А.П. Ковалевский¹, Е.В. Шаталин²**ВЫБОР РЕГРЕССИОННОЙ МОДЕЛИ ЗАВИСИМОСТИ МАССЫ ТЕЛА ОТ РОСТА С ПОМОЩЬЮ ЭМПИРИЧЕСКОГО МОСТА**

Рассмотрен ряд регрессионных моделей зависимости массы тела от роста. Проведено сравнение моделей между собой с помощью конструкции эмпирического моста. В качестве исходных использованы данные Волжского государственного медицинского университета о росте и массе тела студенток первого курса.

Ключевые слова: линейная регрессия, эмпирический мост, зависимость массы тела от роста.

1. Введение и предварительные сведения

Для исследования зависимости массы тела человека W_i от роста H_i в [1] предложена модель пропорциональности массы тела квадрату роста. Отметим, что пропорциональность предполагается для лиц одной возрастной группы и одного пола. В [2] на основании масштабных исследований показано, что эта зависимость является наилучшей в классе степенных зависимостей. Эту пропорциональность можно проинтерпретировать в виде двух различных регрессионных моделей: $\ln W_i = \ln(\alpha H_i^2) + \varepsilon_i$ и $W_i = a + bH_i^2 + \varepsilon_i$. Отметим, что рост и массу тела индивидуума можно считать случайными величинами и предполагать независимость роста H_i и корректирующего фактора ε_i .

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

Если модель неправильно описывает данные, то значения Y_i будут систематически уклоняться от регрессионной кривой, и это отклонение можно выявить суммированием регрессионных остатков (разностей между наблюдаемыми и прогнозируемыми значениями). Для изучения значимости этих отклонений необходимо знать предельное распределение процесса центрированных и самонормированных частичных сумм регрессионных остатков. Этот процесс называется эмпирическим мостом [3]. Нормировка, 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 п.н. (как в изучаемых ниже моделях) центрирования не требуется. Однако общее определение оставлено в статье для использования в других моделях.

¹ Работа выполнена при частичной финансовой поддержке РФФИ (проект 13-01-00661)

² Работа выполнена при финансовой поддержке правительства Новосибирской области (стипендия правительства НСО)

Если для описания данных предложено несколько моделей, то вычисление достигнутых уровней значимости позволяет выбрать модель, наилучшим образом описывающую данные. Подход к анализу соответствия данных вероятностным моделям, основанный на функционалах от эмпирического моста, разрабатывался в [3–5] и применялся к анализу текстов в [3], тестированию моделей цен на недвижимость и автомобили в [6, 7], поиску неоднородностей строительных конструкций в [8].

Предлагаемый авторами подход (в отличие от использования коэффициента детерминации) позволяет сравнивать модели с разным числом параметров. В работе [5] рассмотрен пример, для которого выбором констант коэффициент детерминации может быть сделан сколь угодно близким к 1, а реально достигаемый уровень значимости (для критерия, использующего конструкцию эмпирического моста) с ростом объема выборки стремится к нулю п.н. Этот пример 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, и количество параметров, а также характер зависимости от них влияют только на распределение предельного процесса, на основании которого вычисляется реально достигаемый уровень значимости. Таким образом, F-тест может принимать неправильную модель, а критерий эмпирического моста отвергать ее.

Эмпирический мост для модели выборки слабо сходится к стандартному броуновскому мосту, а сходимость эмпирического моста в модели двухпараметрической линейной регрессии требует доказательства. При этом предельный гауссовский процесс отличается от стандартного броуновского моста.

Рассмотрим две вероятностные модели (одно- и двухпараметрическую):

$$Y_i = \theta + \varepsilon_i, \quad i = 1, \dots, n, \quad n \geq 1,$$

$$Y_i = a + bX_i + \varepsilon_i, \quad i = 1, \dots, n, \quad n \geq 1,$$

где $\theta, a, b \in R$ – неизвестные параметры регрессии, $\varepsilon_1, \dots, \varepsilon_n$ (регрессионные ошибки) – независимые одинаково распределенные случайные величины с нулевым математическим ожиданием и конечной ненулевой дисперсией σ^2 . Также предполагается, что $X_i = \xi_{i,n}$, $i = 1, \dots, n$, – порядковые статистики, где случайные величины $\xi_1, \dots, \xi_n$ независимы, одинаково распределены с функцией распределения F и не зависят от случайных величин $\varepsilon_1, \dots, \varepsilon_n$.

Неизвестные параметры регрессии обычно оценивают по методу наименьших квадратов, получая оценки $\hat{\theta}, \hat{a}, \hat{b}$. На основании регрессионной модели строятся прогнозные значения $\hat{Y}_i = \hat{\theta}$, $\hat{Y}_i = \hat{a} + \hat{b}X_i$. Остатками линейной регрессии называют случайные величины $\hat{\varepsilon}_i = Y_i - \hat{Y}_i$.

Приведем определение эмпирического моста. Эмпирический мост – это кусочно-но-линейная случайная ломаная $Z_n^0 = \{Z_n^0(t), 0 \leq t \leq 1\}$ с узлами в точках

$$\left(\frac{k}{n}, \frac{\hat{\Delta}_k - \frac{k}{n} \hat{\Delta}_n}{\sqrt{\hat{\sigma}^2 n}} \right),$$

где $\hat{\Delta}_k = \hat{\varepsilon}_1 + \dots + \hat{\varepsilon}_k$, $k = 1, \dots, n$, $\hat{\Delta}_0 = 0$, $\hat{\sigma}^2 = \overline{\hat{\varepsilon}^2} - (\bar{\hat{\varepsilon}})^2$.

При условии сходимости $\widehat{\sigma}^2 \xrightarrow{p} \sigma^2$ при $n \rightarrow \infty$ 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, 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

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

Пусть $GL_F(t) = \int_0^t F^{-1}(s) ds$ – теоретическая обобщенная кривая Лоренца (см.

в [9]), $F^{-1}(s) = \sup \{x : F(x) < s\}$ – квантильное преобразование (обобщенная обратная функция) функции распределения $F(x)$, $GL_F^0(t) = GL_F(t) - tGL_F(1)$.

Через $\Rightarrow$ будем обозначать слабую сходимость (сходимость по распределению) в соответствующем пространстве. Так, в теоремах ниже через $\Rightarrow$ обозначена слабая сходимость в пространстве $C(0,1)$, снабженном равномерной метрикой (см. [10, с. 82]).

Из функциональной предельной теоремы (принципа инвариантности) следует (см. [3]) теорема 1.

Теорема 1. Если $Y_i = \theta + \varepsilon_i$, то $Z_n^0 \Rightarrow W^0$, где W^0 – стандартный броуновский мост с ковариационной функцией

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

Следствие 1. Если $Y_i = \theta + \varepsilon_i$, то

$$\int_0^1 (Z_n^0(t))^2 dt \Rightarrow \eta,$$

где η имеет распределение ω^2 , которое представлено в табл. 6.4а на с. 348 в [11] и в табл. 1 на с. 63–64 в [12].

В приложении приведено доказательство следующей теоремы, которая распространяет на двухпараметрическую (парную) линейную регрессию результат работы [5]. Эта теорема также может быть получена применением другой техники доказательства из теоремы 1.2 и следствия 1.3 работы [13]. Для неслучайного регрессора подобные теоремы получены в работах [14, 15].

Теорема 2. Если $Y_i = a + bX_i + \varepsilon_i$, $0 < \text{Var} \xi_1 < \infty$, то $Z_n \Rightarrow Z_F, Z_n^0 \Rightarrow Z_F^0$, где Z_F – центрированный гауссовский процесс с ковариационной функцией

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

Следствие 2. Если $Y_i = a + bX_i + \varepsilon_i$, X_i – порядковые статистики, построенные по выборке из нормального распределения, то

$$\int_0^1 (Z_n^0(t))^2 dt \Rightarrow \hat{\eta} = \int_0^1 Z_\Phi^2(t) dt,$$

где $\hat{\eta}$ имеет распределение $\hat{\omega}^2$, которое представлено в табл. 3 на с. 65 в [12], Z_Φ – центрированный гауссовский процесс с ковариационной функцией

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

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

В [12] получены следующие выражения:

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

(с. 56), где λ_k – собственные числа ядра $\hat{K}$, $\lambda_1 \dots \lambda_{10}$ приведены на с. 37 в [12], для остальных λ_k справедлива эквивалентность $\lambda_k \sim ((2k-1)\pi)^2$, $k > 10$;

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

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

$$I_1(\mu) = \int_0^{1/2} \kappa_1(x) \sin \mu x dx \int_x^{1/2} \kappa_1(x) \cos \mu(1/2 - t) dt;$$

$$\kappa_1(x) = \varphi(\Phi^{-1}(x)).$$

(с. 34–37 в [12]). С помощью данных формул можно вычислить $F_{\hat{\eta}}(x)$, результат приведен в [12, табл. 3].

2. Основные результаты работы

Основной задачей настоящей статьи является применение вышеизложенных теорем для анализа различных регрессионных моделей зависимости массы тела от роста индивидуума.

В качестве исходных данных были взяты сведения о росте (в см) и массе тела (в кг) студентов первого курса Волжского государственного медицинского университета (двумерная выборка объема 750) <http://www.volgmed.ru/ru/>

Предполагается, что есть некоторый статистический ансамбль наблюдений, иными словами, двумерное распределение индивидуумов по росту и весу. Из этого двумерного распределения осуществляется выборка объема n . 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 (конкомитанты) [16].

Мы рассмотрим следующие модели:

$$W_i = \theta + H_i + \varepsilon_i; \quad (1)$$

$$\ln W_i = a + \ln H_i + \varepsilon_i; \quad (2)$$

$$\ln W_i = a + 1.5 \ln H_i + \varepsilon_i; \quad (3)$$

$$\ln W_i = a + 2 \ln H_i + \varepsilon_i; \quad (4)$$

$$\ln W_i = a + 2.5 \ln H_i + \varepsilon_i; \quad (5)$$

$$\ln W_i = a + 3 \ln H_i + \varepsilon_i; \quad (6)$$

$$\ln W_i = a + b \ln H_i + \varepsilon_i; \quad (7)$$

$$W_i = a + bH_i + \varepsilon_i; \quad (8)$$

$$W_i = a + bH_i^{1.5} + \varepsilon_i; \quad (9)$$

$$W_i = a + bH_i^2 + \varepsilon_i; \quad (10)$$

$$W_i = a + bH_i^{2.5} + \varepsilon_i; \quad (11)$$

$$W_i = a + bH_i^3 + \varepsilon_i. \quad (12)$$

Модели (1) – (6) являются однопараметрическими, а модели (7) – (12) двухпараметрическими.

С помощью статистического пакета R на основании критерия Шапиро – Уилка была проверена гипотеза нормальности регрессоров, входящих в модели (1) – (12). Реально достигнутые уровни значимости 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. Так как используемые нами методы основаны на нормальности регрессора, модели (11), (12) далее мы рассматривать не будем.

Статистический критерий показывает, что нормальное приближение хорошо работает и для распределения роста H_i , 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.

С помощью пакета MatLab для каждой из моделей были оценены входящие в них параметры и вычислены значения эмпирического моста в узловых точках. Для сравнения моделей вычисляются выборочные дисперсии остатков, затем статистику

$$\omega_n^2 = \int_0^1 (Z_n^0(t))^2 dt \text{ по формуле}$$

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

которая напрямую следует из определения эмпирического моста, и реально достигаемые уровни значимости (по табл. 1 из [12] для однопараметрических моделей, по табл. 3 из [12] – для двухпараметрических). Результаты приведены в табл. 1.

Таблица 1

Характеристики моделей

Модель	$\widehat{\sigma^2}$	ω_n^2	РДУЗ
1	66.82	7.96	$< 10^{-5}$
2	0.0176	3.316	$< 10^{-5}$
3	0.0171	0.4503	0.053
4	0.0174	0.5928	0.024
5	0.0185	3.591	$< 10^{-5}$
6	0.0203	8.643	$< 10^{-5}$
7	0.0171	0.2697	0.0052
8	57.22	0.2273	0.013
9	57.17	0.2172	0.016
10	57.14	0.2101	0.018

Табл. 1 позволяет сравнить модели между собой. В частности, можно сделать вывод о том, что наилучшей из рассматриваемых является модель 3. Отдельно отметим интересный эффект. Модель 3 лучше с точки зрения критерия согласия ω^2 нежели модель 7, в которой параметр модели b точно оценен. Получается, что в данном случае лучше угадать параметр модели, чем его оценивать. Конечно, 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.

Как показывает табл. 1, ни одна из рассмотренных моделей не демонстрирует высоких реально достигаемых уровней значимости, то есть хорошего соответствия с исследуемыми данными. Поэтому на следующем этапе исследования анализируются выбросы исходных данных относительно предлагаемых моделей и их влияние на изучаемые характеристики. Для исследования выбросов изобразим исходные данные графически (рис. 1).

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

На графике видны выбросы (аномально большие отклонения от любой из предлагаемых регрессионных зависимостей), которые могут привести к искажению результатов. Для устранения данного недостатка мы многократно провели процедуру очистки выборки с помощью правила «трех сигм». Каждый раз, когда несколько значений удалялось, оценки параметров и дисперсии остатков пересчитывались, снова проверялась нормальность выборки 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. 2 (модель 10 исключена, так как на очередном шаге выборка не прошла проверку на нормальность).

Таблица 2

Характеристики моделей после удаления выбросов

Модель	Итераций	К-во удаленных	$\hat{\sigma}^2$	ω_n^2	РДУЗ
1	4	14	53.07	6.87	$< 10^{-5}$
2	2	9	0.0158	4.35	$< 10^{-5}$
3	2	8	0.0158	0.8133	0.0068
4	4	10	0.0149	0.2718	0.165
5	2	9	0.0164	2.84	$< 10^{-5}$
6	2	8	0.0177	7.64	$< 10^{-5}$
7	3	9	0.0151	0.1741	0.0412
8	1	11	47.14	0.1605	0.0563
9	4	20	42.27	0.1674	0.047

Наилучший результат показала модель 4, эмпирический мост регрессионных остатков для нее приведен на рис. 2.

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

Эту модель (модель 4) и следует использовать для анализа отклонений массы тела от нормы. Оценка параметра a равна $\hat{a} = -6,2171\dots$. 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 см.

В частности, исключенные 10 наблюдений являются примерами таких отклонений, значимых на критическом уровне 0.0027 (согласно правилу «трех сигм»).

Приложение. Доказательство теоремы 2

Для простоты восприятия разобьем доказательство на пять последовательных шагов. На первом шаге доказательства покажем, что сумму $\sum_{i=1}^n \frac{\varepsilon_i^0 X_i^0}{\sqrt{n}}$ можно заменить суммой $\sum_{i=1}^n \frac{\varepsilon_i^0 EX_i^0}{\sqrt{n}}$ (величины ε_i^0 и X_i^0 будут введены ниже). На втором шаге докажем слабую сходимость нормированного вектора с координатами $(\hat{\Delta}_{k_1}, \dots, \hat{\Delta}_{k_m})$ к нормированному вектору с координатами $(\Delta_{k_1}, \dots, \Delta_{k_m})$, где Δ_{k_i} будут определены позднее. Третий шаг состоит в доказательстве слабой сходимости соответствующих конечномерных распределений. На четвертом шаге будет показана относительная компактность семейства вероятностных распределений, порожденного семейством $\{Z_n(t), 0 \leq t \leq 1\}$. Последний, пятый шаг, состоит из доказательства сходимости по вероятности выборочной дисперсии $\hat{\sigma}^2$ к истинной дисперсии σ^2 .

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

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

Покажем, что

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

Так как $EX_i^0 = 0$, случайные величины $\{\varepsilon_i\}$ независимы между собой и не зависят от $\{\xi_{i:n}\}$, имеем

$$P \left\{ \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i^0 (X_i^0 - EX_i^0) \right| \geq \delta \right\} \leq \frac{\text{Var} \sum_{i=1}^n \varepsilon_i^0 (X_i^0 - EX_i^0)}{n\delta^2} = \frac{\sigma^2 \sum_{i=1}^n \text{Var} X_i^0}{n\delta^2}.$$

Заметим, что

$$\begin{aligned} \sum_{i=1}^n 2 |\text{cov}(X_i, \bar{X})| &\leq \sum_{i=1}^n 2 \sqrt{\text{Var} X_i \text{Var} \bar{X}} \leq \sum_{i=1}^n 2 \left(\frac{1 + \text{Var} X_i}{2} \right) \sqrt{\frac{\text{Var} X_i}{n}} = \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^n (\sqrt{\text{Var} X_i} (1 + \text{Var} X_i)) \leq n^{-1/2} \sum_{i=1}^n \frac{1 + \text{Var} X_i}{2} + n^{-1/2} \left(\sum_{i=1}^n \text{Var} X_i \right)^{3/2}. \end{aligned}$$

В силу теоремы 1 в [17], $\frac{1}{n} \sum_{i=1}^n \text{Var } X_i \rightarrow 0$ при $n \rightarrow \infty$. Поэтому и $\frac{1}{n} \sum_{i=1}^n \text{Var } X_i^0 \rightarrow 0$, и тем самым (14) доказано.

Шаг 2. Пусть $[t]$ обозначает целую часть числа t . Для каждого m и каждого $0 \leq s_1 < \dots < s_m \leq 1$, $k_i = [ns_i]$, покажем слабую сходимость при $n \rightarrow \infty$ вектора $\bar{\eta} = \frac{1}{\sigma\sqrt{n}}(\hat{\Delta}_{k_1}, \dots, \hat{\Delta}_{k_m})$ к вектору $\bar{Z}_F = (Z_F(s_1), \dots, Z_F(s_m))$.

Из (13), (14) и сходимостей $\overline{(X^0)^2} \rightarrow \text{Var } \xi_1$ п.н., $\frac{1}{n} \sum_{i=1}^{k_j} X_i^0 \rightarrow GL_F^0(s_j)$ п.н. (см. [18]), вектор $\bar{\eta}$ может быть заменен вектором $\bar{\zeta} = \frac{1}{\sigma\sqrt{n}}(\Delta_{k_1}, \dots, \Delta_{k_m})$, где

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

Шаг 3. Покажем слабую сходимость $\bar{\zeta} \Rightarrow \bar{Z}_F$, используя метод характеристических функций. Заметим, что

$$\sum_{j=1}^m t_j \left(\sum_{i=1}^{k_j} (\varepsilon_i - \bar{\varepsilon}) - \frac{GL_F^0(s_j)}{\text{Var } \xi_1} \sum_{i=1}^n \varepsilon_i EX_i^0 \right) = \sum_{i=1}^n \varepsilon_i \sum_{j=1}^m t_j \left(I\{i \leq k_j\} - \frac{k_j}{n} - \frac{GL_F^0(s_j)}{\text{Var } \xi_1} EX_i^0 \right).$$

Из предложения 4 в [5], используя неравенство Гельдера, имеем $EX_i^0 = o(\sqrt{n})$ равномерно по всем $1 \leq i \leq n$.

Определим $\beta_i = \sum_{j=1}^m t_j \left(I\{i \leq k_j\} - \frac{k_j}{n} - \frac{GL_F^0(s_j)}{\text{Var } \xi_1} EX_i^0 \right)$. Тогда характеристическая функция $\varphi_{\bar{\zeta}}(\vec{t})$ может быть представлена как

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

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

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

$$\begin{aligned} \sum_{i=1}^n \frac{\beta_i^2}{n} &= \sum_{i=1}^n \frac{1}{n} \left(\sum_{j=1}^m t_j \left(I\{i \leq k_j\} - \frac{k_j}{n} - \frac{GL_F^0(s_j)}{\text{Var } \xi_1} EX_i^0 \right) \right)^2 = \\ &= \sum_{j_1=1}^m \sum_{j_2=1}^m \sum_{i=1}^n \frac{1}{n} t_{j_1} t_{j_2} \left(I\{i \leq k_{j_1}\} - \frac{k_{j_1}}{n} - \frac{GL_F^0(s_{j_1})}{\text{Var } \xi_1} EX_i^0 \right) \times \end{aligned}$$

$$\begin{aligned} & \times \left(I\{i \leq k_{j_2}\} - \frac{k_{j_2}}{n} - \frac{GL_F^0(s_{j_2})}{\text{Var } \xi_1} EX_i^0 \right) \\ & \rightarrow \sum_{j_1=1}^m \sum_{j_2=1}^m t_{j_1} t_{j_2} \left(\min(s_{j_1}, s_{j_2}) - s_{j_1} s_{j_2} - \frac{GL_F^0(s_{j_1}) GL_F^0(s_{j_2})}{\text{Var } \xi_1} \right). \end{aligned}$$

Таким образом,

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

и

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

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

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

Обозначим $S_k = \sum_{i=1}^k \xi_{i:n}$, $k = 1, \dots, n$, $S_0 = 0$.

В соответствии с теоремой Прохорова (см. гл. 16 в [10]) достаточно показать, что семейство распределений случайных процессов последовательности

$\left\{ \frac{\hat{\Delta}_{[nt]}}{\sigma \sqrt{n}}, 0 \leq t \leq 1 \right\}$ плотно. Обозначим $k = [nt]$.

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

Тогда
$$\hat{\Delta}_k = \hat{\Delta}_k^0 - \frac{k}{n} \hat{\Delta}_n^0.$$

Плотность семейства $\left\{ \frac{\sum_{i=1}^k \varepsilon_i}{\sigma \sqrt{n}}, 0 \leq t \leq 1 \right\}$ была показана в [10] при доказатель-

стве принципа инвариантности Донскера – Прохорова.

Поэтому докажем только плотность семейства

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

В силу теоремы 8.3 в [10] достаточно доказать, что для любых $\varepsilon > 0$, $\alpha > 0$ найдутся $0 < \delta < 1$, $n_0 \in \mathbb{N}$ такие, что

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

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

Заметим, что

$$\frac{\overline{X^0 \varepsilon^0} \sqrt{n}}{\sigma(X^0)^2} \Rightarrow \frac{\zeta}{\sqrt{\text{Var } \xi_1}}$$

и (см. [18])

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

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

Обозначим

$$A(t, \delta) = \sup_{t \leq s \leq t+\delta} |GL_F(s) - GL_F(t)|.$$

Покажем, что для любых $\varepsilon > 0$, $\alpha > 0$ найдется $0 < \delta < 1$ такое, что

$$P \left\{ \frac{|\zeta|}{\sqrt{\text{Var } \xi_1}} A(t, \delta) \geq \varepsilon \right\} \leq \frac{\alpha \delta}{2}, \quad 0 \leq t \leq 1.$$

Левая часть последнего неравенства не превосходит $2 \frac{e^{-x^2/2}}{x\sqrt{2\pi}}$, где $x = \frac{\varepsilon \sqrt{\text{Var } \xi_1}}{A(t, \delta)}$, в силу леммы 2 (см. [19, с. 175]). С помощью неравенства Коши – Буняковского имеем

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

Очевидно, что требуемое δ всегда найдется.

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

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

Это завершает доказательство теоремы 2.

ЛИТЕРАТУРА

1. *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. 1832. V. 7. P. 1–83.
2. *Keys A., Fidanza F., Karvonen M. J., Kimura N., Taylor H. L.* Indices of relative weight and obesity // Journal of Chronic Diseases. 1972. V. 25. No. 6–7. P. 329–343.
3. *Гусарова Г. В., Ковалевский А. П., Макаренко А. Г.* Критерии наличия разладки // Сиб. журн. индустр. матем. 2005. Т. 8. № 4. С. 18–33.
4. *Ковалевский А.П.* Статистические критерии обнаружения разладки регрессии с циклическим трендом // Научный вестник НГТУ. 2013. № 3 (52). С. 55–62.
5. *Ковалевский А.П., Шаталин Е.В.* Асимптотика сумм остатков однопараметрической линейной регрессии, построенной по порядковым статистикам // Теория вероятностей и ее применения. 2014. Т. 59. № 3. С. 452–467.

6. Аркашов Н.С., Ковалевский А.П. Вероятностная модель цен на квартиры // Сиб. журн. индустр. матем. 2012. Т. 15. № 2. С. 11–20.
7. Kovalevskii A. A regression model for prices of second-hand cars // Applied methods of statistical analysis. Applications in Survival Analysis, Reliability and Quality Control. Novosibirsk, 2013. P. 124–128.
8. Ковалевский А.П., Шахраманьян А.М. Анализ дефектов строительных конструкций методом эмпирического моста // Научный вестник НГТУ. 2014. Т. 56. № 3. с. 171–180.
9. Gastwirth J.L. A general definition of the Lorenz curve // Econometrica. 1971. V. 39. P. 1037–1039.
10. Биллингсли П. Сходимость вероятностных мер. М.: Наука, 1977.
11. Большев Л.Н., Смирнов Н.В. Таблицы математической статистики. М.: Наука, 1983.
12. Мартынов Г.В. Критерии омега-квадрат. М.: Наука, 1978.
13. Stute W. Nonparametric model checks for regression // Ann. Statist. 1997. V. 25. P. 613–641.
14. MacNeill I.B. Limit processes for sequences of partial sums of regression residuals // Ann. Prob. 1978. V. 6. No. 4. P. 695–698.
15. Bischoff W. A functional central limit theorem for regression models // Ann. Stat. 1998. V. 26. P. 1398–1410.
16. Davydov Y., Zitakis R. Convex rearrangements of random elements // Fields Institute Communications. 2004. V.44. P. 141–171.
17. Hoeffding W. On the distribution of the expected values of the order statistics // Ann. Math. Statist. 1953. V. 24. No. 1. P. 93–100.
18. Goldie C.M. Convergence theorems for empirical Lorenz curves and their inverses // Advances in Applied Probability. 1977. V. 9. P. 765–791.
19. Феллер В. Введение в теорию вероятностей и ее приложения. Т. 1. М.: Мир, 1984.

Статья поступила 15.04.2015 г.

Kovalevskii A.P., Shatalin E.V. THE CHOICE OF A REGRESSION MODEL OF THE BODY WEIGHT ON THE HEIGHT VIA AN EMPIRICAL BRIDGE

DOI 10.17223/19988621/37/3

An empirical bridge can be used for analysis of correspondance between regression models and observed data. If a model does not describe data correctly, then its response values deviate systematically from the regression curve, and this deviation can be revealed by summing the regression residuals. One needs to know the limiting distribution of the process for centered and normalized partial sums of regression residuals to study significance of these deviations. This process is the empirical bridge. We obtain a limiting process for a simple linear regression model. The main goal of this article is to apply the empirical bridge for the analysis of regression models describing the dependence of an individual's body weight on his height. We considered a number of regression models of this dependence and compared models based on their empirical bridges. We used data on the height and weight of female students of the first course of Volga State Medical University. The study revealed the best model $\ln W_i = a + 2 \ln H_i + \varepsilon_i$. This model should be used for analysis of deviations from the normal body weight.

Keywords: linear regression, empirical bridge, dependence of weight on height.

KOVALEVSKII Artem Pavlovich (Candidate of Physics and Mathematics,
Novosibirsk State Technical University, Novosibirsk,
Novosibirsk State University, Novosibirsk, Russian Federation)
E-mail: pandorra@ngs.ru

SHATALIN Evgeny Viktorovich (Sobolev Institute of Mathematics,
Novosibirsk, Russian Federation)
E-mail: sh_e_v_89@list.ru

REFERENCES

1. Quetelet A. Recherches sur le poids de l'homme aux different ges. *Nouveaux Memoire de l'Academie Royale des Sciences et Belles-Lettres de Bruxelles*, 1832, vol. 7, pp. 1–83.
2. Keys A., Fidanza F., Karvonen M. J., Kimura N., Taylor H. L. Indices of relative weight and obesity. *Journal of Chronic Diseases*, 1972, vol. 25. No. 6–7, pp. 329–343.
3. Gusarova G.V., Kovalevskiy A.P., Makarenko A.G. Kriterii nalichiya razladki. *Sib. zhurn. industr. matem.*, 2005, vol. 8, no. 4, pp. 18–33. (in Russian)
4. Kovalevskiy A.P. Statisticheskie kriterii obnaruzheniya razladki regressii s tsiklicheskim trendom. *Nauchnyy vestnik NGTU*, 2013, no. 3 (52), pp. 55–62. (in Russian)
5. Kovalevskiy A.P., Shatalin E.V. Asimptotika summ ostatkov odnoparametricheskoy lineynoy regressii, postroennoy po poryadkovym statistikam. *Teoriya veroyatnostey i ee primeneniya*, 2014, vol. 59, no. 3, pp. 452–467. (in Russian)
6. Arkashov N.S., Kovalevskiy A.P. Veroyatnostnaya model' tsen na kvartiry. *Sib. zhurn. industr. matem.*, 2012, vol. 15, no. 2. C. 11–20. (in Russian)
7. Kovalevskii A. A regression model for prices of second-hand cars. *Applied methods of statistical analysis. Applications in Survival Analysis, Reliability and Quality Control*. Novosibirsk, 2013, pp. 124–128.
8. Kovalevskiy A.P., Shakhraman'yan A.M. Analiz defektov stroitel'nykh konstruksiy metodom empiricheskogo mosta. *Nauchnyy vestnik NGTU*, 2014, vol. 56, no. 3. pp. 171–180. (in Russian)
9. Gastwirth J.L. A general definition of the Lorenz curve. *Econometrica*, 1971, vol. 39, pp. 1037–1039. (in Russian)
10. Billingsli P. *Skhodimost' veroyatnostnykh mer*. Moscow, Nauka Publ., 1977.
11. Bol'shev L.N., Smirnov N.V. *Tablitsy matematicheskoy statistiki*. Moscow, Nauka Publ., 1983.
12. Martynov G.V. *Kriterii omega-kvadrat*. Moscow, Nauka Publ., 1978.
13. Stute W. Nonparametric model checks for regression. *Ann. Statist.*, 1997, vol. 25, pp. 613–641.
14. MacNeill I.B. Limit processes for sequences of partial sums of regression residuals. *Ann. Prob.*, 1978, vol. 6. No. 4, pp. 695–698.
15. Bischoff W. A functional central limit theorem for regression models. *Ann. Stat.*, 1998, vol. 26, pp. 1398–1410.
16. Davydov Y., Zitikis R. Convex rearrangements of random elements. *Fields Institute Communications*, 2004, vol. 44, pp. 141–171.
17. Hoeffding W. On the distribution of the expected values of the order statistics. *Ann. Math. Statist.*, 1953, vol. 24. No. 1, pp. 93–100.
18. Goldie C.M. Convergence theorems for empirical Lorenz curves and their inverses. *Advances in Applied Probability*, 1977, vol. 9, pp. 765–791.
19. Feller V. *Vvedenie v teoriyu veroyatnostey i ee prilozheniya*. Moscow, Mir Publ., 1984, vol. 1. (in Russian)