САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

На правах рукописи

МАХМУД Саиф Абдель-Рахман Эйд

ЧИСЛЕННЫЕ МЕТОДЫ ИЗУЧЕНИЯ ОШИБОК В НЕКОТОРЫХ СТАТИСТИЧЕСКИХ ЗАДАЧАХ

Специальность

05.13.18 - Математическое моделирование, численные методы и комплексы программ.

А В Т О Р Е Ф Е Р А Т

диссертации на соискание ученой степени кандидата физико-математических наук

Санкт-Петербург
2002

Работа выполнена на кафедре
статистического моделирования
С.-Петербургского гос. Университета

Научный руководитель: доктор физ.-мат. наук,
профессор
ЕРМАКОВ Сергей Михайлович

Официальные оппоненты: доктор физ.-мат. наук,
профессор
НЕВЗОРОВ Валерий Борисович

доктор физ.-мат. наук,
профессор
СЕДУНОВ Валерий Витальевич

Ведущая организация: Институт проблем машиноведения
Российской академии наук

Защита состоится “ ” 2002 г. в час.
на заседании диссертационного Совета Д.212.232.51 по защите диссертаций на соискание ученой степени доктора наук в Санкт-Петербургском государственном университете по адресу: 198504, Санкт-Петербург, Старый Петергоф, Библиотечная площадь, дом 2, математико-механический факультет.

С диссертацией можно ознакомиться в Научной библиотеке им. М. Горького Санкт-Петербургского государственного университета по адресу: 199034, Санкт-Петербург, Университетская набережная, дом 7/9.

Автореферат разослан “ ” 2002 г.

Ученый секретарь
диссертационного Совета
д.ф.-м.н., профессор Б.К. Мартыненко

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Проблема обработки данных занимает важное место в современных научных исследованиях. В частности, важное значение имеет регрессионный анализ, используемый для создания удобных для дальнейшего исследования моделей, сглаживания данных и предсказания поведения сложных систем. Во всех случаях первостепенный интерес имеет гарантированная (с заданным уровнем надежности) оценка погрешности модели и прогноза. Проблема эта достаточно хорошо исследована в так называемом "классическом" случае - в предположении приближенной нормальности и независимости ошибок наблюдений и отсутствии (малости) систематической погрешности. В других случаях имеется ряд нерешённых проблем, которые представляют значительный теоретический и прикладной интерес. Отметим также, что для важного класса задач планирования эксперимента в регрессионном анализе в "неклассическом" случае имеется сравнительно немного результатов.

Актуальность темы. Широкое использование компьютеров при обработке данных делает весьма актуальной задачей построение численных процедур оценки погрешности регрессионного анализа в "неклассическом" случае, построение доверительных и толерантных интервалов для случая зависимых наблюдений в присутствии систематической погрешности и для распределений, отличных от нормального. Разработка численных процедур и программного обеспечения для решения упомянутых задач весьма актуальна при построении и обосновании вероятностных и имитационных моделей в медицине, биологии, экономике, инженерии и в других областях.

Цель исследований. Целью исследований являлось:

а) получение новых теоретических результатов в области регрессионного анализа при наличии систематической погрешности и зависимых ошибках наблюдений. Разработка эффективных вычислительных процедур для оценки погрешности и построения рандомизованных планов эксперимента;

б) создание системы программ для численной реализации полученных теоретических результатов;

в) разработка эффективных численных методов построения толерантных интервалов в случае γ-распределения;

г) создание программного обеспечения для вычисления толерантных интервалов.

Научная новизна. Впервые создано программное обеспечение, реализующее новые методы разделения систематической и случайной компонент погрешности в линейном регрессионном анализе при независимых ошибках наблюдения.

Предложены и изучены рандомизованные оценки метода наименьших квадратов в случае коррелированных наблюдений.

Впервые разработаны численные процедуры построения толерантных интервалов для γ-распределения.

Система программ, реализующая разработанные диссертантом и ранее разработанные методы являются новой.

Научная и практическая ценность.

Построенное в диссертации программное обеспечение и полученные теоретические обобщения результатов Ермакова С.М. и Ермакова С.М., Р. Швабе дают возможность строить новые бутстрап - процедуры для разделения случайной и детерминированной составляющей погрешности функции регрессии. Эти обобщения являются новым направлением исследований в некоторых вопросах регрессионного анализа. Построенные на их основе вычислительные процедуры и система программ может использоваться при изучении широкого класса вероятностных и имитационных моделей.

Численные методы и программное обеспечение для вычисления толерантных интервалов в случае γ-распределения могут быть применены при обработке данных в медицине, биологии, экономике и других областях.

На защиту выносятся.

1. Новые теоретические результаты относительно рандомизации оценок наименьших квадратов в случае зависимых наблюдений и связанных с этой рандомизацией бутстрап процедуры.

2. Методы вычисления толерантных интервалов в случае γ-распределения.

3. Алгоритмы и программы оценки погрешности оценок метода наименьших квадратов и вычисления толерантных интервалов.

Апробация результатов работы.

Результаты работы докладывались на семинаре кафедры статистического моделирования СПбГУ, на семинаре кафедры информационного менеджмента СПбГУ, институте машиноведения РАН.

Публикации.

Содержание работы отражено в одной опубликованной ранее работе.

Структура и объем диссертации.

Диссертация написана на английском языке, содержит 150 стр. текста и приложение на 50 страницах.

СОДЕРЖАНИЕ РАБОТЫ

Первые две главы диссертации посвящены исследованию и обобщению специального распределения узлов интерполирования так называемого Δ^2-распределения. Это распределение было введено в рассмотрение как аппарат уменьшения дисперсии при вычислении интегралов методом Монте - Карло.

Пусть μ есть σ-конечная мера, определенная на подмножестве некоторого множества $\mathfrak{X}$, $x \in \mathfrak{X}$, $f(x)$ интегрируемая с квадратом относительно меры μ функция, а $\varphi_1(x), ..., \varphi_m(x)$ множество ортонормированных на $\mathfrak{X}$ по отношению к μ функций.

Обозначим $\Delta(Q) = det\|\varphi_i(x_j)\|_{i,j=1}^m, \quad Q = (x_1, ..., x_m)$. Будем говорить, что случайная точка $P = (y_1, ..., y_m), \quad y_i \in \mathfrak{X}$ имеет Δ^2 -распределение (относительно меры μ^n) если плотность её распределения относительно $\mu^n(dQ)$ выражается формулой

$$\Delta^2(Q) = \frac{1}{m!}\Big(det\Big\|\varphi_i(x_j)\Big\|_{i,j=1}^m\Big)^2. \tag{1}$$

Плотность (1) обладает в частности следующими свойствами.

Если $C_1, ..., C_m$ случайные величины, определяемые из системы уравнений

$$\sum_{j=1}^m C_j\varphi_j(y_i) = f(y_i), \quad i = 1, ..., m,$$

где y_i случайные точки, распределенные с плотностью (1), то

1. Математическое ожидание C_j выражается формулой

$$EC_j = \int f(x)\varphi_j(x)\mu(dx).$$

(в левой части равенства j-тый коэффициент Фурье функции f по системе $\{\varphi_i\}$.)

2. Для дисперсии C_j справедливо неравенство

$$DC_j \leqslant \int \Big[f(x) - \sum_{i=1}^m \varphi_j(x)EC_j\Big]^2\mu(x).$$

При этом знак равенства достигается в том случае, когда система $\{\varphi_i\}$ удовлетворяет условию (регулярности):

$$\mu^m\{Q : \Delta(Q) = 0\} = 0.$$

Δ^2-распределение имеет также приложения в регрессионном анализе, если предположить, что функция f имеет случайную ошибку - наблюдаются значения

$$\varsigma(x_i) = f(x_i) + \varepsilon_i,$$

и считать меру μ сосредоточенной в конечном множестве N точек в области $\mathfrak{D} \in R^s$.

Именно свойство Δ^2-распределения в связи с задачами регрессионного анализа изучаются в первых двух главах диссертации.

Глава 1 посвящена случаю некоррелированных ошибок наблюдений. После изложения известных фактов, необходимых для дальнейшего (пп 1.1-1.4), описываются методы и алгоритмы моделирования Δ^2-распределения.

Обсуждается два способа проведения эксперимента. Первый, это т.н. активный, когда получаемые в результате моделирования случайные векторы $P = (y_1, ..., y_1)$ рассматриваются как рандомизованные планы регрессионного эксперимента (случайные реплики). (Заметим, что модой Δ^2-распределения является точный D-оптимизированный план.) Если возможно повторение эксперимента в одних и тех же точках, то можно использовать хорошо известные процедуры разделения систематической и случайной составляющей дисперсии.

Второй подход относится к случаю, когда данные эксперимента получены заранее и повторение эксперимента в использованных ранее точках невозможно или нецелесообразно. В этом случае также указывается процедура случайного отбора данных, позволяющая разделить компоненты погрешности. Процедура основана на результатах, полученных ранее (Ермаков, Швабе). В диссертации дана её программная реализация. Для модельного эксперимента приведены примеры использования этой процедуры. В заключительном параграфе дается прямое (не использующее предположений об ортогональности φ_i) доказательство несмещённости оценок метода наименьших квадратов. Имеет место следующее утверждение:

Лемма. 1.1 *Пусть $\varsigma(x)$ наблюдаемые в точке x значения случайной функции ς,*

$$\varsigma(x) = f(x) + \varepsilon(x), \quad E\varepsilon(x) = 0, \quad E\Big(\varepsilon(x_i)\varepsilon(x_j)\Big) = \begin{cases} \sigma^2 & \text{при } x_i = x_j \\ 0 & \text{при } x_i \neq x_j \end{cases},$$

и $\varphi_1, ..., \varphi_m$ заданные линейно независимые на носителе меры μ функции.

Если случайные величины $C_1, ..., C_m$ находятся как решения системы уравнений:

$$\sum_{j=1}^{m} C_j \varphi_j(x_i) = \varsigma(x_i), \qquad i = 1, ..., m,$$

где $(x_1, ..., x_m)$ *выбраны случайно в соответствии с* Δ^2*-распределением, то справедливы равенства*

$$E(C_j|\varepsilon_1, ..., \varepsilon_m) = \widehat{C}_j, \qquad EC_j = \widetilde{C}_j,$$

Здесь

$$(\widehat{C}_1, ..., \widehat{C}_m) = \underset{d_1,...,d_m}{argmin} \int \Big[\varsigma(x) - \sum_{l=1}^{m} d_l\varphi_l(x)\Big]^2 \mu(dx).$$

$$(\widetilde{C}_1, ..., \widetilde{C}_m) = \underset{d_1,...,d_m}{argmin} E \int \Big[\varsigma(x) - \sum_{l=1}^{m} d_l\varphi_l(x)\Big]^2 \mu(dx) =$$

$$= \underset{d_1,...,d_m}{argmin} \int \Big[f(x) - \sum_{l=1}^{m} d_l\varphi_l(x)\Big]^2 \mu(dx).$$

Вторая глава посвящена рандомизации оценок наименьших квадратов в случае зависимых наблюдений, когда имеют место равенства

$$E\varepsilon(x_i)\varepsilon(x_j) = b_{ij}. \tag{2}$$

При этом матрица $B = \|b_{ij}\|$ не является, вообще говоря, диагональной и положительно определена.

Разделы 2.1-2.5 содержат вспомогательные сведения из линейной алгебры и теории метода наименьших квадратов. В разделе 2.5 строится аналог Δ^2-распределения в этом случае. При этом используется известное разложение $B = \Gamma\Gamma^T$, где Г-треугольная матрица. В разделах 2.6-2.7 описываются алгоритмы моделирования обобщенного Δ^2-распределения и приводятся примеры приложения разработанных процедур к задаче обработки экспериментальных данных. Как и в случае некоррелированных наблюдений полученные результаты и алгоритмы оказываются полезными при построении оптимальных планов эксперимента. Заключительный раздел 2.8 посвящен общей задаче построения аналога Δ^2-распределения для случая, когда ошибка является стационарным гауссовским случайным процессом (полем).

Корреляционная функция $B(x, y) = E(\varepsilon(x), \varepsilon(y))$ гауссовского случайного процесса есть билинейный положительный функционал в гильбертовом пространстве и, следовательно, имеет представление

$$\mathfrak{B}(\xi, \eta) = (\widetilde{\mathfrak{B}}\xi, \eta),$$

где $\widetilde{\mathfrak{B}}$ есть линейный положительный оператор. В предположении строгой положительности также существует обратный интегральный оператор Г, ядро которого мы обозначим $\gamma(x, y)$.

Пусть теперь константы $\widehat{C}_1, ..., \widehat{C}_m$ и $\widetilde{C}_1, ..., \widetilde{C}_m$ найдены из условий $(\widehat{C}_1, ..., \widehat{C}_m) = \underset{d_1,...,d_m}{argmin}\ \mathsf{F}(\varsigma, d_1, ..., d_m)$ и $(\widetilde{C}_1, ..., \widetilde{C}_m) = \underset{d_1,...,d_m}{argmin}\ \mathsf{F}(f, d_1, ..., d_m)$ соответственно где $\mathsf{F}(f, d_1, ..., d_m) = \int\int [g(x) - \sum d_l\varphi_l(x)][g(y) - \sum d_l\varphi_l(y)]\mu(dx)\mu(dy)$, а $\varphi_1, ..., \varphi_m$ заданные линейно независимые функции на носителе меры μ.

Доказана следующая теорема.

Теорема. 2.5. *Пусть точка $P = (y_1, ..., y_m)$ выбирается случайно в соответствии с плотностью*

$$\Delta_d^2(Q) = \frac{det\|\varphi_l(x_i)\|_{1,l=1}^m \cdot det\|\psi_s(x_j)\|_{1,j=1}^m}{\int \mu^m(dQ) \cdot det\|\varphi_l(x_i)\|_{1,l=1}^m \cdot det\|\psi_s(x_j)\|_{1,j=1}^m},$$

где $\psi_k(x) = \int \gamma(x, y)\varphi_k(y)\mu(dy)$.

Тогда, если C_l $(l = 1, ..., m)$ определяется из системы уравнений

$$\sum_{l=1}^{m} C_l\varphi_l(y_i) = \varsigma(y_i),$$

то $E(C_l|\varepsilon(y_i), ..., \varepsilon(y_m)) = \widehat{C}_l$ и $EC_l = \widetilde{C}_l$.

Вторая часть диссертации посвящена построению и иследованию толерантных интервалов для γ-распределения.

Толерантный интервал - это интервал со случайными границами, который с заданной заранее вероятностью P содержит не менее заданной доли вероятностной массы непрерывного распределения, которому подчиняются элементы выборки. Алгоритмы численного построения толерантных интервалов для конкретных распределений оказываются достаточно сложными. Наиболее изученным является случай нормального распределения. Случай γ-распределения несмотря на его практическую важность (приложения в биологии, медицине, страховании) изучен мало.

Глава 3 диссертации содержит общие концепции интервалов прогноза и толерантных интервалов, их классификацию. Понятия β-средних и β-содержащих толерантных интервалов иллюстрируется на примере нормального распределения. В случае β-средних интервалов мы ожидаем, что 100% β популяции будет находиться внутри этого интервала $(0 < \beta < 1)$.

В случае β-содержащих толерантных интервалов мы ожидаем что по крайней мере 100% β популяции будет лежать в указанном интервале при доверительном уровне γ.

Раздел 3.7 посвящен краткому обзору литературы.

Глава 4 содержит введение, сравнительно простой случай экспоненциального распределения (Раздел 4.2) и затем весьма подробное рассмотрение толерантных интервалов для γ-распределения.

Пусть $(Q_1 v, Q_2 v)$, (Qv, ∞) и $(0, Qv)$ являются двусторонним и односторонним (нижним и верхним) соответственно толерантными интервалами. В разделе 4.4 их свойства изучаются и алгоритмы для их вычисления строятся.

Теоремы 4.1-4.5 сформулированные и доказанные в этих разделах посвящены вопросам единственности толерантных интервалов. Даются явные выражения констант Q, Q_1, Q_2.

Приведем в качестве примера следующую теорему.

Теорема. 4.1. *Константа Q для одностороннего верхнего β-среднего толерантного интервала для γ-распределения единственна и равна*

$$Q = \frac{1 - Be(n\alpha_1, \alpha_2, 1 - \beta)}{Be(n\alpha_1, \alpha_2, 1 - \beta)},$$

где $Be(n\alpha_1, \alpha_2, 1 - \beta)$ означает β квантиль для $Be(n\alpha_1, \alpha_2)$ (Be есть β-распределение).

Раздел 4.5 посвящен вопросам статистического моделирования для построения толерантных интервалов в случае γ-распределения. Заключительный раздел 4.6 содержит результаты сравнения толерантных интервалов в случае γ-распределений. Приведенный подробный анализ обширных вычислительных экспериментов показывает, что за редкими исключениями (равенство средних при большом объеме выборки) между толерантными интервалами для нормального и γ-распределений имеется весьма существенное различие.

Приложение к диссертационной работе содержит тексты программ на языке ПАСКАЛЬ с комментариями. Программы могут быть использованы как при обработке реальных экспериментальных данных, так и в исследовательской работе при решении модельных задач. Они могут быть использованы в учебном процессе в курсах прикладной статистики и обработки данных.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ И ВЫВОДЫ

1. Создана система программ моделирования Δ^2-распределения для случая независимых ошибок наблюдений. На её базе созданы программы осуществляющие бустрап процедуру разделения систематической и случайной компонент погрешности для оценок наименьших квадратов.

2. Получено обобщение результатов С.М. Ермакова и В.Г. Золотухина на случай линейно-независимых (не являющихся ортонормированными) базисных функций линейной регрессии.

3. Указано обобщение Δ^2-распределения на случай зависимых наблюдений. Создана система программ моделирующих указанное распределение (Δ_d^2-распределение)

4. Разработаны алгоритмы и система программ для построения различных типов толерантных интервалов в случае γ-распределения.

5. Проведено детальное сравнение толерантных интервалов для случаев нормального и γ-распределений. Показано, что за редкими исключениями между толерантными интервалами в этих двух случаях имеется существенное различие.

   Таким образом в диссертации решен ряд теоретических и вычислительных задач статистического моделирования и представлена система программ, которая может служить полезным дополнением к существующим пакетам программ обработки данных в особенности в той части, которая относится к анализу погрешности прогноза с помощью оценки параметров регрессиии построения доверительных интервалов.

ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

1. S.M. Ermakov, Mahmoud S. Eid., β-expectation tolerance intervals for Gamma distribution, Проблемы оптимизации дискретных систем. Сб. Под редакцией док. физ. мат. наук, проф, М.К. Чиркова. СПб. Изд-во НИИХ СПбГУ 2001. стр. 45-59.