БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

УДК 519.22

АГЕЕВА
Елена Сергеевна

СТАТИСТИЧЕСКИЕ ВЫВОДЫ О РЕГРЕССИОННЫХ МОДЕЛЯХ ПРИ НАЛИЧИИ КЛАССИФИКАЦИИ НАБЛЮДЕНИЙ

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата физико-математических наук

по специальности 01.01.05 — теория вероятностей
и математическая статистика

Минск, 2017

Работа выполнена в Белорусском государственном университете.

Научный руководитель — **Харин Юрий Семенович**, доктор физико-математических наук, профессор, член-корреспондент НАН Беларуси, заведующий кафедрой математического моделирования и анализа данных Белорусского государственного университета.

Официальные оппоненты: **Янович Леонид Александрович**, доктор физико-математических наук, профессор, член-корреспондент НАН Беларуси, главный научный сотрудник отдела нелинейного и стохастического анализа ГНУ "Институт математики НАН Беларуси";

Цеховая Татьяна Вячеславовна, кандидат физико-математических наук, доцент, доцент кафедры теории вероятностей и математической статистики Белорусского государственного университета.

Оппонирующая организация — ОУ "Гомельский государственный университет им. Ф. Скорины".

Защита состоится 15 сентября 2017 года в 10.00 на заседании совета по защите диссертаций Д 02.01.07 при Белорусском государственном университете по адресу: 220030, Республика Беларусь, г. Минск, ул. Ленинградская, 8 (корпус юридического факультета), ауд. 407, тел. ученого секретаря: (017) 209-57-09.

С диссертацией можно ознакомиться в Фундаментальной библиотеке Белорусского государственного университета.

Автореферат разослан " " июня 2017 года.

Ученый секретарь совета
по защите диссертаций Д 02.01.07
кандидат физ.-мат. наук доцент Е. М. Радыно

ВВЕДЕНИЕ

Теория статистических выводов исследует математические модели и методы построения статистических оценок параметров, статистических решений, прогнозов на основе наблюдений (статистических данных). Регрессионная модель наблюдений широко используется для описания многих процессов в технике, экономике, медицине, геонауке и других приложениях. Классическая регрессионная модель хорошо изучена в математической статистике: для нее построены статистические оценки параметров, критерии для статистической проверки гипотез согласия и гипотез о коэффициентах линейной регрессии. Однако на практике часто наблюдаются различные отклонения от этой хорошо изученной модели, которые требуется учитывать при построении статистических выводов. Для того чтобы преодолеть затруднения при статистическом анализе данных, связанные с наличием в выборке выбросов, пропусков, искажений, огрублений, требуется разработка специальных сложных моделей данных, примерами которых являются модели с гетероскедастичностью, модели с засорениями Тьюки-Хьюбера, модели с цензурированием или пропусками. Развитием этого направления в математической статистике, получившего название "робастная статистика" (англ. robust — "крепкий", "сильный", "устойчивый"), занимаются P. J. Huber, F. Hampel, P. J. Rousseeuw, E. M. Ronchetti, U. Gather, P. Filzmoser, С. А. Айвазян и другие ученые. В Беларуси статистическому анализу моделей с неполными данными и другими типами искажений посвящены работы Ю. С. Харина, Е. Е. Жука, М. С. Абрамовича, В. И. Малюгина, А. Ю. Харина, И. А. Бодягина, В. А. Волошко.

Одним из возможных искажений классической модели является группирование наблюдаемых значений. В обзорной статье D. F. Heitjan[1] выделены три типа группирования:

1) округление, то есть замена истинного значения средней точкой интервала, которому оно принадлежит;

2) интервальное цензурирование, например, время обнаружения рецидива заболевания, о котором известно только, что оно наступило между моментами посещений врача;

3) группирование как таковое, то есть огрубление, вообще говоря, непрерывных переменных путем разбиения их на категории (классы) во время сбора данных или их обработки.

В диссертационной работе рассматривается регрессионная модель при наличии особого вида группирования наблюдений. Вместо точного значения зависимой переменной наблюдается только номер одного из заранее заданных непересекающихся числовых промежутков (интервалов), в который это

[1] *Heitjan, D. F.* Inference from Grouped Continuous Data: A Review / D. F. Heitjan // Statistical Science. — 1989. — Vol. 4, Issue 2. — P. 164–183.

значение попало. Такое искажение данных будем называть классификацией. Классификация естественным образом возникает при сборе данных, например, при проведении социологических опросов: респондентам не всегда нужно указывать точный ответ, достаточно только выбрать соответствующий интервал. Другая причина появления классифицированных наблюдений связана с удобством хранения информации, когда точное значение не представляет интерес.

Стоит отметить так называемую интервальную регрессию[2], которую также называют "регрессией при наличии группирования". Эта модель предполагает, что все значения зависимой переменной наблюдаются с точностью до некоторого интервала. Исследуемая в диссертационной работе регрессионная модель при наличии классификации наблюдений подпадает под это определение, однако, интервальная регрессия является более общим случаем, в котором на интервалы не накладываются никакие ограничения, в частности, их неслучайность и зафиксированность. В связи с этим статистический анализ для интервальной регрессии не получил теоретического развития и ограничивается численными методами построения оценок максимального правдоподобия и некоторых их аппроксимаций (см. работы J. Burridge [3] , S. B. Caudill and J. D. Jackson [4] , M. B. Stewart [5]), при этом вероятностные свойства полученных оценок не исследованы.

Таким образом, актуальной задачей является развитие вероятностно-статистического анализа регрессионной модели при наличии классификации наблюдений с учетом ее особенностей.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Связь работы с крупными научными программами (проектами) и темами

Тема диссертации соответствует направлению фундаментальных и прикладных научных исследований "12.1. Физические и математические методы и их применение для решения актуальных проблем естествознания, техники, новых технологий, экономики и социальных наук", определенному Перечнем приоритетных направлений фундаментальных и прикладных научных исследований Республики Беларусь в области естественных, технических, гумани-

[2] *Long, J. S.* Regression Models for Categorical and Limited Dependent Variables Using Stata / J. S. Long, J. Freese. — 2-nd ed. — Stata Press, 2014. — 589 p.

[3] *Burridge, J.* A Note on Maximum Likelihood Estimation for Regression Models using Grouped Data / J. Burridge // J.R. Statist.Soc. Series B (Methodological). — 1981. — Vol. 43, Issue 1. — P. 41–45.

[4] *Caudill, S. B.* Heteroscedasticity and grouped data regression / S. B. Caudill, J. D. Jackson // Southern Economic Journal. — 1993. —Vol. 60, Issue 1. — P. 128–135.

[5] *Stewart, M. B.* On least squares estimation when the dependent variable is grouped / M. B. Stewart // Review of Economic Studies. — 1983. — Vol. 50. — P. 737–753.

тарных и социальных наук на 2011–2015 годы. Результаты диссертационного исследования использованы при выполнении в Белорусском государственном университете и в Учреждении БГУ "Научно-исследовательский институт прикладных проблем математики и информатики" следующих научно-исследовательских работ:

1. НИР "Разработка методов робастного статистического анализа случайных последовательностей и полей при наличии неоднородностей и цензурирования" ГПНИ "Конвергенция", подпрограмма "Математические методы" (2011–2015 гг.), номер госрегистрации 20111047;
2. НИР "Разработка вероятностных моделей, методов и алгоритмов статистического анализа и прогнозирования дискретных временных рядов" ГПНИ "Конвергенция", подпрограмма "Методы математического моделирования сложных систем" (2016–2020 гг.), номер госрегистрации 20162616;
3. НИР "Робастные статистические выводы и их применение в компьютерных системах моделирования и анализа данных" (2011–2015 гг.), номер госрегистрации 20120355;
4. Грант Министерства образования Республики Беларусь 2014, НИР "Регрессионное прогнозирование при наличии классификации зависимой переменной" (2014 г.), номер госрегистрации 20140775;
5. Грант Министерства образования Республики Беларусь 2015, НИР "Робастное оценивание параметров множественной регрессии при наличии классификации зависимой переменной" (2015 г.), номер госрегистрации 20150722;
6. Грант БГУ 2011, НИР "Робастный статистический анализ марковских случайных последовательностей" (2011 г.);
7. Грант БГУ 2012, НИР "Статистический анализ данных сложной структуры" (2012 г.);
8. Грант БГУ 2013, НИР "Малопараметрические вероятностно-статистические модели и их применения в задачах анализа данных" (2013 г.);
9. Грант БГУ 2014, НИР "Вероятностно-статистический анализ марковских зависимостей в дискретных данных" (2014 г.);
10. Грант БГУ 2015, НИР "Статистический анализ дискретных временных рядов и случайных последовательностей" (2015 г.).

Цель и задачи исследования

Цель диссертационной работы: построение статистических выводов (статистических оценок, решающих правил и прогнозирующих статистик) о регрессионных моделях при наличии классификации наблюдений. Для достижения поставленной цели требуется решить следующие основные задачи.

Р1. Установить условия идентифицируемости (разрешимости задачи статистического анализа) регрессионной модели при наличии классификации наблюдений.

Р2. Исследовать вероятностные свойства (состоятельность и асимптотическую нормальность) оценки максимального правдоподобия параметров регрессионной модели при наличии классификации наблюдений.

Р3. Для регрессионной модели при наличии классификации наблюдений построить прогнозирующую статистику и исследовать ее вероятностные характеристики (смещение и среднеквадратический риск).

Р4. Разработать статистические критерии проверки гипотез об истинном значении параметров регрессионной модели при наличии классификации наблюдений и исследовать их вероятностные характеристики.

Р5. Разработать статистические критерии проверки гипотез согласия для регрессионной модели при наличии классификации наблюдений.

Объектом исследования является регрессионная модель при наличии классификации наблюдений, которая часто возникает на практике при решении задач статистического анализа данных. Предмет исследования — оценки максимального правдоподобия параметров моделей, подстановочные прогнозирующие статистики, статистические критерии и их вероятностные характеристики.

Научная новизна

В диссертации развита теория статистического анализа регрессионных моделей по неполным данным: впервые установлены условия идентифицируемости регрессионной модели при наличии классификации наблюдений; построены новые прогнозирующие статистики, новые статистические критерии проверки параметрических гипотез и гипотез согласия; найдены вероятностные характеристики статистических оценок параметров, прогнозирующих статистик и статистических критериев.

Положения, выносимые на защиту

1. Необходимые и достаточные условия идентифицируемости регрессионной модели при наличии классификации наблюдений.
2. Достаточные условия сильной состоятельности и асимптотической нормальности оценки максимального правдоподобия параметров регрессионной модели при наличии классификации наблюдений.
3. Асимптотические выражения смещения и среднеквадратического риска подстановочной прогнозирующей статистики для регрессионной модели при наличии классификации наблюдений.

4. Статистические критерии проверки гипотез об истинных значениях параметров и гипотез согласия функции регрессии с заданным параметрическим семейством для регрессионной модели при наличии классификации наблюдений с оценками вероятностей ошибок I-ого рода и мощностей построенных критериев.

Личный вклад соискателя

Основные результаты диссертационной работы получены соискателем самостоятельно. Научному руководителю в совместных работах принадлежат выбор направлений исследования, предметные постановки задач, обсуждение результатов. Соавтором Рудаковской А.В. получены результаты по другой математической модели, которые не включены в диссертацию.

Апробация диссертации и информация об использовании ее результатов

Основные результаты диссертации были представлены и обсуждались на заседаниях Республиканского научного семинара кафедры математического моделирования и анализа данных БГУ и НИИ прикладных проблем математики и информатики "Математическое моделирование сложных систем, анализ данных и защита информации", научных конференциях студентов и аспирантов БГУ, 3-ей Международной конференции "Modern Stochastics: Theory and Applications" (10–14 сентября 2012 года, Киев, Украина), 11-ой и 12-ой Белорусских математических конференциях (4–9 ноября 2012 года и 5–10 сентября 2016 года, Минск), Международных конференциях по робастной статистике ICORS2013 (8–12 июля 2013, Санкт-Петербург, РФ) и ICORS2014 (10–15 августа 2014, Галле, Германия), 10-ой и 11-ой Международных конференциях "Computer data analysis and modeling" (10–14 сентября 2013 года и 6–10 сентября 2016 года, Минск), Международных конгрессах по информатике: информационные системы и технологии CSIST'2013 и CSIST'2016 (4–7 ноября 2013 года и 24–27 октября 2016 года, Минск), 11-ой Международной Вильнюсской конференции по теории вероятностей и математической статистике (30 июня–3 июля 2014 года, Вильнюс, Литва), 15-ой и 16-ой Международных конференциях "Проблемы прогнозирования и государственного регулирования социально-экономического развития" (23–24 октября 2014 года и 23 октября 2015 года, Минск), 7-ом Международном семинаре "Data Analysis Methods for Software Systems" (3–5 декабря 2015, Друскининкай, Литва).

Результаты диссертации внедрены в учебный процесс на факультете прикладной математики и информатики БГУ (имеется акт внедрения).

Опубликованность результатов диссертации

Основные результаты диссертационной работы опубликованы в 25 научных работах. Из них 4 статьи в научных журналах в соответствии с пунктом 18 Положения о присуждении ученых степеней и присвоении ученых званий в Республике Беларусь (общим объемом 1.68 авторского листа), в том числе одна статья в Lithuanian Mathematical Journal (Impact Factor 0.314). Кроме того, опубликовано 3 статьи в сборниках научных трудов, из которых одна статья в ВАК-овском сборнике научных статей "Экономика. Моделирование. Прогнозирование", 10 статей в сборниках материалов научных конференций и 8 тезисов докладов.

Структура и объем диссертации

Диссертационная работа состоит из перечня условных обозначений, введения, общей характеристики работы, четырех глав, заключения, библиографического списка и приложения. Полный объем диссертации составляет 108 страниц, включая 10 рисунков на 6 страницах, 3 таблицы на 2 страницах, 1 приложение на 2 страницах. Библиографический список содержит 112 наименований, включая собственные публикации соискателя ученой степени.

Во введении обосновывается актуальность решаемых в диссертационной работе задач. В первой главе представлена регрессионная модель при наличии классификации наблюдений; для этой модели установлены условия идентифицируемости и построена оценка максимального правдоподобия параметров модели; проведен аналитический обзор регрессионных моделей, в которых происходит огрубление зависимой переменной. Во второй главе установлены условия, при которых оценка максимального правдоподобия является сильно состоятельной и асимптотически нормальной; в асимптотике растущего объема выборки найдены асимптотические смещение и среднеквадратический риск для подстановочной прогнозирующей статистики. В третьей главе построены статистические критерии проверки простой нулевой гипотезы против простой и сложной альтернативных гипотез об истинном значении параметров регрессионной модели при наличии классификации наблюдений, исследованы их вероятности ошибок первого рода и мощности. Четвертая глава посвящена построению статистических критериев проверки гипотез согласия для исследуемой модели на основе модифицированной χ^2-статистики. В заключении приводятся основные научные результаты диссертации и рекомендации по их практическому использованию. В приложении представлен акт о практическом использовании результатов диссертационной работы в учебном процессе на факультете прикладной математики и информатики БГУ.

ОСНОВНОЕ СОДЕРЖАНИЕ

В **главе 1** решена задача **P1**: установлены необходимые и достаточные условия идентифицируемости регрессионных моделей при наличии классификации наблюдений.

В **разделе 1.1** дано определение регрессионной модели при наличии классификации наблюдений. На вероятностном пространстве $(\Omega, \mathcal{F}, \mathbf{P})$ рассматривается модель множественной регрессии:

$$Y_t = F(X_t; \theta^0) + \xi_t,\ t = 1, \ldots, n, \tag{1}$$

где n — объем выборки; $X_t = (X_{t,1}, \ldots, X_{t,N})' \in \mathbf{X} \subseteq \mathbb{R}^N$ — наблюдаемый N-мерный вектор-столбец регрессоров; $\theta^0 = (\theta^0_1, \ldots, \theta^0_m)' \in \Theta \subset \mathbb{R}^m$ — неизвестный m-мерный вектор-столбец регрессионных параметров; $F(\cdot;\cdot) : \mathbf{X} \times \Theta \to \mathbb{R}$ — известная с точностью до векторного параметра функция регрессии; $Y_t \in \mathbb{R}$ — зависимая переменная; $\xi_t \in \mathbb{R}$ — случайная ошибка наблюдения, распределенная по нормальному закону с нулевым математическим ожиданием и неизвестной дисперсией $0 < \mathbf{D}\{\xi_t\} = (\sigma^0)^2 < +\infty$: $\mathcal{L}\{\xi_t\} = \mathcal{N}(0, (\sigma^0)^2)$. Ошибки наблюдений $\{\xi_t\}_{t=1}^n$ предполагаются независимыми в совокупности случайными величинами. Составной вектор-столбец неизвестных параметров модели (1) обозначим $\delta^0 = ((\theta^0)', (\sigma^0)^2)' \in \Delta \subseteq \mathbb{R}^{m+1}$, где Δ — множество всевозможных значений параметров.

Пусть задано разбиение числовой прямой $\mathbb{R}$ на K $(2 \leq K < +\infty)$ непересекающихся числовых промежутков (интервалов):

$$A_k = (a_{k-1}, a_k],\ k \in \mathbf{K} = \{1, \ldots, K\}, \tag{2}$$

где $-\infty = a_0 < a_1 < \cdots < a_K = +\infty$ — упорядоченный набор границ. Набор интервалов (2) определяет классификацию зависимой переменной Y_t:

$$Y_t \textit{ относится к классу номер } \nu_t,\ \textit{если } Y_t \in A_{\nu_t},\ \nu_t \in \mathbf{K}. \tag{3}$$

Неполнота регистрируемых статистических данных заключается в том, что вместо истинных значений зависимой переменной $Y_1, \ldots, Y_n$ наблюдаются лишь соответствующие номера классов $\nu_1, \ldots, \nu_n \in \mathbf{K}$.

Модель (1)–(3) назовем регрессионной моделью при наличии классификации наблюдений.

В **разделе 1.2** найдены необходимые и достаточные условия регулярности модели (1)–(3). Вводятся обозначения: $\mathbf{P}_A\{\cdot\}$, $\mathbf{E}_A\{\cdot\}$, $\mathbf{D}_A\{\cdot\}$ — символы вероятности, математического ожидания и дисперсии при фиксированном A, $\Phi(\cdot)$ — функция распределения стандартного нормального закона $\mathcal{N}(0,1)$,

$$P(k; \delta, X) = \Phi\left(\frac{a_k - F(X;\theta)}{\sigma}\right) - \Phi\left(\frac{a_{k-1} - F(X;\theta)}{\sigma}\right),$$

где $k \in \mathbf{K}$, $\delta \in \Delta$, $X \in \mathbf{X}$.

В силу модельных предположений (1)–(3) наблюдения $\{\nu_t\}_{t=1}^n$ независимы в совокупности и дискретное распределение вероятностей случайной величины $\nu_t \in \mathbf{K}$ имеет вид:

$$\mathbf{P}_{X_t,\delta}\{\nu_t = k\} = \mathbf{P}_{X_t,\delta}\{Y_t \in A_k\} = P(k;\delta,X_t),\ t = 1,\ldots,n.$$

Модель (1)–(3) называется *идентифицируемой при плане эксперимента* $\mathcal{X} = \{X_1,\ldots,X_n\}$, если для любых $\delta,\ \tilde{\delta} \in \Delta$ распределение вероятностей случайных величин $\nu_1,\ldots,\nu_n$, соответствующее параметру δ, совпадает с распределением вероятностей, соответствующим параметру $\tilde{\delta}$:

$$\forall k_1,\ldots,k_n \in \mathbf{K},\quad \prod_{t=1}^n P(k_t;\delta,X_t) = \prod_{t=1}^n P(k_t;\tilde{\delta},X_t)$$

тогда и только тогда, когда $\delta = \tilde{\delta}$. В противном случае модель называется *неидентифицируемой*.

Теорема 1.1.[3] Если число классов $K > 2$, то для идентифицируемости модели (1)–(3) при плане эксперимента $\mathcal{X}$ необходимо и достаточно, чтобы равенство

$$F(X;\theta) = F(X;\tilde{\theta})$$

выполнялось сразу для всех $X \in \mathcal{X}$ тогда и только тогда, когда $\theta = \tilde{\theta}$.

Теорема 1.2.[3] Если число классов $K = 2$, $A_1 = (-\infty, a]$, $A_2 = (a, +\infty)$, где $a \in \mathbb{R}$ — граничная точка, то для идентифицируемости модели (1)–(3) при плане эксперимента $\mathcal{X}$ необходимо и достаточно, чтобы равенство

$$a - F(X,\theta) = c\,(a - F(X,\tilde{\theta})),\ c > 0,$$

выполнялось сразу для всех $X \in \mathcal{X}$ тогда и только тогда, когда $\theta = \tilde{\theta}$, $c = 1$.

В **разделе 1.3** строится оценка максимального правдоподобия параметров регрессионной модели при наличии классификации наблюдений. **Лемма 1.3** определяет логарифмическую функцию правдоподобия для модели (1)–(3):

$$l(\delta;\mathcal{X},\mathcal{H}) = \sum_{t=1}^n \ln\left(\Phi\left(\frac{a_{\nu_t} - F(X_t;\theta)}{\sigma}\right) - \Phi\left(\frac{a_{\nu_t-1} - F(X_t;\theta)}{\sigma}\right)\right),$$

где $\mathcal{X} = \{X_1,\ldots,X_n\} \in \mathbb{R}^{nN}$, $\mathcal{H} = \{\nu_1,\ldots,\nu_n\} \in \mathbf{K}^n$.

Оценка максимального правдоподобия (ОМП) для модели (1)–(3) является решением следующей экстремальной задачи:

$$\hat{\delta}^n = ((\hat{\theta}^n)', (\hat{\sigma}^n)^2)' :\quad l(\hat{\delta}^n;\mathcal{X},\mathcal{H}) = \sup_{\delta\in\Delta} l(\delta;\mathcal{X},\mathcal{H}). \tag{4}$$

Раздел 1.4 содержит аналитический обзор регрессионных моделей, в которых зависимые переменные наблюдаются не полностью. В частности, рассмотрены особенности порядковых логит- и пробит-моделей, а также моделей с искажениями типа "округление" и "цензурирование" [1].

В **главе 2** решены задачи **P2** и **P3**: найдены условия, при которых ОМП для регрессионных моделей при наличии классификации наблюдений являются сильно состоятельными и асимптотически нормально распределенными при растущем объеме выборки $n \to +\infty$; найдены оценки для смещения и среднеквадратического риска подстановочной прогнозирующей статистики.

Вводятся следующие обозначения: $\mathbf{I}_N$ — единичная матрица порядка N, $0_{N\times M}$ — нулевая $(N \times M)$-матрица, $\mathbb{I}\{A\}$ — индикатор события A.

В **разделе 2.1** исследуется состоятельность ОМП. Отличительной чертой регрессионной модели при наличии классификации наблюдений является неодинаковая распределенность независимых случайных величин $\{\nu_t\}_{t=1}^n$, обусловленная изменением регрессора X_t. В теоремах, доказанных Hoadley [6] и Chao [7], представлены достаточные условия, при которых ОМП будет соответственно состоятельной и сильно состоятельной в случае независимых неодинаково распределенных случайных величин для общей вероятностной модели выборки. Однако эти условия являются трудно проверяемыми и не учитывают особенностей модели (1)–(3), поэтому в **подразделе 2.1.1** доказана теорема, формулирующая достаточные условия сильной состоятельности ОМП $\hat{\delta}^n$ в общем случае [3], которые затем используются в **подразделе 2.1.2** для нахождения достаточных условий, учитывающих особенности модели (1)–(3) и легко проверяемых на практике.

Теорема 2.2.[2,3] Пусть выполнены следующие условия.

SC1. Количество классов больше двух: $K > 2$.

SC2. Множество всевозможных значений параметров $\Delta \subset \mathbb{R}^{m+1}$ — компакт.

SC3. Пространство регрессоров $\mathbf{X} \subset \mathbb{R}^N$ — компакт.

SC4. Функция $F(X;\theta)$ непрерывна на $\mathbf{X} \times \Theta$.

SC5. Для любого $\varepsilon > 0$ существует $\gamma = \gamma(\varepsilon) > 0$ такое, что при любом $\theta \in \Theta$, $|\theta - \theta^0| \geq \varepsilon$, выполнено следующее предельное соотношение:

$$\varliminf_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} \mathbb{I}\left\{\left|F(X_t;\theta^0) - F(X_t;\theta)\right| \geq \gamma\right\} = b,$$

где $b = b(\theta, \theta^0, \gamma, F(\cdot)) > 0$.

---

[6] *Hoadley, B.* Asymptotic properties of the maximum likelihood estimators for the independent not identically distributed case / B. Hoadley // Ann. Math. Statist. — 1971. — Vol. 42, Issue 4. — P. 1977–1991.

[7] *Chao, M. T.* Strong consistency of maximum likelihood estimators when the observations are independent but not identically distributed / M. T. Chao // Studies and Essays presented to Yu-Why Chen on his 60-th Birthday / K. Fan [et al.]. — Math Research Center, Taiwan, 1970. — P. 57–63.

Тогда ОМП $\hat{\delta}^n$ является сильно состоятельной: $\hat{\delta}^n \xrightarrow[n\to\infty]{\mathbf{P}=1} \delta^0$.

Раздел 2.2 посвящен исследованию асимптотической нормальности ОМП. Обозначим: $\Gamma_n(\delta) = \sum_{t=1}^n \mathbf{E}_{X_t,\delta}\{\nabla_\delta \ln P(\nu_t;\delta,X_t)(\nabla_\delta \ln P(\nu_t;\delta,X_t))'\}$ — $((m+1)\times(m+1))$-мерная информационная матрица Фишера для наблюдаемой выборки, $\bar{\Gamma}_n(\delta) = \frac{1}{n}\Gamma_n(\delta)$ — усредненная по плану эксперимента матрица.

Теорема 2.4.[3] Пусть выполнены следующие условия.

A1. ОМП $\hat{\delta}^n$ является состоятельной оценкой вектора параметров δ^0.

A2. Функция $F(X;\theta)$, $X \in \mathbf{X}$, трижды непрерывно дифференцируема по θ.

A3. Для любого $\delta \in \Delta$ функции $F(X;\theta)$, $\dfrac{\partial F(X;\theta)}{\partial\theta_i}$, $\dfrac{\partial^2 F(X;\theta)}{\partial\theta_i\partial\theta_j}$, $\dfrac{\partial^3 F(X;\theta)}{\partial\theta_i\partial\theta_j\partial\theta_s}$, $i,j,s = 1,\ldots,m$, равномерно ограничены на $\mathbf{X}$.

A4. Матрица $\bar{\Gamma}_n(\delta^0)$ положительно определена: $\bar{\Gamma}_n(\delta^0) \succ 0$.

A5. Для матрицы $\bar{\Gamma}_n(\delta^0)$ существует предел $\varliminf\limits_{n\to\infty} \left|\bar{\Gamma}_n(\delta^0)\right| = b > 0$.

Тогда ОМП $\hat{\delta}^n$ асимптотически нормально распределена:

$$\mathcal{L}\left\{n^{\frac{1}{2}}\left(\bar{\Gamma}_n(\delta^0)\right)^{\frac{1}{2}}(\hat{\delta}^n - \delta^0)\right\} \xrightarrow[n\to\infty]{} \mathcal{N}_{m+1}(0_{m+1}, \mathbf{I}_{m+1}).$$

В **разделе 2.3** для модели (1)–(3) рассматривается задача прогнозирования будущего значения Y_{n+1} в точке X_{n+1} по выборке $\{X_t,\nu_t\}_{t=1}^n$, то есть задача построения точечного прогноза $\hat{Y}_{n+1}$ [7]. Для этого используется подстановочная прогнозирующая статистика:

$$\hat{Y}_{n+1} = F(X_{n+1};\hat{\theta}^n). \tag{5}$$

Согласно **следствию 2.1** прогноз $\hat{Y}_{n+1}$ является асимптотически несмещенным и для его среднеквадратического риска $R = \mathbf{E}_{X_{n+1},\delta^0}\left\{\left(\hat{Y}_{n+1} - Y_{n+1}\right)^2\right\}$ справедливо следующее асимптотическое разложение:

$$R = (\sigma^0)^2 + n^{-1}\lambda_n(\delta^0, X_{n+1}) + o\left(n^{-1}\right), \tag{6}$$

где $\lambda_n(\delta^0,X) = (\nabla_\theta F(X;\theta^0))'\left(\bar{\Gamma}_n(\delta^0)\right)^{-1}_{(1,1)}(\nabla_\theta F(X;\theta^0))$, $\left(\bar{\Gamma}_n(\delta^0)\right)^{-1}_{(1,1)}$ — верхний левый блок размера $m\times m$ обратной матрицы $\left(\bar{\Gamma}_n(\delta^0)\right)^{-1}$, являющейся ограниченной величиной.

В **разделе 2.4** приведены результаты компьютерных экспериментов, иллюстрирующие свойство сильной состоятельности ОМП $\hat{\delta}^n$. При этом с увеличением количества классов K ОМП по классифицированным наблюдениям приближается по точности к оценкам по методу наименьших квадратов по полным регрессионным наблюдениям. Численные результаты демонстрируют достаточную точность апроксимации среднеквадратического риска R с помощью формулы (6).

В **главе 3** решена задача **P4**: построены статистические критерии (решающие правила) проверки гипотез об истинном значении параметров регрессионных моделей при наличии классификации наблюдений.

Для этого в **разделе 3.1** формализуется задача **P4**. Множество всевозможных значений параметров Δ разбивается на два непересекающихся подмножества Δ_0 и Δ_1. Рассматриваются две гипотезы H_0, H_1:

$$H_0 = \{\delta^0 \in \Delta_0\},\ H_1 = \{\delta^0 \in \Delta_1\}.$$

Построение статистических критериев опирается на принцип оптимальности Неймана-Пирсона.

Раздел 3.2 посвящен проверке простых гипотез, т.е. случаю, когда множество возможных значений параметров Δ состоит из двух элементов δ_1 и δ_2, $\delta_1 \neq \delta_2$: $\Delta_0 = \{\delta_1\}$, $\Delta_1 = \{\delta_2\}$. В **подразделе 3.2.1** показано, что критерий Неймана-Пирсона для модели (1)–(3) имеет экспоненциальную по n вычислительную сложность $O(K^n)$. В **подразделе 3.2.2** построен статистический критерий, вычислительная сложность которого линейна по n: $O(nK)$.

Вводятся обозначения для статистики отношения правдоподобия:

$$L = L(\mathcal{H}, \mathcal{X}) = \prod_{t=1}^{n} \frac{P(\nu_t; \delta_2, X_t)}{P(\nu_t; \delta_1, X_t)},\ \zeta_t = \ln \frac{P(\nu_t; \delta_2, X_t)}{P(\nu_t; \delta_1, X_t)},$$

$$E_{t,1} = \mathbf{E}_{X_t,\delta_1}\{\zeta_t\},\ D_{t,1} = \mathbf{D}_{X_t,\delta_1}\{\zeta_t\}.$$

Теорема 3.2.[4] Пусть выполнены следующие условия.

SH1. Количество классов больше двух: $K > 2$.

SH2. Функции $F(X;\theta_1)$, $F(X;\theta_2)$ непрерывны на $\mathbf{X}$.

SH3. Пространство регрессоров $\mathbf{X} \subset \mathbb{R}^N$ — компакт.

SH4. Существует такое $\rho > 0$, что

$$\varliminf_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} \mathbb{I}\big\{|{\sigma_1}^2 - {\sigma_2}^2| + |F(X_t;\theta_1) - F(X_t;\theta_2)| \geq 2\rho\big\} = b,$$

где $b = b(\delta_1, \delta_2, \rho, F(\cdot)) > 0$.

Тогда при любом наперед заданном уровне значимости $\varepsilon \in (0,1)$ решающее правило

$$d_{ANP} = d_{ANP}(\mathcal{H}, \mathcal{X}) = \begin{cases} 0, & L(\mathcal{H}, \mathcal{X}) < c_{ANP}; \\ 1, & L(\mathcal{H}, \mathcal{X}) \geq c_{ANP}, \end{cases}$$

где $c_{ANP} = \exp\left(\Phi^{-1}(1-\varepsilon)\sqrt{\sum_{t=1}^{n} D_{t,1}} + \sum_{t=1}^{n} E_{t,1}\right)$, обладает следующими свойствами:

1) вероятность ошибки I-ого рода α_{ANP} стремится к ε: $\alpha_{ANP} \xrightarrow[n\to\infty]{} \varepsilon$;

2) мощность w_{ANP} стремится к 1: $w_{ANP} \xrightarrow[n\to\infty]{} 1$;

3) решающее правило d_{ANP} имеет наибольшую мощность w_{ANP} среди всех тестов, вероятность ошибки I-ого рода которых не превосходит вероятности ошибки I-ого рода теста d_{ANP}.

В **разделе 3.3** рассмотрен случай простой нулевой гипотезы H_0 и сложной альтернативы H_1. В этом случае $\Delta_0 = \{\bar{\delta}\}$, $\Delta_1 = \Delta \setminus \Delta_0$, где $\bar{\delta} \in \Delta$ — предполагаемое (гипотетическое) значение параметра δ^0. Вводятся обозначения: статистика отношения правдоподобия

$$\Lambda = \Lambda(\mathcal{H}, \mathcal{X}, \bar{\delta}) = \frac{P(\mathcal{H}; \bar{\delta}, \mathcal{X})}{\sup\limits_{\delta\in\Delta} P(\mathcal{H}; \delta, \mathcal{X})} \in [0, 1],$$

где $P(\mathcal{H}; \delta, \mathcal{X}) = \prod_{t=1}^{n} P(\nu_t; \delta, X_t)$; $F^{-1}_{\chi^2_m}(\cdot)$ — квантиль χ^2-распределения с m степенями свободы.

Теорема 3.3.[12] Пусть выполнены условия теоремы 2.4. Тогда для любого наперед заданного $\varepsilon \in (0, 1)$ и $c_{LR} = F^{-1}_{\chi^2_{m+1}}(1-\varepsilon)$ предел при $n \to \infty$ вероятности ошибки I-ого рода α_{LR} решающего правила

$$d_{LR} = d_{LR}(\mathcal{H}, \mathcal{X}) = \begin{cases} 0, & -2\ln\Lambda(\mathcal{H}, \mathcal{X}, \bar{\delta}) < c_{LR}; \\ 1, & -2\ln\Lambda(\mathcal{H}, \mathcal{X}, \bar{\delta}) \geq c_{LR} \end{cases} \tag{7}$$

равен ε: $\alpha_{LR} \xrightarrow[n\to\infty]{} \varepsilon$.

Теорема 3.4.[12] Пусть выполнены следующие условия.

CH1. Количество классов больше двух: $K > 2$.

CH2. Выполнены условия теоремы 2.4.

CH3. Пространство регрессоров $\mathbf{X} \subset \mathbb{R}^N$ — компакт.

CH4. Для любого $\rho > 0$ существует $\gamma = \gamma(\rho) > 0$ такое, что при любом $\theta \in \Theta$, $|\bar{\theta} - \theta| \geq \rho$, выполнено следующее предельное соотношение:

$$\varliminf_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} \mathbb{I}\left\{\left|F(X_t; \bar{\theta}) - F(X_t; \theta)\right| \geq \gamma\right\} = b,$$

где $b = b(\bar{\theta}, \theta, \gamma, F(\cdot)) > 0$.

Тогда для любого фиксированного $\delta^1 \neq \bar{\delta}$ мощность w_{LR} решающего правила (7) стремится к 1: $w_{LR}(\delta^1) \xrightarrow[n\to\infty]{} 1$.

В разделе 3.4 приведены результаты компьютерных экспериментов по проверке гипотез об истинном значении параметров, которые иллюстрируют теоретические свойства: вероятности ошибок I-ого рода колеблются около заданного уровня значимости, мощности стремятся к 1.

В **главе 4** решена задача **P5**: построены статистические критерии проверки гипотез согласия для регрессионных моделей при наличии классификации наблюдений на основе модифицированной χ^2-статистики.

В **разделе 4.1** формализуется задача **P5**. Вид регрессионной функции $F(\cdot,\cdot)$ полагается неизвестным, т.е. на вероятностном пространстве $(\Omega,\mathcal{F},\mathbf{P})$ рассматривается регрессионная модель

$$Y_t = f^0(X_t) + \xi_t,\ t = 1,\dots,n,$$

где n — объем выборки; $X_t \in \mathbf{X} \subseteq \mathbb{R}^N$ — наблюдаемый вектор-столбец регрессоров; $\xi_t \in \mathbb{R}$ — ненаблюдаемая случайная величина ошибок с нормальным распределением вероятностей $\mathcal{N}(0,(\sigma^0)^2)$. Случайные величины $\{\xi_t\}_{t=1}^n$ независимы в совокупности. Неизвестными характеристиками модели являются дисперсия ошибок $(\sigma^0)^2$ и функция регрессии $f^0(\cdot)$. Набор интервалов (2) определяет классификацию зависимой переменной Y_t в виде (3).

Рассматриваемая в данной главе задача заключается в том, чтобы построить статистический тест для проверки гипотезы согласия функции регрессии с параметрическим семейством $H_0 = \{f^0(\cdot) \in \mathcal{F}\}$ против альтернативы $H_1 = \{f^0(\cdot) \notin \mathcal{F}\}$, где $\mathcal{F} = \{F(X;\theta), \theta \in \Theta, X \in \mathbf{X}\}$ — предполагаемое параметрическое семейство функций регрессии.

Подраздел 4.2.1 посвящен статистической проверке гипотез согласия в случае, когда гипотеза H_0 — простая, т.е. параметрическое семейство функций регрессии $\mathcal{F}$ состоит только из одной функции: $\mathcal{F} = \{F(\cdot;\theta^0)\}$ и истинное значение дисперсии ошибки наблюдений $(\sigma^0)^2$ известно.

Вводятся следующие обозначения ($k,s \in \mathbf{K}$):

$$q(k) = \sum_{t=1}^{n} \mathbb{I}\{\nu_t = k\}, \quad p_k(\delta) = \frac{1}{n}\sum_{t=1}^{n} P(k;\delta,X_t),$$

$$d_{k,s}(\delta) = \mathbb{I}\{k=s\}p_k(\delta) - \frac{1}{n}\sum_{t=1}^{n}\left(P(k;\delta,X_t)\right)^2,$$

$$\phi(\delta) = \left(\frac{q(k) - np_k(\delta)}{\sqrt{n}}\right)_{k=1}^{K-1}, \quad D(\delta) = (d_{k,s})_{k,s=1}^{K-1}.$$

Матрица $D(\delta^0)$, $\delta^0 = ((\theta^0)', (\sigma^0)^2)'$, предполагается невырожденной. Предложена следующая модификация классической χ^2-статистики:

$$\tilde{\chi}^2(\delta^0) = (\phi(\delta^0))'(D(\delta^0))^{-1}\phi(\delta^0).$$

Теорема 4.1.[17] Пусть матрица $D(\delta^0)$ положительно определена и для нее существует предел $\varlimsup_{n\to\infty} \left|D(\delta^0)\right| = b > 0$. Тогда при верной гипотезе H_0 имеет место сходимость по распределению:

$$\mathcal{L}_{H_0}\left\{\tilde{\chi}^2(\delta^0)\right\} \xrightarrow[n\to\infty]{} \chi^2_{K-1}.$$

Теорема 4.2.[17] Пусть выполнены условия теоремы 4.1. Тогда для любого наперед заданного уровня значимости $\varepsilon \in (0,1)$ и $c^0_{\chi^2} = F^{-1}_{\chi^2_{K-1}}(1-\varepsilon)$

при $n \to \infty$ предел вероятности ошибки I-ого рода $\alpha^0_{\chi^2}$ для решающего правила

$$d^0_{\chi^2} = d^0_{\chi^2}(\mathcal{H}, \mathcal{X}) = \begin{cases} 0, & \tilde{\chi}^2(\delta^0) < c^0_{\chi^2}; \\ 1, & \tilde{\chi}^2(\delta^0) \geq c^0_{\chi^2}, \end{cases}$$

равен ε: $\alpha^0_{\chi^2} \xrightarrow[n\to\infty]{} \varepsilon$.

В **подразделах 4.2.2** и **4.2.3** рассматривается случай сложной гипотезы $H_0 = \{f^0(\cdot) \in \mathcal{F}\}$. Для оценивания неизвестного вектора параметров $\delta^0 = ((\theta^0)', (\sigma^0)^2)'$ при верной гипотезе H_0 используется метод максимального правдоподобия. Отметим, что в случае верной гипотезы H_0 мы находимся в рамках регрессионной модели при наличии классификации наблюдений, исследованной в главах 1–3, поэтому справедливы все результаты, полученные в главе 2 для ОМП (4).

Обобщенным χ^2-распределение с параметрами S и C называется распределение вероятностей квадратичной формы $\zeta = z'\,C\,z$, где C — $(L \times L)$-матрица, а случайный вектор $z \in \mathbb{R}^L$ имеет L-мерное нормальное распределение с нулевым математическим ожиданием и ковариационной матрицей S: $\mathcal{L}\{z\} = \mathcal{N}_L(0_L, S)$.

Однако это определение избыточно: в случае симметричной неотрицательно определенной матрицы C случайную величину ζ можно представить в виде $\sum\limits_{j=1}^{L} \lambda_j u_j^2$, где $\{u_j :\ j = 1, \ldots, L\}$ — независимые в совокупности стандартные нормальные случайные величины, а $\lambda_1, \ldots, \lambda_L$ — собственные значения матрицы $(S^{\frac{1}{2}})'CS^{\frac{1}{2}}$. Тогда параметрами *обобщенного χ^2-распределения* являются L *и* $\lambda = (\lambda_1, \ldots, \lambda_L)'$.

Вводятся следующие обозначения: $((K-1) \times (m+1))$-матрица

$$P^{(1)}(\delta) = \left(\frac{\partial p_i(\delta)}{\partial \delta_j} \right)_{\substack{i=1,\ldots,K-1, \\ j=1,\ldots,m+1}},$$

$F_{\chi^2}(\cdot; L, \lambda)$ — функция обобщенного χ^2-распределения с параметрами L и $\lambda \in \mathbb{R}^L$, вектор $\lambda^n = (\lambda^n_1, \ldots, \lambda^n_{K-1})'$, состоящий из собственных значений матрицы $\mathbf{I}_{K-1} - P^{(1)}(\hat{\delta}^n)(\bar{\Gamma}_n(\hat{\delta}^n))^{-1}(P^{(1)}(\hat{\delta}^n))'D(\hat{\delta}^n)^{-1}$.

Строится подстановочная модифицированная χ^2-статистика:

$$\tilde{\chi}^2(\hat{\delta}^n) = (\phi(\hat{\delta}^n))'(D(\hat{\delta}^n))^{-1}\phi(\hat{\delta}^n).$$

Теорема 4.3.[17] Пусть выполнены следующие условия.

G1. Функция $F(X;\theta)$, $X \in \mathbf{X}$, трижды непрерывно дифференцируема по θ.

G2. Для любого $\delta \in \Delta$ функции $F(X;\theta)$, $\dfrac{\partial F(X;\theta)}{\partial \theta_i}$, $\dfrac{\partial^2 F(X;\theta)}{\partial \theta_i \partial \theta_j}$, $\dfrac{\partial^3 F(X;\theta)}{\partial \theta_i \partial \theta_j \partial \theta_s}$, $i, j, s = 1, \ldots, m$, равномерно ограничены на $\mathbf{X}$.

G3. Матрица $\bar{\Gamma}_n(\delta^0)$ положительно определена и существует предел $\varliminf\limits_{n\to\infty} \left|\bar{\Gamma}_n(\delta^0)\right| = b > 0$.

G4. Блочная $((m+K)\times(m+K))$-матрица

$$\left(\begin{array}{c|c} D(\delta^0) & P^{(1)}(\delta^0) \\ \hline \left(P^{(1)}(\delta^0)\right)' & \bar{\Gamma}_n(\delta^0) \end{array}\right)$$

является положительно определенной, при этом матрицы $(D(\delta^0))^{-1}$ и $\left(\bar{\Gamma}_n(\delta^0) - (P^{(1)}(\delta^0))'(D(\delta^0))^{-1}P^{(1)}(\delta^0)\right)^{-1}$ являются равномерно ограниченными по n.

G5. Существуют $((K-1)\times(K-1))$-матрицы $H(\delta^0)$ и $J(\delta)$ такие, что

$$D(\delta^0) - P^{(1)}(\delta^0)(\bar{\Gamma}_n(\delta^0))^{-1}(P^{(1)}(\delta^0))' \xrightarrow[n\to\infty]{} H(\delta^0),$$

$$D(\delta) \underset{n\to\infty}{\rightrightarrows} J(\delta) \text{ для любого } \delta \in \Delta.$$

Тогда при верной гипотезе H_0 для состоятельной ОМП $\hat{\delta}^n$, определяемой (4), распределение вероятностей подстановочной модифицированной статистики $\tilde{\chi}^2(\hat{\delta}^n)$ сходится к обобщенному χ^2-распределению с матричными параметрами $H(\delta^0)$ и $(J(\delta^0))^{-1}$.

Теорема 4.4.[17] В условиях теоремы 4.3 статистический критерий

$$d_{\chi^2} = d_{\chi^2}(\mathcal{H}, \mathcal{X}) = \begin{cases} 0, & p_{\chi^2} \geq \varepsilon; \\ 1, & p_{\chi^2} < \varepsilon, \end{cases}$$

где $p_{\chi^2} = 1 - F_{\chi^2}(\tilde{\chi}^2(\hat{\delta}^n); K-1, \lambda^n)$, обладает вероятностью ошибки I-ого рода α_{χ^2}, сходящейся к заданному уровню значимости $\varepsilon \in (0,1)$: $\alpha_{\chi^2} \xrightarrow[n\to\infty]{} \varepsilon$.

В разделе 4.3 приведены результаты компьютерных экспериментов по проверке гипотез согласия, которые иллюстрируют теоретические свойства: вероятности ошибок I-ого рода колеблются около заданного уровня значимости. Проиллюстрирована также состоятельность построенных решающих правил.

ЗАКЛЮЧЕНИЕ

Основные научные результаты диссертации

1. Для регрессионной модели при наличии классификации наблюдений впервые установлены необходимые и достаточные условия идентифицируемости при различном числе интервалов классификации, что позволяет проверять условия разрешимости задачи статистического оценивания параметров такой модели [1, 3, 5, 8, 21, 22, 23].

2. Для регрессионной модели при наличии классификации наблюдений доказаны новые достаточные условия сильной состоятельности и асимптотической нормальности оценки максимального правдоподобия параметров модели, что дает возможность делать выводы о точности оценивания коэффициентов регрессии и дисперсии случайных ошибок наблюдений [2, 3, 9, 13, 15, 16, 18, 19, 21, 22, 23, 24].

3. Для подстановочной прогнозирующей статистики в рамках регрессионной модели при наличии классификации наблюдений построены новые оценки асимптотического смещения и среднеквадратического риска, что позволяет не только построить прогноз зависимой переменной, но и оценить точность этого прогноза [2, 3, 7, 13, 16, 23].

4. Впервые построены статистические критерии проверки простой нулевой гипотезы против простой и сложной альтернатив об истинном значении параметров регрессионной модели при наличии классификации наблюдений; доказано, что с увеличением объема выборки вероятности ошибок I-ого рода этих критериев стремятся к заданному уровню значимости, а мощности критериев стремятся к единице, что позволяет контролировать точность принимаемых решений о значениях параметров модели [4, 10, 11, 12, 20].

5. Впервые построены статистические критерии проверки гипотез согласия функции регрессии с некоторым параметрическим семейством для регрессионной модели при наличии классификации наблюдений, основанные на модификации χ^2-статистики, в случае простой и сложной нулевых гипотез; доказано, что с увеличением объема выборок вероятности ошибок I-ого рода этих критериев стремятся к заданному уровню значимости, что дает возможность контролировать точность принимаемых решений о принадлежности функции регрессии заданному параметрическому семейству функций [2, 3, 6, 14, 17, 21, 22, 23, 25].

Рекомендации по практическому использованию результатов

Практическая значимость диссертационной работы состоит в том, что теоретические результаты могут применяться при статистическом анализе и прогнозировании регрессионных данных при наличии классификации наблюдений в таких областях, как экономика, финансы, техника, медицина. Полученные результаты также целесообразно использовать в учебном процессе при чтении специальных курсов по математической и прикладной статистике на математических факультетах университетов.

СПИСОК ПУБЛИКАЦИЙ СОИСКАТЕЛЯ УЧЕНОЙ СТЕПЕНИ

Статьи в научных изданиях в соответствии с пунктом 18 Положения о присуждении ученых степеней и присвоении ученых званий в Республике Беларусь

1. *Агеева, Е. С.* Статистическое оценивание параметров множественной линейной регрессии при наличии случайного цензурирования / Е. С. Агеева, Ю. С. Харин // Вестн. Беларус. гос. ун-та. Сер. 1, Физ., матем., информат. — 2012. — № 1. — С. 72–78.
2. *Агеева, Е. С.* Состоятельность оценки максимального правдоподобия параметров множественной регрессии по классифицированным наблюдениям / Е. С. Агеева, Ю. С. Харин // Доклады Национальной академии наук Беларуси. — 2012. — Т. 56, № 5. — С. 11–19.
3. *Ageeva, H.* ML estimation of multiple regression parameters under classification of the dependent variable / H. Ageeva, Yu. Kharin // Lithuanian Mathematical Journal. — 2015. — Vol. 55, №. 1. — P. 48–60.
4. *Агеева, Е. С.* Проверка простых гипотез для регрессионной модели при наличии классификации зависимой переменной / Е. С. Агеева // Труды института математики. — 2015. — Т. 23, № 1. — С. 3–11.

Статьи в сборниках научных трудов

5. *Агеева, Е. С.* О статистическом оценивании параметров регрессии при наличии случайного цензурирования / Е. С. Агеева, Ю. С. Харин // Современные информационные компьютерные технологии: сборник научных статей / ГрГУ им. Я. Купалы. — Гродно, 2011. — С. 199–204.
6. *Агеева Е. С.* Критерий согласия для регрессионной модели при наличии классификации зависимых наблюдений / Е. С. Агеева // Теория вероятностей, случайные процессы, математическая статистика и их приложения: сборник научных статей / РИВШ; под редакцией Н.Н. Труша, Г.А. Медведева, Ю.С. Харина. — Минск, 2014. — С. 11–16.
7. *Агеева, Е. С.* Прогнозирование регрессионных временных рядов при наличии классификации наблюдений / Е. С. Агеева, Ю. С. Харин // Экономика. Моделирование. Прогнозирование: сб. науч. тр. / НИЭИ Мин-ва экономики Респ. Беларусь. — Минск, 2015. — Вып. 9. — С. 178–183.

Статьи в сборниках материалов научных конференций

8. *Агеева, Е. С.* Статистический анализ множественной линейной регрессии при наличии случайного цензурирования / Е. С. Агеева, Ю. С. Харин // Информационные системы и технологии: материалы VI Международной конференции, Минск, 24–25 ноября, 2010 г. / Белорус. гос. ун-т; редкол.: А. Н. Курбацкий [и др.]. — Минск, 2010. — С. 120–124.

9. *Агеева, Е. С.* Статистическое оценивание параметров множественной регрессии при наличии классификации наблюдений / Е.С. Агеева, Ю.С. Харин // Международный конгресс по информатике: информационные системы и технологии: материалы международного научного конгресса, Минск, 31 октября–3 ноября 2011 г.: в 2 ч. / БГУ, ОИПИ НАН Беларуси, НТА «Инфопарк»; редкол.: С. В. Абламейко (отв. ред.) [и др.]. — Минск, 2011. — Ч. 1. — С. 22–26.

10. *Агеева, Е. С.* Проверка сложных гипотез для множественной регрессии при наличии классификации наблюдений / Е. С. Агеева, Ю. С. Харин // 70-ая научная конференция студентов и аспирантов Белорус. гос. ун-та: сб. работ, Минск, 15–18 сентября, 2013 г.: в 3 ч. — Минск, 2013. — Ч. 1. — С. 160–164.

11. *Агеева, Е. С.* Проверка простой нулевой гипотезы и сложной альтернативной гипотезы для множественной регрессии при наличии классификации наблюдений / Е. С. Агеева // Международный конгресс по информатике: информационные системы и технологии: материалы международного научного конгресса, Минск, 4-7 ноября, 2013 г. — Минск: БГУ, 2013. — С. 12–16.

12. *Aheyeva, H.* On hypothesis testing for regression model under classification of dependent variable / H. Aheyeva // Computer data analysis and modeling: proceeding of the 10th International conference, Minsk, September 11–14, 2013: in 2 vol. / Ministry of Education of the Republic of Belarus; ed. board: Prof. Dr. S. Aivazian, Prof. Dr. P. Filzmoser, Prof. Dr. Yu. Kharin. — Minsk, 2013. — Vol. 1. — P. 52–55.

13. *Агеева, Е. С.* Прогнозирование линейных регрессионных временных рядов при классификации зависимой переменной / Е. С. Агеева, Ю. С. Харин // Проблемы прогнозирования и государственного регулирования социально-экономического развития: материалы XV Междунар. науч. конференции, Минск, 23–24 октября, 2014 г. — Минск, 2014. — С. 188–189.

14. *Агеева, Е. С.* Проверка сложных гипотез согласия для регрессионной модели при наличии классификации зависимой переменной / Е. С. Агеева // Статистические методы анализа экономики и общества: труды 6-ой Междунар. научно-практической конференции студентов и аспирантов, Москва, 12–15 мая, 2015 г. — Москва, Россия, 2015. — С. 37–38.

15. *Харин, Ю. С.* Об одном подходе к построению робастной оценки линейной регрессионной модели на основе классификации зависимой переменной / Ю. С. Харин, Е. С. Агеева // Проблемы прогнозирования и государственного регулирования социально-экономического развития: материалы XVI Междунар. науч. конф., Минск, 23 октября, 2015 г.: в 3 т. / НИЭИ М-ва экономики Респ. Беларусь; редкол.: А.В. Червяков [и др.]. — Минск, 2015. — Т. 1. — С. 150–155.

16. *Ageeva, H.* Forecasting of regression model under classification of the dependent variable / H. Ageeva // Computer data analysis and modeling: proceeding of the 11th International conference, Minsk, September 6–10, 2016. / Publishing center of BSU; ed. board: Prof. Dr. S. Aivazian, Prof. Dr. P. Filzmoser, Prof. Dr. Yu. Kharin. — Minsk, 2016. — P. 117–120.

17. *Агеева, Е. С.* Хи-квадрат критерий для регрессионной модели при наличии классификации наблюдений / Е. С. Агеева, Ю. С. Харин // Международный конгресс по информатике: информационные системы и технологии — International Congress on Computer Science: Information Systems and Technologies [Электронный ресурс] : материалы междунар. науч. конгресса, Республика Беларусь, Минск, 24–27 окт. 2016 г. / редкол.: С. В. Абламейко (гл. ред.), В. В. Казаченок (зам. гл. ред. [и др.]. — Минск : БГУ, 2016. — С. 416–420. — 1 электрон. опт. диск (DVD-RW).

Тезисы докладов

18. *Aheyeva, H. S.* On asymptotic properties of ML estimators for the regression parameters under classification of observations / H. S. Aheyeva, Yu. S. Kharin // Modern Stochastics: Theory and Applications III: conference materials of the Intern. conf., Kiev, Sept. 10–14, 2012. — Kiev, Ukraine, 2012. — P. 77.

19. *Агеева, Е. С.* Об асимптотических свойствах ОМП регрессионных коэффициентов при наличии классификации наблюдений / Е. С. Агеева, Ю. С. Харин // XI Белорусская математическая конференция: сб. тезисов Междун. науч. конф., Минск, 4–9 ноября, 2012 г. — Минск, 2012. — С. 43.

20. *Aheyeva, H. S.* Hypothesis testing in the multiple regression model under classification of the dependent variable / H. S. Aheyeva, Yu. S. Kharin // International Conference on Robust Statistics: book of abstr. of the Intern. conf., St. Petersburg, July 8–12, 2013. — St. Petersburg, Russia, 2013. — P. 11.

21. *Ageeva, H.* Statistical analysis for regression model under grouping distortion / H. Ageeva, Yu. Kharin // 11th International Vilnius Conference on Probability Theory and Mathematical Statistics: book of abstr. of the Intern. conf., Vilnus, 30 June – 3 July, 2014. — Vilnus, Lithuania, 2014. — P. 25.

22. *Ageeva, H.* Statistical inferences for multiple regression under grouping / H. Ageeva, Yu. Kharin // International Conference on Robust Statistics: book of abstr. of the Intern. conf., Martin-Luther-University Halle-Wittenberg, August 10–15, 2014. — Halle (Saale), Germany, 2014. — P. 18.

23. *Ageeva, H.* Forecasting of regression time series under classification of the dependent variable / H. Ageeva, Yu. Kharin // Probability, Reliability and Stochastic Optimization: book of abstr. of the Intern. conf., Kyiv, April 7–10, 2015. — Kiev, Ukraine, 2015. — P. 37.

24. *Kharin, Yu.* Statistical Analysis of Time Series Based on Incomplete Discrete Data / Yu. Kharin, H. Ageeva, H. Rudakouskaya // Data Analysis Methods for Software Systems: book of abstr. of 7th Intern. Workshop, Druskininkai, Lithuania, December 3–5, 2015. — Druskininkai, Lithuania, 2015. — P. 30–31.

25. *Агеева, Е. С.* Проверка гипотез согласия для регрессионной модели при наличии классификации зависимой переменной / Е. С. Агеева // XII Белорусская математическая конференция: материалы Международной научной конференции. Минск, 5–10 сентября, 2016 г.: в 5 ч. — Минск, 2016. — Ч. 4. — С. 3.

РЭЗЮМЭ

Агеева Алена Сяргееўна

Статыстычныя вывады аб рэгрэсійных мадэлях пры наяўнасці класіфікацыі назіранняў

Ключавыя словы: рэгрэсія, групаванне, класіфікацыя, ацэнка максімальнай праўдападобнасці, параметрычныя гіпотэзы, гіпотэзы адпаведнасці, χ^2-статыстыка.

Мэтай дысертацыйнай работы з'яўляецца пабудова статыстычных вываданў (статыстычных ацэнак, рашаючых правілаў і прагназуючых статыстык) аб рэгрэсійных мадэлях пры наяўнасці класіфікацыі назіранняў. Пры даследаванні выкарыстоўваліся метады тэорыі імавернасцей, матэматычнай статыстыкі, рэгрэсійнага аналізу, камп'ютэрнага мадэлявання.

У дысертацыйнай рабоце атрыманы наступныя новыя навуковыя вынікі. Для рэгрэсійнай мадэлі пры наяўнасці класіфікацыі назіранняў устаноўлены неабходныя і дастатковыя ўмовы ідэнтыфікацыі мадэлі, знойдзены дастатковыя ўмовы моцнай кансістэнтнасці і асімптатычнай нармальнасці ацэнкі максімальнай праўдападобнасці. Для падстаноўчай прагназуючай статыстыкі прапанаваны ацэнкі яе асімптатычнага зрушэння і сярэднеквадратычнага рыску. Пабудаваны статыстычныя крытэрыі для праверкі простай нулявой гіпотэзы супраць простай і складанай альтэрнатыў аб сапраўдных значэннях параметраў рэгрэсійнай мадэлі пры наяўнасці класіфікацыі назіранняў; даказана, што іх імавернасці памылкі I-ага роду імкнуцца да зададзенага ўзроўню значнасці, а магутнасці імкнуцца да 1. Пабудаваны статыстычныя крытэрыі для праверкі гіпотэз адпаведнасці для рэгрэсійнай мадэлі пры наяўнасці класіфікацыі назіранняў, заснаваныя на мадыфікацыі χ^2-статыстыкі; даказана, што іх імавернасці памылкі I-ага роду імкнуцца да зададзенага ўзроўню значнасці.

Атрыманыя вынікі могуць быць выкарыстаны пры статыстычным аналізе і прагназаванні рэгрэсійных дадзеных пры наяўнасці класіфікацыі назіранняў у такіх галінах, як эканоміка, фінансы, тэхніка, медыцына, а таксама ў навучальным працэсе.

РЕЗЮМЕ

Агеева Елена Сергеевна

Статистические выводы о регрессионных моделях при наличии классификации наблюдений

Ключевые слова: регрессия, группирование, классификация, оценка максимального правдоподобия, параметрические гипотезы, гипотезы согласия, χ^2-статистика.

Цель диссертационной работы заключается в построении статистических выводов (статистических оценок, решающих правил и прогнозирующих статистик) о регрессионных моделях при наличии классификации наблюдений. При исследовании использовались методы теории вероятностей, математической статистики, регрессионного анализа, компьютерного моделирования.

В диссертационной работе получены следующие новые научные результаты. Для регрессионной модели при наличии классификации наблюдений установлены необходимые и достаточные условия идентифицируемости, найдены достаточные условия сильной состоятельности и асимптотической нормальности оценки максимального правдоподобия. Для подстановочной прогнозирующей статистики построены оценки ее асимптотического смещения и среднеквадратического риска. Построены статистические критерии для проверки простой нулевой гипотезы против простой и сложной альтернатив об истинных значениях параметров регрессионной модели при наличии классификации наблюдений; доказано, что их вероятности ошибок I-ого рода стремятся к заданному уровню значимости, а мощности стремятся к 1. Построены статистические критерии проверки гипотез согласия для регрессионной модели при наличии классификации наблюдений, основанные на модификации χ^2-статистики; доказано, что их вероятности ошибок I-ого рода стремятся к заданному уровню значимости.

Полученные результаты могут быть использованы при статистическом анализе и прогнозировании регрессионных данных при наличии классификации наблюдений в таких областях, как экономика, финансы, техника, медицина, а также в учебном процессе.

SUMMARY

Ageeva Helena

Statistical inferences on regression models under classification of observations

Key words: regression, grouping, classification, maximum likelihood estimator, parametric hypotheses, goodness-of-fit testing, χ^2 statistic

The goal of this dissertation is to construct statistical inferences (statistical estimators, statistical tests and forecasting statistics) on regression models under classification of observations. The techniques used include methods of probability theory, mathematical statistics, regression analysis, computer simulation.

In this dissertation the following new scientific results are obtained. Necessary and sufficient conditions for identifiability of the regression model under the classification of observations are established. Sufficient conditions for consistency and asymptotic normality of the MLE are found. Asymptotic bias and squared error risk are obtained for the plug-in forecasting statistic for the considered model. Statistical tests for simple null hypothesis versus simple and composite alternative hypotheses for true value of the regression model parameter under the classification of observations are constructed; their probabilities of type I error are proven to converge to the given significance level and their powers are proven to converge to 1. Statistical tests for goodness-of-fit hypotheses for the regression model under the classification of observations are constructed based on the modification of χ^2 statistic; their probabilities of type I error are proven to converge to the given significance level.

The obtained results can be used in statistical analysis and forecasting of real data under classification of observations in areas such as economics, finance, technology, medicine, as well as in the educational process.