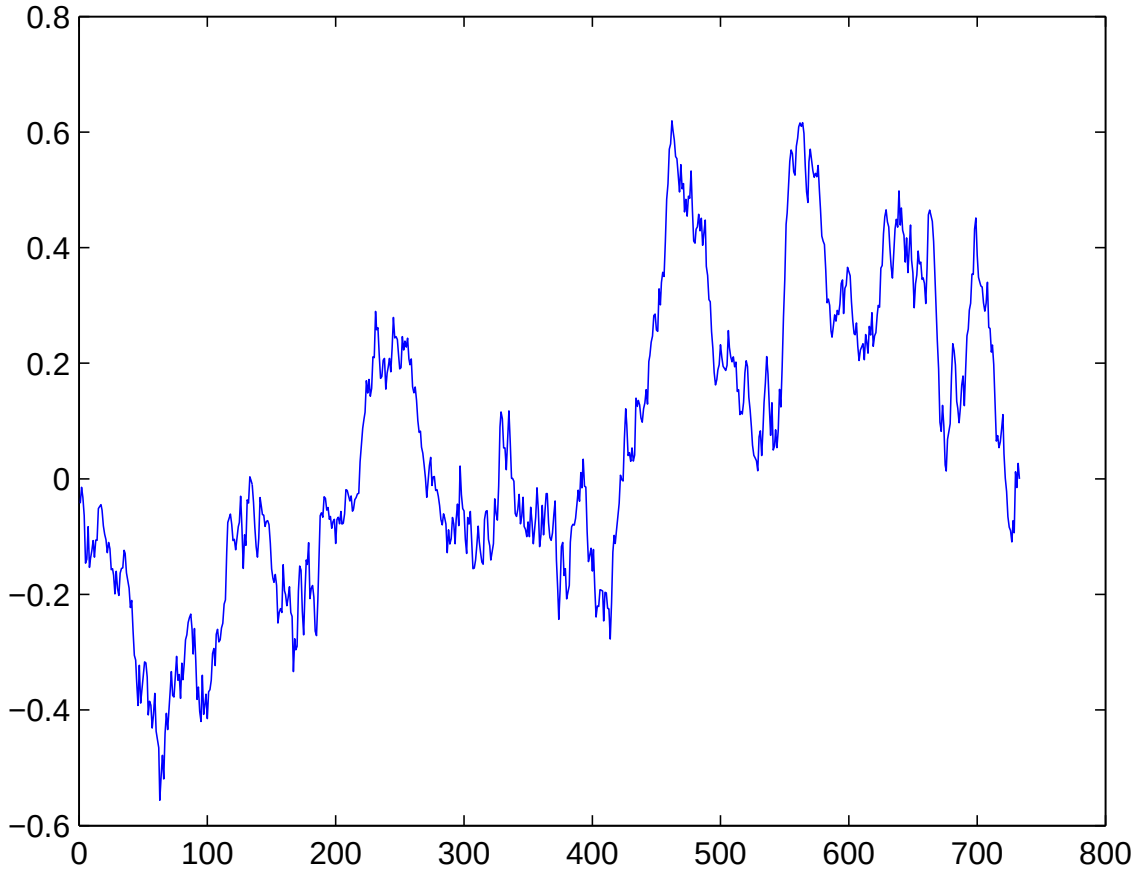


Рис. 4: Эмпирический мост после окончательной фильтрации выборки

Заключение и благодарности

Итак, подводя итог, сформулируем еще раз основные результаты диссертационного исследования, которые определяются следующими положениями:

- Разработан и обоснован новый алгоритм (а на его основе два решающих правила) анализа адекватности одно- и двухпараметрических линейных регрессионных моделей на порядковые статистики, основанный на доказанных предельных теоремах и классических статистических критериях типа хи-квадрат и омега-квадрат и ориентированный на практическое применение;

- Проведено сравнение предлагаемого алгоритма с известным F-тестом; приведен пример, когда применение построенного алгоритма предпочтительнее чем использование F-теста;

- Проиллюстрирована практическая применимость предлагаемого алгоритма к разнообразным реальным прикладным задачам анализа данных, а именно проведено исследование зависимости массы человеческого тела и его роста, длины прыжка с места и роста человека и зависимости курсов американского доллара и евро с помощью конструкции эмпирического моста;

- Даны полные методические рекомендации по практическому применению предложенного алгоритма к прикладным задачам анализа данных.

Как было отмечено во введении, задачи регрессионного анализа имеют важное прикладное значение, а количество публикаций и научных изысканий на эту тему говорят о растущей актуальности данных вопросов. Конструкция эмпирического моста действительно дает исследователю в руки весьма эффективный и полезный инструмент анализа регрессионных зависимостей, который еще на

первом этапе исследования позволяет ранжировать модели по их пригодности и отсеивать неадекватные выборочным данным модели.

Исследование регрессионных зависимостей с использованием порядковых статистик также является весьма актуальным, так как задачи с упорядочиванием наблюдений, в свою очередь, постоянно возникают на практике, так как графическое изображение данных (а следовательно их упорядочение) является весьма полезным, удобным и естественным способом исследования в целом ряде приложений.

Исследование модели, в которой ошибки управляются цепью Маркова, позволяют отказаться от классического предположения регрессионного анализа — гомоскедастичности. Полученная в диссертации предельная теорема показывает универсальность конструкции эмпирического моста и для случая „неклассической“ регрессии. Что, в свою очередь, сильно расширяет возможности исследования.

Примеры применения полученных результатов, приведенные в главе 2 показывают всю широту и универсальность применения предложенных конструкций в различных областях современной науки, техники, отраслях народного хозяйства.

Следующим этапом исследования является обобщение полученных в работе результатов на случай многопараметрической регрессии, а также на случай нелинейной зависимости. Таким образом, настоящая диссертация является отправной точкой большого и перспективного исследования.

Автор благодарит научного руководителя С. Г. Фосса и своего соавтора А. П. Ковалевского за постановку задач, многочисленные полезные обсуждения и совместные исследования, важность которых сложно переоценить, а также моральную поддержку на пути написания настоящей диссертации. Также автор выражает признательность сотрудникам лаборатории теории вероятностей и математической статистики института математики им. С.Л. Соболева, и лично А. И. Саханенко за полезные обсуждения, ссылки на неизвестные ранее автору

публикации.

Литература

[1] *Андерсон Т.* Статистический анализ временных рядов. М.: Мир, 1976.

[2] *Аркашов Н.С., Ковалевский А. П.* Вероятностная модель цен на квартиры // Сиб. журн. индустр. матем., Т. 15 №2, стр. 11–20, 2012.

[3] *Биллингсли П.* Сходимость вероятностных мер. М.: Наука, 1977.

[4] *Болдин М.В., Симонова Г.И., Тюрин Ю.Н.* Знаковый статистический анализ линейных моделей. М.: Наука, 1997.

[5] *Большев Л. Н., Смирнов Н.В.* Таблицы математической статистики. М.: Наука, 1983.

[6] *Боровков А.А.* Математическая статистика. М.: Наука, 1984.

[7] *Боровков А.А.* Теория вероятностей. М.: Наука, 1998.

[8] *Бродский Б. Е., Дарховский Б. С.* Проблемы и методы вероятностной диагностики // Автомат. и телемех., №8, стр. 3–50, 1999.

[9] *Булинский А.В., Ширяев А.Н.* Теория случайных процессов. М.: Физматлит, 2005.

[10] *Буркатовская Ю. Б., Воробейчиков С. Э.* Обнаружение разладки процесса авторегрессии, наблюдаемого с помехами // Автомат. и телемех., №3, стр. 76–89, 2000.

[11] *Вентцель А.Д* Курс теории случайных процессов. М.: Физматлит, 1996.

[12] *Воробейчиков С. Э., Конев В. В.* Последовательный метод обнаружения разладок случайных процессов рекуррентного типа // Автомат. и телемех., №5, стр. 27–38, 1984.

[13] *Воробейчиков С. Э., Конев В. В.* Об обнаружении разладок в динамических системах // Автомат. и телемех., №3, стр. 56–68, 1990.

[14] *Воробейчиков С. Э., Конев В. В.* Характеристики процедуры обнаружения разладки процесса авторегрессии с неизвестным распределением помехи // Автомат. и телемех., №2, стр. 68–75, 1992.

[15] *Гихман И.И., Скороход А.В.* Введение в теорию случайных процессов. М.: Наука, 1977.

[16] *Гусарова Г. В. , Ковалевский А. П., Макаренко А. Г.* Критерии наличия разладки // Сиб. журн. индустр. матем., 8:4, стр. 18—33, 2005.

[17] *Дамодаран А.* Инвестиционная оценка. Инструменты и методы оценки любых активов. М.: Альпина Бизнес Букс, 2004.

[18] *Доугерти К.* Введение в эконометрику. М.: ИНФРА-М, 1999.

[19] *Дрейпер Н., Смит Г.* Прикладной регрессионный анализ. Т. 1. М.: Финансы и статистика, 1986.

[20] *Дэйвид Г.* Порядковые статистики. М.: Наука, 1979.

[21] *Зигангиров К. М.* Задача поиска в системе с конечным числом позиций // Радиотехника и электроника,Т. 8, № 1. стр. 16-24, 1963.

[22] *Ковалевский А. П.* Статистические критерии обнаружения разладки регрессии с циклическим трендом // Научный вестник НГТУ, №3(52), стр. 55–62, 2013.

[23] *Ковалевский А. П., Шахраманьян А. М.* Анализ дефектов строи- тельных конструкций методом эмпирического моста // Научный вестник НГТУ, №3(56), стр. 171–180, 2014.

[24] *Конев В. В., Дмитриенко А. А.* О гарантированном оценивании параметров авторегрессии при неизвестной дисперсии помех // Автомат. и телемех., №2, стр. 87—99, 1994.

[25] *Конев В. В., Пергаменщиков С. М.* Гарантированное оценивание параметров авторегрессии на основе последовательного корреляционного метода // Тр. МИАН, Т. 202, стр. 149—169, 1993.

[26] *Конев В. В., Пергаменщиков С. М.* Об оценивании параметра авторегрессии на основе обобщенного метода наименьших квадратов // УМН, Т. 5, вып. 6(306), стр. 187—188, 1995.

[27] *Конев В. В., Пергаменщиков С. М.* О гарантированном оценивании параметров линейной регрессии при зависимых помехах // Автомат. и телемех., №2, стр. 75—87, 1997.

[28] *Коршунов Д. А., Фосс С. Г., Эйсымонт И. М.* Сборник задач и упражнений по теории вероятностей. СПб.: Лань, 2004.

[29] *Крамер Г.* Математические методы статистики. М.: Мир, 1975.

[30] *Мартынов Г. В.* Критерии омега-квадрат. М.: Наука, 1978.

[31] *Муганцева Л. А.* Проверка нормальности в схемах одномерной и многомерной линейной регрессии // ТВП, 22:3, 603—614, 1977.

[32] *Мудров В.И., Кушко В.Л.* Методы обработки измерений. Квазиправдоподобные оценки. М.: Радио и связь, 1983.

[33] *Прохоров Ю.В.*. Сходимость случайных процессов и предельные теоремы теории вероятностей // ТВП, 1:2, стр. 177–238, 1956.

[34] *Розанов Ю.А.* Случайные процессы (краткий курс). М.: Наука, 1979.

[35] *Сархан А.Е., Гринберг Б.Д.* Введение в теорию порядковых статистик. М.: Статистика, 1970.

[36] *Трухачева Н.В.* Математическая статистика в медико-биологических исследованиях с применением пакета Statistica. ГЭОТАР-Медиа, 2012.

[37] *Тырсин А.Н., Соколов Л.А.* Оценивание линейной регрессии на основе обобщенного метода наименьших модулей // Вестн. Сам. гос. техн. ун-та. Сер. Физ.-мат. науки, выпуск 5(21), стр. 134-–142, 2010.

[38] *Феллер В.* Введение в теорию вероятностей и ее приложения. Т. 1. М.: Мир, 1984.

[39] *Ширяев А. Н.* Об оптимальных методах в задачах скорейшего обнаружения // ТВП, 8:1, стр. 24–51, 1963.

[40] *Ширяев А. Н.* Статистический последовательный анализ. М.: Наука, 1976.

[41] *Ширяев А. Н.* Вероятность — 1. М.: МЦНМО, 2004.

[42] *Abdalla S.Z.S.* Modelling exchange rate volatility using garch models: Empirical evidence from arab countries // International Journal of Economics and Finance, Vol. **4**, №3., pp. 216–229, 2012.

[43] *Arnold J. R., Libby W. F.* Age determinations by radiocarbon content: checks with samples of known age // Science, Vol. 110 (2869), pp. 678–680, 1949.

[44] *Aue A., Horvath L., Huskova M., Kokoszka P.* Testing for change in polynomial regression // Bernoully 14(3), pp. 637–660, 2008.

[45] *Bhattacharya P. K.* Convergence of sample paths of normalized sums of induced order statistics // The Annals of Statist., 2, pp. 1034–1039, 1974.

[46] *Bildirici M., Alp E. A., Ersin O.* TAR-cointegration neural network model: An empirical analysis of exchange rates and stock returns // Expert Systems with Applications, Vol. **37**, Issue 1, pp. 2–11, 2010.

[47] *Bischoff W.* A functional central limit theorem for regression models // Ann. of Stat., Vol. 26, № 4, pp. 1398–1410, 1997.

[48] *Box G.E.P, Cox D.R.* An analysis of transformation. // Journal of the Royal Statistical Society Series B, 26(2), pp. 211–243, 1964.

[49] *David, H.A.* Concomitants of order statistics // Bull. Internat. Statist. Inst., 45, pp. 295–300, 1973.

[50] *Davydov Y., Zitikis R.* Functional limit theorems for induced order statistics // Mathematical Methods of Statistics, 9(3), pp. 297–313, 2000.

[51] *Davydov Y., Zitikis R.* Convex rearrangements of random elements // Fields Institute Communications, Vol.44, pp. 141–171, 2004.

[52] *Deheuvels P., Martynov G. V.* Cramer-von Mises-type tests with applications to tests of independence for multivariate extreme-value distributions // Comm. Stat. — Theory and Methods, Vol.25, No. 4, pp. 871–908, 1996.

[53] *Durbin J., Knott M., Taylor C. C.* Components of Cramer - von Mises Statistics II // J. Roy. Statist. Soc., B 37, pp. 216–237, 1975.

[54] *Galton F.* Regression towards mediocrity in hereditary stature // Journal of the Anthropological Institute of Great Britain and Ireland, Vol. 15, pp. 246–263, 1885.

[55] *Gastwirth J. L.* A general definition of the Lorenz curve // Econometrica, Vol. 39, pp. 1037–1039, 1971.

[56] *Goldie C. M.* Convergence theorems for empirical Lorenz curves and their inverses // Advances in Applied Probability, Vol. 9, pp. 765–791, 1977.

[57] *Hausman J.A.* Specification tests in econometrics. // Econometrica, 46(6), pp. 1251–1272, 1978.

[58] *Heisenberg. W.* Uber den anschaulichen Inhalt der quantentheoretischen Kinematik und Mechanik // Zeitschrift fu"r Physik., Vol. 43, № 3—4, pp. 172–198, 1927.

[59] *Hoeffding W.* On the distribution of the expected values of the order statistics // Ann. Math. Statist., Vol. 24, № 1, pp. 93–100, 1953.

[60] *Keys A., Fidanza F., Karvonen M. J., Kimura N., Taylor H. L.* Indices of relative weight and obesity // Journal of Chronic Diseases 25 (6–7), pp. 329–343, 1972.

[61] *Kingman, J. F. C.* The ergodic theory of subadditive stochastic processes. // J. R. Statist. Soc. 30, pp. 499–510, 1968.

[62] *Kovalevskii, A.* A regression model for prices of second-hand cars. // Applied methods of statistical analysis. Applications in survival analysis, reliability and quality control, 124–128, 2013.

[63] *MacNeill I. B.* Limit processes for sequences of partial sums of regression residuals // Ann. Prob., Vol. 6, № 4, pp. 695–698, 1978.

[64] *McLeish, D. L.* Invariance principles for dependent variables. Zeitschrift füjr Wahrscheinlichkeitstheorie und Verwandte Gebiete 32, pp. 165–178. 1975.

[65] *Nwankwo Steve C.* Autoregressive Integrated Moving Average (ARIMA) Model for Exchange Rate (Naira to Dollar) // Academic Journal of Interdisciplinary Studies MCSER Publishing, Vol. **3**, №4, pp. 429–433, 2014.

[66] *Pacelli V., Bevilacqua V., Azzollini M.* An Artificial Neural Network Model to Forecast Exchange Rates // Journal of Intelligent Learning Systems and Applications, Vol. **3**, №2/2011, pp. 57–69, 2011.

[67] *Pacelli V.* Forecasting Exchange Rates: a Comparative Analysis // International Journal of Business and Social Science, Vol. **3**, №10, pp. 145-156, 2012.

[68] *Piackett R. L.* Studies in the history jf probability and statistics. XXIX. The discovery of the method of least squares // Biometrika, 59, pp. 239–251, 1972.

[69] *Quandt R. E.* The estimation of parameters of a linear regression system obeying two separate regimes // J. Amer. Statist. Assoc., Vol. 50, pp. 873–880, 1958.

[70] *Quandt R. E.* Tests of the hypothesis that a linear regression system obeys two separate regimes // J. Amer. Statist. Assoc., Vol. 55, pp. 324–330, 1960.

[71] *Quetelet A.* Recherches sur le poids de l'homme aux different âges // Nouveaux Memoire de l'Academie Royale des Sciences et Belles-Lettres de Bruxelles, p. VII, 1832.

[72] *Ramsey J.B.* Tests for Specification Errors in Classical Linear Least Squares Regression Analysis. // Journal of the Royal Statistical Society Series B, 31(2), pp. 350–371, 1969.

[73] *Ramzan S., Ramzan S., Zahid F.M.* Modeling and forecasting exchange rate dynamics in Pakistan using arch family of models // Electron. J. App. Stat. Anal., Vol. **5**, Issue 1, pp. 15–29, 2012.

[74] *Shapiro S.S., Wilk M.B.* An analysis of variance test for normality (complete samples) // Biometrika, Vol.52, pp. 591–611, 1965.

[75] *Shapiro S.S., Francia R.S.* An appriximate analysis of variance test fo normality // J. Amer. Statist. Assoc., 337, pp. 215–216, 1972.

[76] *Stute, W.* Nonparametric model checks for regression // Ann. Statist. 25, 613–641, 1997.

[77] *Zarembka P.* Functional form in the demand for money. // Journal of the American Statistical Association, 63(322), pp. 502–5111, 1968.

[78] *Ковалевский А. П., Шаталин Е.В.* Асимптотика сумм остатков однопараметрической линейной регрессии, построенной по порядковым статистикам // Теория вероятностей и ее применения, 59:3. – 2014. – С. 452–467. DOI: 10.4213/tvp4579 (входит в РИНЦ).

Перевод: A. P. Kovalevskii and E. V. Shatalin Asymptotics of Sums of Residuals of One-Parameter Linear Regression on Order Statistics // Theory of Probability and Its Applications, Vol. 59, No. 3 – 2015. – pp. 375-387. DOI: 10.1137/S0040585X97T987193 (входит в Web of Science, Scopus).

[79] *Шаталин Е.В.* Исследование регрессионных моделей зависимости курсов американского доллара и евро с помощью эмпирического моста // Сибирский журнал чистой и прикладной математики, №3, стр. 91–97, 2015. DOI: 10.17377/PAM.2015.15.308 (входит в РИНЦ).

[80] *Ковалевский А. П., Шаталин Е.В.* Выбор регрессионной модели зависимости массы тела от роста с помощью эмпирического моста // Вестник Томского государственного университета. Математика и механика, №5(37), стр. 35–47, 2015. DOI 10.17223/19988621/37/3 (входит в РИНЦ).

[81] *Kovalevskii A. P., Shatalin E. V.* A limit process for a sequence of partial sums of residuals of a simple regression on order statistics with Markovmodulated noise // Probability and Mathematical Statistics, Vol. 36.1, pp. 113–120, 2016. (входит в Scopus).

[82] *Шаталин Е.В.* Асимптотика эмпирического моста по остаткам регрессии на порядковые статистики // Материалы XLIX международной научной студенческой конференции „Студент и научно-технический прогресс.“ Новосибирск: НГУ, стр. 205, 2011.

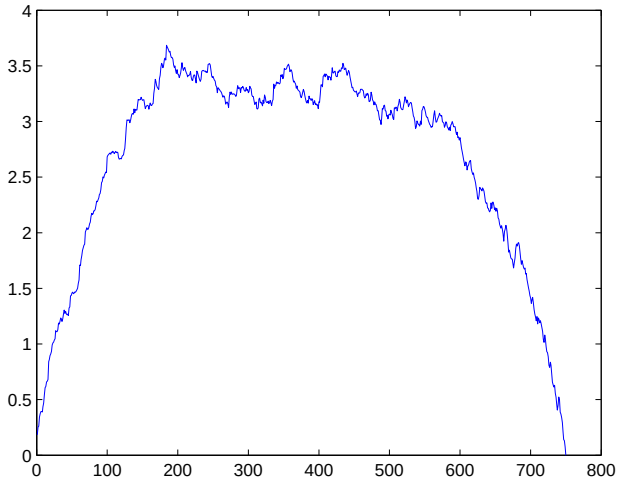
[83] *Ковалевский А. П., Шаталин Е.В.* Asymptotic distribution of empirical bridge for regression on order statistics // Programme of V International Conference „Limit Theorems in Probability Theory and Their Applications“. Novosibirsk: Sobolev Institute of Mathematics, pp. 26, 2011.

[84] *Шаталин Е.В., Ковалевский А.П.* Асимптотика эмпирического моста в линейных регрессионных моделях, построенных по порядковым статистикам // Материалы XIV всероссийского симпозиума по прикладной и промышленной математике (осенняя сессия). Великий Новгород, стр. 573–574, 2013.

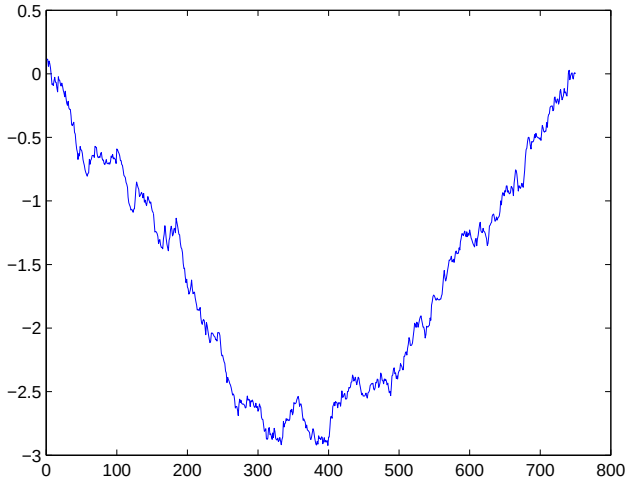
[85] *Шаталин Е.В.* Предельные процессы для частичных сумм остатков регрессии на порядковые статистики с ошибками, управляемыми цепями Маркова // Материалы 52-й международной научной студенческой конференции МНСК-2014. Новосибирск: НГУ, стр. 241, 2014.

[86] *Kovalevskiy A., Shatalin E.* Limit processes for sequences of partial sums of residuals of regressions against order statistics with Markov-modulated noise // Conference program and abstract book of 11th International conference on ordered statistical data. Bedlewo(Poland), pp. 37-38, 2014.

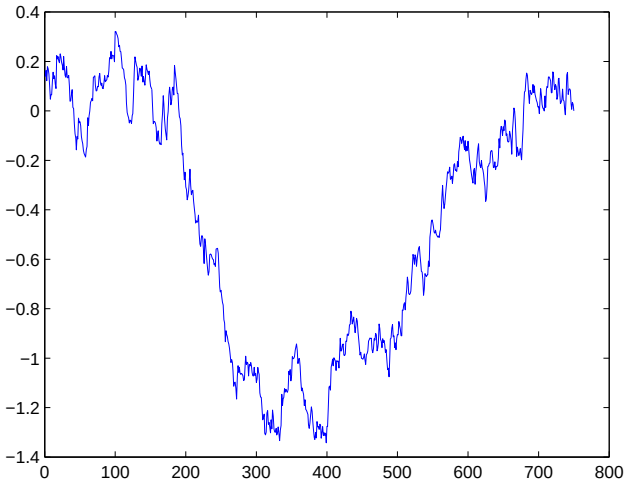
Приложение (графики эмпирических мостов)



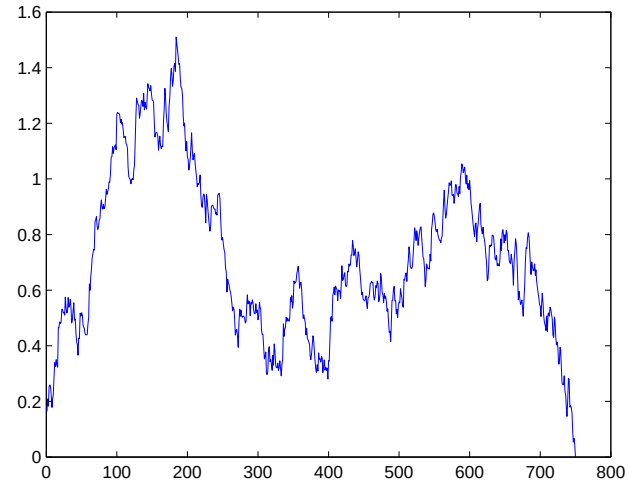
$$W_i = \theta + H_i + \varepsilon_i$$



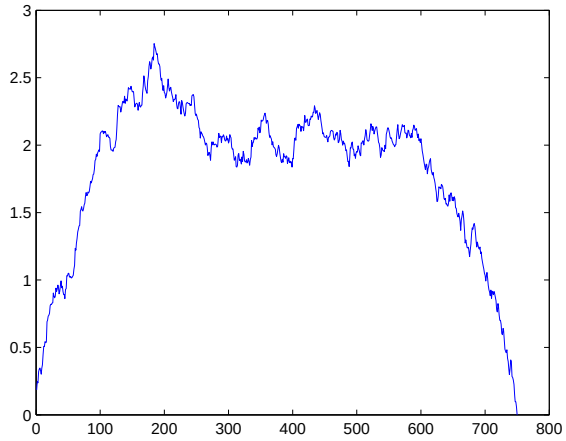
$$\ln W_i = a + \ln H_i + \varepsilon_i$$



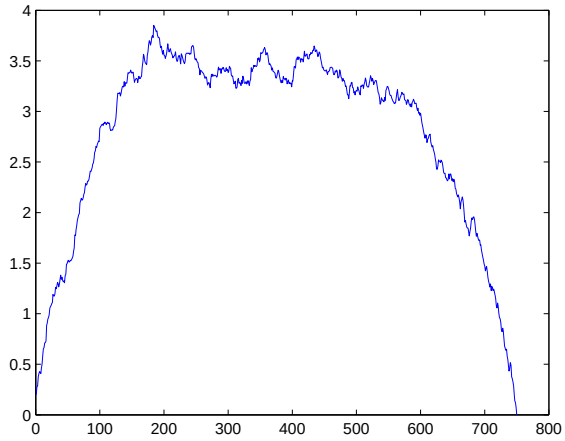
$$\ln W_i = a + 1.5 \ln H_i + \varepsilon_i$$



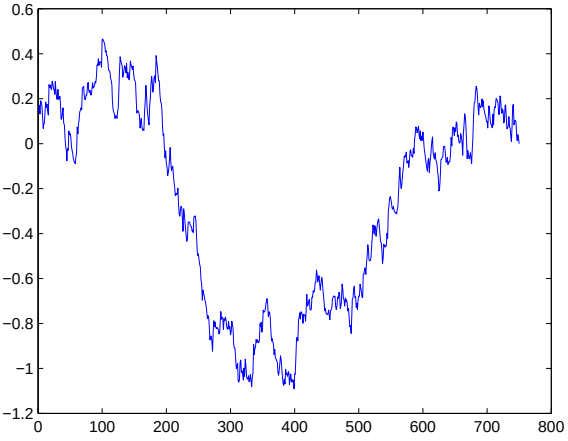
$$\ln W_i = a + 2 \ln H_i + \varepsilon_i$$



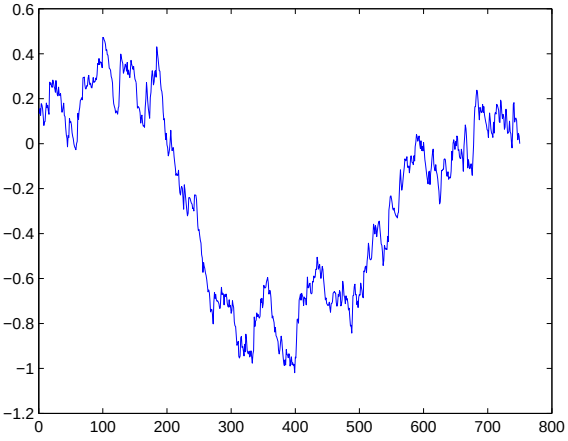
$$\ln W_i = a + 2.5 \ln H_i + \varepsilon_i$$



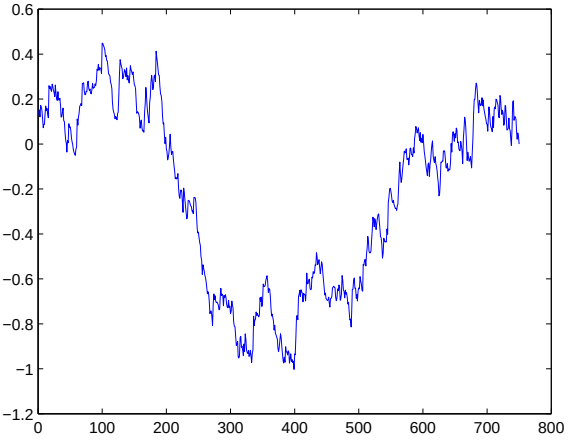
$$\ln W_i = a + 3 \ln H_i + \varepsilon_i$$



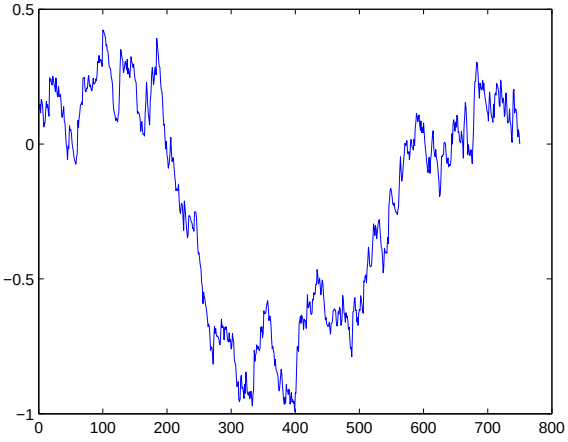
$$\ln W_i = a + b \ln H_i + \varepsilon_i$$



$$W_i = a + bH_i + \varepsilon_i$$



$$W_i = a + bH_i^{1.5} + \varepsilon_i$$



$$W_i = a + bH_i^2 + \varepsilon_i$$