На правах рукописи

Шаталин Евгений Викторович

ЭМПИРИЧЕСКИЙ МОСТ И ЗАДАЧИ ТЕСТИРОВАНИЯ АДЕКВАТНОСТИ РЕГРЕССИОННЫХ МОДЕЛЕЙ АНАЛИЗА ДАННЫХ

05.13.17 — Теоретические основы информатики

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата физико-математических наук

Новосибирск
2017

Работа выполнена в Федеральном государственном бюджетном учреждении науки Институте математики им. С. Л. Соболева Сибирского отделения Российской академии наук в лаборатории теории вероятностей и математической статистики

Научный руководитель:
доктор физико-математических наук, профессор
Фосс Сергей Георгиевич

Официальные оппоненты:
Цициашвили Гурами Шалвович, доктор физико-математических наук, профессор, Федеральное государственное бюджетное учреждение науки «Институт прикладной математики Дальневосточного отделения Российской академии наук», главный научный сотрудник
Хрущев Сергей Евгеньевич, кандидат физико-математических наук, Новосибирский государственный университет экономики и управления, заведующий кафедрой математики и естественных наук

Ведущая организация: Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский Томский государственный университет»

Защита состоится 8 ноября 2017 г. в 15.00 на заседании диссертационного совета Д 999.082.03 на базе Федерального государственного бюджетного учреждения науки Института систем информатики им. А.П. Ершова Сибирского отделения Российской академии наук (ИСИ СО РАН) по адресу 630090, г. Новосибирск, проспект Академика Лаврентьева, 6, комн. 245.

С диссертацией можно ознакомиться в библиотеке и на сайте ИСИ СО РАН http://www.iis.nsk.su/files/shatalin_e_v.pdf

Автореферат разослан «___» _____ 2017 года.

Ученый секретарь
диссертационного совета
канд. физ.-мат. наук  Мурзин Федор Александрович

Общая характеристика работы

Актуальность темы. Объектом исследования настоящей работы являются 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 работы МакНилла (1978). 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].

Основные результаты. Основные результаты диссертационного исследования 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.

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.

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

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

Достоверность и обоснованность полученных результатов. Все полученные в диссертации результаты имеют строгое математическое обоснование в форме утверждений, лемм, теорем и следствий из них. Применимость и эффективность полученных результатов подтверждена практическим их применением к реальным прикладным задачам анализа данных.

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.

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.

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 промышленной 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 введе-

ния, 2 глав, заключения, списка литературы и приложения с графиками. Общий объем диссертации составляет 102 страниц машинописного текста. Библиография содержит 86 наименований, в том числе 9 работ автора по теме диссертации.

Содержание работы

Во **введении** раскрывается актуальность исследуемой в данной диссертации проблемы, дается исторический очерк по данной теме, приводится обзор известных результатов, излагается содержание работы, обосновывается теоретическая и практическая значимость полученных в работе результатов.

В **первой главе** приводится алгоритм анализа адекватности линейных регрессионных моделей на порядковые статистики, а также рассматриваются обосновывающие его предельные теоремы для эмпирического моста. А именно, доказаны три предельные теоремы для случаев однопараметрической и двух видов двухпараметрической модели.

В **параграфе 1.1** вводятся необходимые понятия и формулируются основные теоретические результаты диссертационной работы.

Прежде чем формулировать основные результаты, приведем необходимые обозначения и условия, общие для всех трех рассматриваемых регрессионных моделей. Рассмотрим последовательность независимых, одинаково распределенных случайных величин $\xi_1, \ldots, \xi_n$ с общей функцией распределения F. По данным величинам построим их вариационный ряд, то есть порядковые статистики, $X_{ni} = \xi_{i:n},\ i = 1, \ldots, n$, где $\xi_{1:n} \leqslant \ldots \leqslant \xi_{n:n}$.

Кроме того, рассмотрим другую, не зависящую от первой, последовательность независимых, одинаково распределенных случайных величин $\varepsilon_1, \ldots, \varepsilon_n$ с нулевым математическим ожиданием и конечной ненулевой дисперсией σ^2.

В моделях далее величины $\{X_{ni}\}$ будут выступать в роли регрессоров, а $\{\varepsilon_i\}$ в роли случайных ошибок или помех (с некоторыми видоизменениями в модели 3).

Далее, заменяя в регрессионных моделях параметры их оценками (по методу наименьших квадратов) и откидывая случайный член, мы приходим к величинам $\widehat{Y}_{ni}$ — прогнозным значениям отклика, а затем и к остаткам регрессии $\widehat{\varepsilon}_{ni} = \ Y_{ni} - \widehat{Y}_{ni}$.

Теперь мы можем ввести основной объект нашего исследования — эмпирический мост. 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 $\widehat{\Delta}_{nk} = \widehat{\varepsilon}_{n1} + \ldots + \widehat{\varepsilon}_{nk},\ k = 0, 1, \ldots, n,\ \widehat{\Delta}_{n0} = 0,\ \widehat{\sigma^2} = \overline{\widehat{\varepsilon}^2} - (\overline{\widehat{\varepsilon}})^2$. В дальнейшем мы будем опускать в двойных индексах индекс n, где это не вызывает недоразумений.

Отметим попутно, что при условии сходимости оценки дисперсии к ее истинному значению, для нахождения предельного процесса для эмпирического моста достаточно найти предельный процесс для более простой случайной ломаной. А именно, для ломаной Z_n, построенной по точкам

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

так как эмпирический мост получается из нее непрерывным в равномерной метрике на $[0, 1]$ преобразованием.

Сформулируем кратко основные шаги разработанного в диссертации алгоритма анализа адекватности регрессионных моделей:

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.

Наконец, введем последний объект, который понадобиться нам для формулирования предельных теорем. $GL_F(t) = \int\limits_0^t F^{-1}(s)\, ds$ – *теоретическая обобщенная кривая Лоренца*[2], где $F^{-1}(s) = \sup\{x : \ F(x) < s\}$

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

— квантильное преобразование функции распределения $F(x)$. Обозначим $GL_F^0(t) = GL_F(t) - tGL_F(1)$.

На протяжении всей работы символ $\Longrightarrow$ обозначает 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[3].

Модель 1 (однопараметрическая модель)

Рассмотрим однопараметрическую ($\theta \in \mathbf{R}$) модель линейной регрессии на порядковые статистики:

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

Оценим неизвестный параметр регрессии по методу наименьших квадратов $\widehat{\theta} = \overline{XY}/\overline{X^2}$. Подставив оценку параметра в модель, получим прогнозные значения зависимой переменной $\widehat{Y}_{ni} = \widehat{\theta} X_{ni}$, а затем и остатки регрессии $\widehat{\varepsilon}_{ni} = Y_{ni} - \widehat{Y}_{ni}$.

Предельное поведение эмпирического моста $\widehat{Z}_n$ описывается в следующей теореме.

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

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

Модель 2 (двухпараметрическая модель)

Рассмотрим двухпараметрическую ($a, b \in \mathbf{R}$) линейную регрессионную модель:

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

Для неизвестных параметров a и b введем так называемые оценки наименьших квадратов (или, как еще их называют, оценки Гаусса-Маркова): $\widehat{b}_n = \frac{\overline{XY} - \overline{X}\,\overline{Y}}{\overline{X^2} - \overline{X}^2}$, $\widehat{a}_n = \overline{Y} - \widehat{b}_n\,\overline{X}$.

[3]Биллингсли П. Сходимость вероятностных мер. М.: Наука, 1977, стр.82.

Массивы прогнозных значений $\{\widehat{Y}_{ni}\}$ и регрессионных остатков $\{\widehat{\varepsilon}_{ni}\}$ вводятся аналогично: $\widehat{Y}_{ni} = \widehat{a}_n + \widehat{b}_n X_{ni},\ \widehat{\varepsilon}_{ni} = Y_{ni} - \widehat{Y}_{ni}$.

В этом случае предельное поведение эмпирического моста $\widehat{Z}_n$ дается следующей теоремой.

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

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

Модель 3 (двухпараметрическая модель, управляемая цепью Маркова.)

Для описания модели, нам потребуются 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:

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

Как и в случае модели (2), в качестве оценок неизвестных параметров выступают классические оценки Гаусса-Маркова.

В данном случае предельное поведение эмпирического моста $\widehat{Z}_n$ описывается следующей теоремой.

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 ковари-

ационной функцией

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

В **параграфах 1.2-1.4** приведены доказательства теорем 1-3 соответственно.

В **параграфе 1.5** проведено сравнение обсуждаемого в работе алгоритма с известным F-тестом. На конкретном примере показано преимущество подхода с использованием эмпирического моста перед F-тестом. А именно построен пример, когда F-тест ошибочно принимает, а эмпирический мост отвергает заведомо ложную модель.

Во **второй главе** излагаются прикладные аспекты полученных теоретических результатов. Также проиллюстрировано практическое применение методов и результатов работы к реальным примерам, а именно исследованы зависимости (а) массы человеческого тела от его роста, (б) длины прыжка с места и роста человека и (в) зависимость курсов единой европейской валюты (евро) и американского доллара.

В **параграфе 2.1** описаны аспекты практического применения полученных в работе результатов.

Для того, чтобы применять, например, 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.

В параграфе построена статистика, слабо сходящаяся 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 получения состоятельного критерия построен критерий типа омега-квадрат. Предельное распределение для лежащей в его основе статистики 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

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

Пусть $d > 0$ — целое число. Обозначим

$$\vec{q}_n = \left(\widehat{Z}_n\left(\frac{1}{d+1}\right), \ldots, \widehat{Z}_n\left(\frac{d}{d+1}\right)\right), \vec{q}_F = \left(Z_F^0\left(\frac{1}{d+1}\right), \ldots, Z_F^0\left(\frac{d}{d+1}\right)\right).$$

Обозначим через A ковариационную матрицу вектора $\vec{q}_F$ и ее эмпирический аналог — матрицу $\widehat{A_n} = (\widehat{a}_{ij})_{i,j=1}^d$,где

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

В итоге, 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.

Отметим, что 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 статистик для критериев типа омега-квадрат.

Будем использовать критерий, 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 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[4] 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 (из данного выше определения ω_F^2) 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 может быть найдена в явном виде. В параграфе 3.3[5] 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. Ряд результатов такого рода для предельного распределения статистики хи-квадрат при проверке сложных гипотез получен Г. В. Мартыновым[6].

Также в параграфе разобраны примеры, когда ξ_1 имеет нормальное и сдвинутое показательное распределение. Распределение функционала ω_F^2 в обоих случаях найдено и табулировано Мартыновым[7].

Поговорим теперь о применении теоремы 2. Из нее вытекает следующее следствие (сохранена нумерация диссертации).

Следствие 3 Если $Y_i = a + bX_i + \varepsilon_i$, где 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 вычислено и табулировано[8], а 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

[4]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.

[5]Там же

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

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

[8]Там же, табл. 3 на с. 65

распределения, соответственно.

Также Мартыновым[9] получены выражения, с помощью которых можно вычислить $F_{\widehat{\eta}}(x)$.

В **параграфе 2.2** показаны возможности применения обсуждаемого в диссертации метода анализа регрессионных зависимостей на примере анализа зависимости курсов евро и доллара. Было рассмотрено две гипотезы: евро следует за долларом или доллар следует за евро с точностью до шума. Все расчеты проведены с помощью пакета для MatLab.

Отметим также, что в процессе изучения данного примера были показаны возможности использования конструкции эмпирического моста для решения известной задачи о разладке. Анализ графиков эмпирического моста позволяет выявить моменты разладки регрессионной модели. А именно, если на каком-то из участков наблюдается стремительный рост или снижение графика, то резонно говорить о разладке в регрессионной модели (смене параметра) в точке экстремального значения эмпирического моста. После этого выборка может быть разбита на ряд кусков (в точках экстремальных значений моста), на каждом из которых строится 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.

Описанный метод и был использован при анализе курсов валют, что привело к дроблению исходной выборки на три непересекающихся последовательных интервала с разной зависимостью курсов на каждом из них. В результате с помощью применения критерия хи-квадрат была выбрана модель, наиболее согласующаяся с эмпирическими данными: 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).

В **параграфе 2.3** исследована зависимость массы человеческого тела от его роста. В качестве исходной была взята биометрическая двумерная выборка объемом 750 наблюдений, содержащая сведения о росте (в сантиметрах) и массе тела (в килограммах) студенток первого курса лечебного факультета ГБОУ ВПО «Волгоградский государственный медицинский университет». Было рассмотрено 12 вариантов регрессионных моделей, описывающих исследуемую зависимость. Для проведения расчетов,

[9]Там же.

помимо упомянутого пакета MatLab, использовался пакет R.

В результате с помощью применения критерия омега-квадрат была выбрана модель, наиболее согласующаяся с эмпирическими данными: логарифм роста пропорционален логарифму веса с коэффициентом пропорциональности равным 2.

В **параграфе 2.4** была проверена гипотеза о наличии линейной зависимости длины прыжка с места от роста человека. В качестве исходной была взята биометрическая двумерная выборка объемом 743 наблюдения, содержащая сведения о росте (в сантиметрах) и длине прыжка с места (в сантиметрах) студенток первого курса лечебного факультета ГБОУ ВПО «Волгоградский государственный медицинский университет». Для проведения расчетов, как и в параграфе 2.3 применялся пакет MatLab.

В результате с помощью применения критерия омега-квадрат была принята гипотеза о наличии линейной зависимости исследуемых величин.

Заключение содержит список основных результатов, полученных в работе, которые изложены в разделах основных результатов и научной новизны настоящего автореферата.

В приложении приведены графики эмпирических мостов для моделей из параграфа 2.3 диссертационной работы.

Работы автора по теме диссертации

Статьи в журналах, рекомендованных ВАК:

[1] 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).

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

[3] 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). – 2015. – С. 35–47. 91–97. DOI 10.17223/19988621/37/3 (входит в РИНЦ).

[4] 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 Markov-modulated noise // Probability and Mathematical Statistics, Vol. 36.1. – 2016. – P. 113–120. (входит в Scopus).

Другие публикации:

[5] 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, – 2011. – С. 205.

[6] 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, Novosibirsk: Sobolev Institute of Mathematics, – 2011. – P. 26.

[7] 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, – 2013. – С. 573-574.

[8] 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, – 2014. – С. 241.

[9] 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). – 2014. – P. 37-38.