На правах рукописи

ХАЙЛЕНКО ЕКАТЕРИНА АЛЕКСЕЕВНА

Алгоритмы оценивания параметров регрессионных моделей и планирования эксперимента при наличии выбросов и неоднородности распределения ошибок

Специальность 05.13.17 — Теоретические основы информатики

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата технических наук

НОВОСИБИРСК – 2013

Работа выполнена в Федеральном государственном бюджетном образовательном учреждении высшего профессионального образования «Новосибирский государственный технический университет»

Научный руководитель: доктор технических наук, доцент
Тимофеев Владимир Семенович

Официальные оппоненты: Хабаров Валерий Иванович
доктор технических наук, профессор
Федеральное государственное бюджетное образовательное учреждение высшего профессионального образования «Сибирский государственный университет путей и сообщений», заведующий кафедрой «Информационные технологии на транспорте»;

Осипов Александр Леонидович
кандидат технических наук, доцент
Федеральное государственное бюджетное образовательное учреждение высшего профессионального образования «Новосибирский государственный университет экономики и управления», заведующий кафедрой «Прикладные информационные технологии»

Ведущая организация: Федеральное государственное образовательное бюджетное учреждение высшего профессионального образования «Сибирский государственный университет телекоммуникаций и информатики»

Защита состоится «23» мая 2013 г. в 14-00 часов на заседании диссертационного совета Д 212.173.06 при Федеральном государственном бюджетном образовательном учреждении высшего профессионального образования «Новосибирский государственный технический университет» по адресу: 630073, Новосибирск, пр-т К. Маркса, 20.

С диссертацией можно ознакомиться в библиотеке Новосибирского государственного технического университета.

Автореферат разослан « 19 » апреля 2013 г.

Ученый секретарь
диссертационного совета Чубич Владимир Михайлович

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность работы. В различных отраслях науки и техники исследователям часто приходится сталкиваться с необходимостью анализа данных и получения достоверной, максимально согласуемой с его природой, информации об исследуемом процессе (явлении). При современном уровне развития науки и техники это приводит к постановке сложных и дорогостоящих экспериментов. При их проведении исследователь пытается извлечь наибольшее количество информации об изучаемых процессах при наименьших затратах. Одним из способов получения такой информации является решение задачи оценивания параметров регрессионных моделей, которое позволяет спрогнозировать поведение наблюдаемого объекта в дальнейшем.

Классическим методом оценивания параметров регрессионных зависимостей является метод максимального правдоподобия (ММП), однако его применение требует наличие априорной информации о виде распределения ошибок наблюдений. Другим популярным методом оценивания параметров является метод наименьших квадратов (МНК), преимущество которого состоит в простоте вычислительной процедуры получения оценок. Однако при появлении в выборке грубых ошибок наблюдений (выбросов) либо при отклонении распределения ошибок от нормального закона оценки, полученные классическими методами, перестают обладать оптимальными свойствами. Для решения проблемы оценивания параметров регрессионного уравнения при появлении выбросов был разработан ряд устойчивых методов оценивания. Исследованиями в данной области занимались Хьюбер П., Хампель Ф., Rousseeuw P.J, K. van Driessen, Болдин М.В., Тюрин Ю.Н и др. При негауссовском распределении ошибок наблюдений возможно применение адаптивных методов оценивания параметров регрессионных зависимостей. В данной области можно отметить работы Hogg R.V., Lenth R.V., Денисова В.И., Лисицина Д.В. Многообразие возможных распределений случайной ошибки привело к идее применения ММП на основе универсальных распределений, одним из которых является обобщенное лямбда–распределение (GL–распределение), описывающее целый класс рас-

пределений, таких как нормальное, экспоненциальное, Вейбулла, Гамма–, Бета– и др. В результате появляется возможность оценивания параметров регрессионных моделей для любых распределений случайных ошибок, представимых в рамках GL-распределения.

Хорошо известно, что качество оценок параметров также зависит от информативности точек, в которых проводились измерения, т.е. можно получить большее количество информации об исследуемом процессе путем использования планов эксперимента. Наиболее известными исследователями в данной области являются Федоров В.В., Адлер Ю.П., Фишер Р., Налимов В.В., Денисов В.И., Попов А.А., Хабаров В.И. и др. Однако классические алгоритмы построения оптимальных планов эксперимента позволяют учитывать лишь неоднородность дисперсий на области планирования, но в ряде случаев на различных ее участках могут быть разные распределения. Поэтому необходимы алгоритмы синтеза оптимальных планов в условиях неоднородности формы распределения ошибок наблюдений на всей области планирования, построение которых также предлагается провести на основе GL-распределения.

Цель работы состоит в обеспечении возможности устойчивого и адаптивного оценивания параметров регрессионных моделей и синтеза оптимальных планов эксперимента при различных распределениях ошибок наблюдений.

Для достижения данной цели поставлены и решены следующие задачи:

- разработка, реализация и исследование модификаций метода наименьших уравновешенных квадратов (LTS), рангового метода и алгоритмов построения оценочных подмножеств, близких к A- и D-оптимальному плану для схемы LTS-оценивания;
- разработка, реализация и исследование адаптивного метода оценивания параметров регрессионного уравнения на основе GL-распределения;
- вывод соотношений для вычисления элементов информационной матрицы Фишера на основе GL-распределения и реализация на их основе нового алгоритма построения оптимального плана эксперимента;

- разработка программной системы устойчивого и адаптивного оценивания параметров регрессионных моделей и планирования эксперимента;
- применение разработанных алгоритмов устойчивого, адаптивного оценивания параметров и планирования эксперимента для задачи оценивания кривой провисания троса и прогнозирования покупательского спроса.

Методы исследования. Исследование проводилось с использованием методов регрессионного анализа, теории планирования эксперимента, математического анализа и линейной алгебры, численных методов, методов оптимизации и методов статистического моделирования.

Достоверность и обоснованность научных выводов и рекомендаций подтверждается корректными применением аналитических методов, соответствием выводов хорошо известным теоретическим законам, а также путем подтверждения полученных выводов и работоспособности алгоритмов результатами вычислительных экспериментов.

Научная новизна состоит в следующем:

- предложены модификации рангового метода на основе расстояния Махаланобиса и метода LTS на основе расстояний Махаланобиса, Кука, Велша-Куха и робастного расстояния, способ формирования оценочного подмножества исходя из критериев A- и D-оптимальности, применение предложенных алгоритмов позволяет проводить устойчивое оценивание параметров уравнения регрессионной зависимости по наиболее информативным наблюдениям;
- разработан алгоритм адаптивного метода оценивания параметров на основе GL-распределения, применение которого позволяет получить оценки максимального правдоподобия параметров регрессионных моделей при различных распределениях ошибок наблюдений на участках области планирования;
- получены соотношения для вычисления элементов информационной матрицы Фишера на основе универсального лямбда–распределения, предложен обобщенный алгоритм планирования эксперимента, который позволяет учитывать форму распределения ошибок;

- разработана программная система устойчивого и адаптивного оценивания параметров регрессии и планирования эксперимента.

Практическая значимость. Разработанные подходы позволяют восстанавливать регрессионные зависимости и планировать эксперимент в условиях отклонения ошибок от нормального закона, что дает возможность применять предложенные алгоритмы для широкого спектра практических задач. Разработанная программная система, позволяющая применить алгоритмы оценивания параметров регрессии и планирования эксперимента на практике, зарегистрирована в виде объекта интеллектуальной собственности как программа ЭВМ (№ гос. рег. 2011614692).

Реализация результатов работы. Научные и практические результаты нашли свое применение в ООО «ЗапСибГеоПроект» и в учебном процессе НГТУ, о чем имеются соответствующие акты внедрения.

Основные положения, выносимые на защиту:

- алгоритмы формирования оценочных подмножеств метода LTS на основе расстояний Кука, Велша-Куха, Махаланобиса и робастного расстояния;
- алгоритм построения оценочного подмножества, близкого к оптимальному плану, для схемы LTS-оценивания;
- алгоритм метода адаптивного оценивания параметров регрессионных зависимостей на основе обобщенного лямбда-распределения;
- способ вычисления элементов информационной матрицы Фишера на основе GL–распределения, обобщенный алгоритм синтеза планов с использованием универсального лямбда-распределения.

Апробация работы. Основные результаты работы докладывались и обсуждались на пятой международной научно-практической конференции «Высокие технологии, фундаментальные и прикладные исследования, образование», Санкт-Петербург, 2008г; на всероссийской научной конференции молодых ученых «Наука. Технологии. Инновации», Новосибирск, 2008-2010гг; на десятой международной научно-технической конференции «Актуальные проблемы электронного приборостроения» АПЭП-2010, Новосибирск, 2010г. Так-

же некоторые результаты проведенных исследований опубликованы в депонированных отчетах по научно-исследовательской работе.

Работа выполнена при поддержке стипендии Президента Российской Федерации на 2011-2012 учебный год согласно приказу Министерства образования и науки Российской Федерации № 2659 от 11.10.2011 г., ФЦП «Научные и научно-педагогические кадры инновационной России» на 2009-2013 гг. (проекты № П263, № 14.B37.21.0698), стипендии Правительства Новосибирской области на 2011г., научного студенческого гранта НГТУ 2008-2009 гг.

Публикации. По результатам исследований опубликовано 15 научных работ, общим объемом 5,49 п.л. (из них авторских 3,03 п.л.), включая: входящие в перечень рецензируемых научных журналов и изданий – 6, сборники научных трудов – 1, материалы трудов научно-технических конференций – 7, свидетельство о государственной регистрации программы для ЭВМ – 1.

Структура и объем работы. Диссертация состоит из введения, 5 разделов, заключения и списка литературы, состоящего из 95 источников, 3 приложений. Диссертация изложена на 175 страницах основного текста, содержит 46 рисунков и 41 таблицу.

СОДЕРЖАНИЕ РАБОТЫ

В разделе 1 представлен обзор наиболее известных подходов к поиску оценок неизвестных параметров регрессионного уравнения. В п.1.1 описана модель «черного ящика» и приведена постановка задач регрессионного анализа, планирования эксперимента в рамках активного и пассивного эксперимента.

Рассмотрим регрессионное уравнение вида

$$y = X\theta + \varepsilon, \qquad (1)$$

где $X = \begin{bmatrix} f_1(x_{1i_1}) & \cdots & f_m(x_{1i_m}) \\ \vdots & \ddots & \vdots \\ f_1(x_{ni_1}) & \cdots & f_m(x_{ni_m}) \end{bmatrix}$ – матрица плана, имеющая полный столбцовый ранг, т.е. $rg(X) = m$, m – количество регрессоров, n – количество испытаний, $f_1(x),...,f_m(x)$ – известные действительные функции, x_{ji_l} – заданные зна-

чения входных факторов $F_1, F_2, ..., F_k$ в n наблюдениях, $i_1, ..., i_m$ – номера входных факторов, $y = (y_1, ..., y_n)^T$ – вектор значений отклика, $\theta = (\theta_1, ..., \theta_m)^T$ – вектор неизвестных параметров, подлежащих оцениванию; $\varepsilon = (\varepsilon_1, ..., \varepsilon_n)^T$ – вектор независимых ошибок наблюдений. Имеют место предположения:

$$E(\varepsilon) = 0,\ E(\varepsilon\varepsilon^T) = \sigma^2 I,\ \sigma^2 < \infty, rg(X) = m. \qquad (2)$$

Кроме того для распределения случайной ошибки существуют и конечны первые четыре момента.

В диссертационной работе рассматриваются две схемы проведения эксперимента: пассивная и активная. В рамках пассивного эксперимента оценивание параметров регрессионных зависимостей проводится по заданной выборке наблюдений, полученной в соответствии с нормированным планом $\xi = \begin{Bmatrix} x_1 & x_2 & ... & x_s \\ p_1 & p_2 & ... & p_s \end{Bmatrix}$, где $\sum_{i=1}^{s} p_i = 1$, $p_i = \frac{n_i}{n}$, s – количество точек в спектре плана и n_i – число повторных наблюдений в i-ой точке. Если измерения проводились по случайным наблюдениям, то $s = n$, $p_i = 1/n$, $i = 1, 2, ..., n$. Задача регрессионного анализа состоит в том, чтобы по имеющимся исходным данным (значениям отклика и входных факторов) как можно точнее оценить вектор неизвестных параметров уравнения регрессии (1). Задача планирования в рамках активного эксперимента состоит в том, чтобы по имеющимся исходным данным построить оптимальный нормированный план эксперимента ξ.

В п. 1.2–1.8 рассмотрены наиболее известные подходы и методы поиска оценок вектора неизвестных параметров регрессионного уравнения, такие как МНК, ММП, метод наименьших модулей, М-оценки, оценки Хьюбера, знаковый метод, метод LTS, метод LMS, L_V-оценки, оценки на основе универсальных распределений и др. Также рассмотрены базовые понятия теории планирования эксперимента, рассмотрены критерии A-, D-, E-оптимальности плана эксперимента и критерий экстраполяции в точку, приведено описание классического алгоритма построения планов эксперимента. В данном разделе также

проведен обзор существующих программных систем, применимых для оценивания параметров регрессии и построения планов эксперимента.

Раздел 2 посвящен описанию предложенных автором модификаций метода LTS, рангового метода и алгоритма адаптивного оценивания на основе GL-распределения.

Рассмотрим модель наблюдений вида (1). При классической реализации метода LTS формирование оценочного подмножества размерности h $\left((n+m+1)/2 \le h \le n\right)$ производится только исходя из величины остатков. Однако при формировании оценочного подмножества можно использовать информацию о разбросе и точности наблюдений, например, сопоставляя расстояния Махаланобиса и соответствующие стандартизированные остатки.

Расстояние Махаланобиса MD_i вычисляется для каждой точки исходных данных по соотношению:

$$MD_i = \sqrt{P_{ii}(n-1) - 1 + 1/n}\,,\ i = 1,...,n\,,$$

где P_{ii} – i-й диагональный элемент проекционной матрицы $P = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$.

В зависимости от величины стандартизованных остатков $r_i = \dfrac{e_i}{\hat{\sigma}}$ ($\hat{\sigma}$ – оценка среднеквадратического отклонения ε_i, e_i – i-й остаток) и соответствующих значений MD_i наблюдения можно подразделить на четыре класса: регулярные наблюдения, вертикальные выбросы, «хорошие» горизонтальные выбросы, «плохие» горизонтальные выбросы. Автором предлагается в оценочное подмножество добавлять наблюдения как представлено на рис.1. Особенностью данного способа формирования оценочного подмножества является то, что в первую очередь в него включаются регулярные наблюдения, затем «хорошие» горизонтальные выбросы и далее наблюдения из зон вертикальных и «плохих» горизонтальных выбросов, соответствующие минимальным значениям r_i.

Для учета разброса наблюдений вместо расстояния Махаланобиса можно использовать робастное расстояние, которое в каждой точке вычисляется следующим образом:

$$RD_i = \sqrt{\left(\overline{F}_i - \overline{T}\right)^T \left(Cov\left(\overline{F}\right)\right)^{-1} \left(\overline{F}_i - \overline{T}\right)},\ i = 1,...,n,$$

где $\overline{F}_i$ – вектор, элементами которого являются i-ые реализации каждого входного фактора (см. п.1.1), $\overline{T}$ – вектор, состоящий из робастных оценок средних, которые при использовании метода LTS следует определять по h близким наблюдениям для каждого входного фактора F_j, $j = 1,...,k$, $Cov\left(\overline{F}\right)$ – ковариационная матрица входных факторов F_j. В оценочное подмножество наблюдения предложено добавлять как представлено на рис.1.

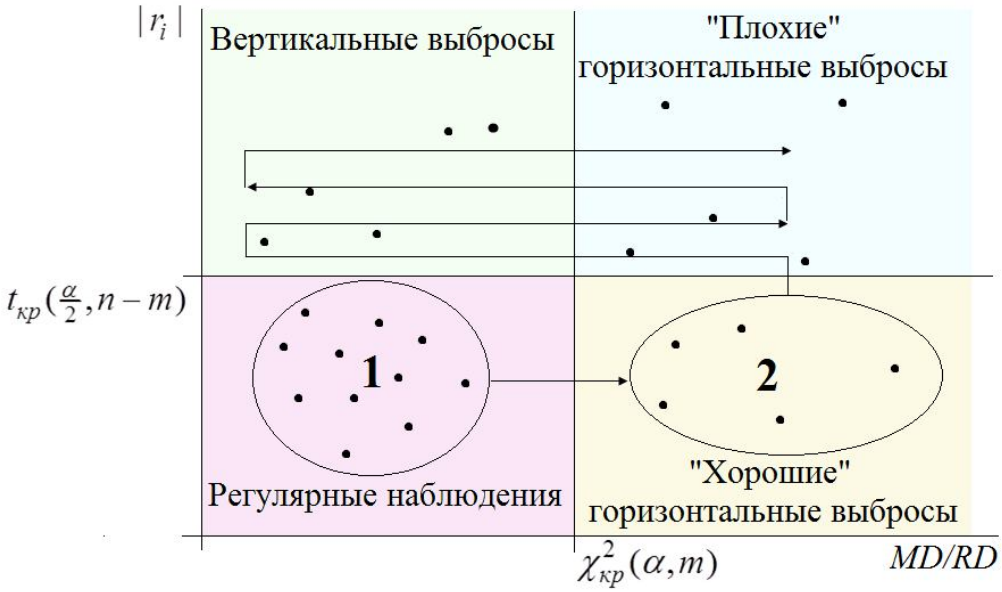


Рис.1. Схема формирования оценочного подмножества
с использованием иформации о характере наблюдений

Учитывать разброс наблюдений и величину остатков можно путем вычислений расстояний Кука и Велша-Куха, которые определяются по формулам:

$$C_i = \frac{1}{m}\frac{P_{ii}}{1-P_{ii}}(r_i^t)^2 \text{ и } WK_i = \left|r_i^{t*}\right|\sqrt{\frac{P_{ii}}{1-P_{ii}}},\ i = 1,...,n,$$

где стьюдентизированные остатки r_i^t и внешние стьюдентизированнные остатки r_i^{t*} вычисляются по формулам:

$$r_i^t = \frac{e_i}{\hat{\sigma}\sqrt{1-P_{ii}}} \text{ и } r_i^{t*} = r_i^t\sqrt{\frac{n-m-1}{n-m-(r_i^t)^2}},\ i = 1,...,n.$$

В этом случае в оценочное подмножество автором предлагается добавлять наблюдения с минимальным значением расстояния Кука либо Велша-Куха.

Кроме того в работе [12] предложена модификация рангового метода на основе расстояния Махаланобиса, алгоритм которой представлен ниже. Такой подход позволяет учесть величину остатков и разброс наблюдений, используя информацию о ранге остатка и о расстоянии Махаланобиса.

Шаг 1. В качестве начального приближения выбирается МНК-оценка $\hat{\theta}^0 = \left(X^T X\right)^{-1} Xy$.

Шаг 2. Вычисляются остатки e_i и расстояния Махаланобиса MD_i, $i = 1,...,n$.

Шаг 3. Остатки упорядочиваются по модулю и вычисляются их ранги $\pi(i)$.

Шаг 4. Вычисляются веса каждого наблюдения w_i

$$w_i = \begin{cases} 1, \text{ если } \pi(i) < h^*, \\ 0.75, \text{ если } h^* \le \pi(i) \le h \text{ и } MD_i \le \chi^2_{кр}(\alpha, m), \\ 0.5, \text{ если } h^* \le \pi(i) \le h \text{ и } MD_i > \chi^2_{кр}(\alpha, m), \\ 0, \text{ если } \pi(i) > h, \end{cases} \text{ где } h^* = \frac{n+m+1}{2}.$$

Шаг 5. Вычисляется оценка $\hat{\theta}^1 = (X^T W X)^{-1} X^T W y$, где $W = diag(w_i)$;

Шаг 6. Вычисление $\Delta = \max_j \left|\hat{\theta}^0_j - \hat{\theta}^1_j\right|$, $j = 1,..,m$, $\hat{\theta}^0 = \hat{\theta}^1$.

Шаг 7. Если не достигнута требуемая точность δ ($\Delta > \delta$), переход на шаг 2, иначе вычисления заканчиваются и $\hat{\theta} = \hat{\theta}^0$.

Для ситуации, когда распределение ошибок наблюдений отклоняется от нормального закона, в данном разделе предложен алгоритм адаптивного метода оценивания параметров регрессионных зависимостей на основе GL-распределения, который представлен на рис.2.

Так как ошибки наблюдений являются независимыми, то логарифм функции правдоподобия в данном случае имеет вид:

$$\ln(L(e,\theta)) = \sum_{i=1}^{n} \ln(g_{e_i}(z_i)),$$

где функция плотности GL-распределения определяется как

$$g_{e_i}(z_i) = \frac{\lambda_2}{u_i^{\lambda_3 - 1} + (1-u_i)^{\lambda_4 - 1}},\ 0 \le u_i \le 1,$$

$$z_i = y_i - X_i\theta = Q_{e_i}(u, \lambda_1, \lambda_2, \lambda_3, \lambda_4) = \lambda_1 + \frac{1}{\lambda_2}\left[\frac{u_i^{\lambda_3}}{\lambda_3} - \frac{(1-u_i)^{\lambda_4}}{\lambda_4}\right].$$

Величина шага $\Delta\hat{\theta}^k$ задается алгоритмом решения оптимизационной задачи поиска максимума логарифма функции правдоподобия, в качестве которого автором использовался симплексный метод Нелдера-Мида.

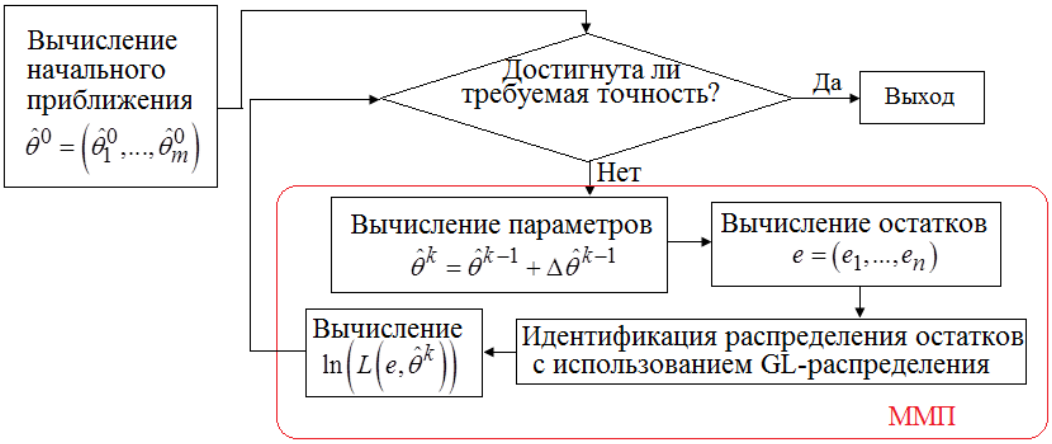


Рис.2. Алгоритм адаптивного оценивания параметров регрессионных моделей

В п. 2.5. представлены результаты исследования предложенных алгоритмов. В качестве исследуемой использовалась следующая модель:

$$y_i = \theta_1 + \theta_2 x_{i2} + \theta_3 x_{i3} + \varepsilon_i,\ i = 1, \ldots, n, \qquad (3)$$

где $m = 3$, $n = 200$, значения входных факторов x_{ij} выбирались из интервала $(0,1)$, $\theta_{ucm} = (25,25,25)^T$, ε_i независимые случайные величины, функция распределения которых имеет вид:

$$F(x) = (1-\mu)F_1(x,0,\sigma_1) + \mu F_2(x,0,\sigma_2), \qquad (4)$$

где $F_i(x,0,\sigma_i)$ – функция нормального распределения с нулевым математическим ожиданием и дисперсией σ_i^2, μ – доля выбросов, $\mu \in [0,1]$, $i = 1,2$. При

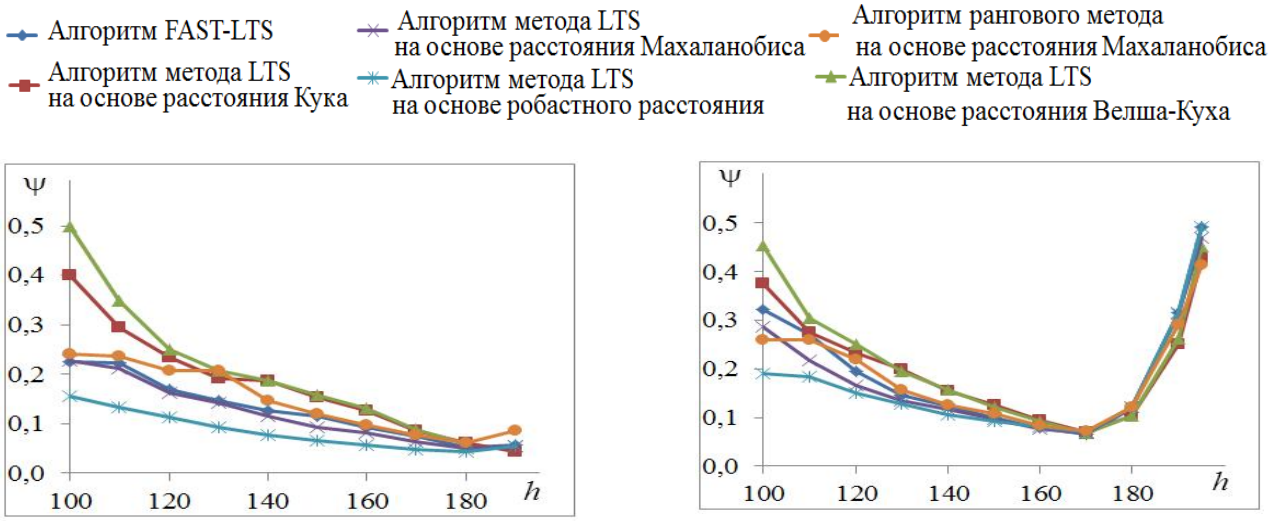
моделировании задавались не сами значения дисперсий, а соответствующие уровни шума, которые определяются отношением «шум»/«сигнал» в %.

В качестве показателя точности оценивания неизвестных параметров использовалась величина:

$$\psi = \left(\hat{\theta}_i - \theta_i^{ucm}\right)^T \left(\hat{\theta}_i - \theta_i^{ucm}\right). \qquad (5)$$

Для различных комбинаций μ и h проводилось по 100 вычислительных экспериментов. Каждый эксперимент заключался в моделировании выборки исходных данных в соответствии с моделью (3) и последующим оцениванием ее параметров. В качестве итоговых показателей точности оценивания ψ использовалось усредненное по 100 экспериментам значение.

На рис. 3 представлены зависимости качества оценивания параметров от размера оценочного подмножества для случаев, когда в выборке присутствуют выбросы, доля которых составляет 10% (см. рис.3 а)) и 20% (см. рис.3 б)) соответственно, с уровнем шума ошибок $\rho_1 = 5\%$ и выбросов – $\rho_2 = 50\%$.



а) б)

Рис.3. Зависимость качества оценивания параметров регрессии от размера оценочного подмножества, доля выбросов а) – 10%; б) – 20%

По рис. 3 видно, что при малых размерах оценочного подмножества наилучшие оценки параметров дают методы LTS на основе расстояний Махаланобиса и робастного. Все модификации метода LTS показали наиболее точ-

ные результаты оценивания при размере оценочного подмножества $h = (1-\mu)n$. При размере оценочного подмножества $h > (1-\mu)n$ наиболее точные результаты оценивания показывают алгоритмы метода LTS на основе расстояний Кука и Велша-Куха. В работах [1,11] представлены более полные результаты проведенных исследований.

В п. 2.5.2 – 2.5.4 приведены результаты исследования алгоритмов метода LTS, LMS и рангового метода при различных распределениях ошибок наблюдений. В частности показано, что при асимметричном распределении ошибок наблюдений методы LTS и LMS дают менее точные результаты оценивания, чем в случае представления ошибок в виде (4), поэтому в таком случае для оценивания можно предложить использовать адаптивный метод. В табл.1 приведены результаты оценивания параметров регрессионной модели (1), полученные с использованием алгоритмов методов LTS, LMS, МНК и адаптивного оценивания для случаев, когда функция распределения ошибки представлена в виде (4) с уровнями шума $\rho_1 = 5\%$ и $\rho_2 = 15\%$, долей выбросов $\mu = 0.1$ и когда ошибка имеет GL-распределение с параметрами (0,1,0.04,0.3).

Таблица 1

Точность оценивания параметров при различных распределениях ошибок

Распределение ошибок	Смесь двух нормальных, $\rho_1 = 5\%$ и $\rho_2 = 15\%$, $\mu = 0.1$				GLD(0,1,0.04,0.3)			
Метод оценивания	$\hat{\theta}_0$	$\hat{\theta}_1$	$\hat{\theta}_2$	ψ	$\hat{\theta}_0$	$\hat{\theta}_1$	$\hat{\theta}_2$	ψ
МНК	25,02	24,92	25,06	1,07E-02	24,77	25,22	25,00	9,96E-02
LTS	24,99	24,96	24,95	4,63E-03	25,04	25,15	25,18	5,59E-02
LMS	25,15	24,82	24,84	8,05E-02	25,03	24,96	25,05	4,82E-03
Адаптивное оценивание	25,00	24,96	25,04	3,48E-03	25,03	25,22	25,05	5,05E-02

Как видно из табл.1, при появлении выбросов наиболее точные результаты дают методы адаптивного оценивания и LTS. Это свидетельствует о том, что предложенный алгоритм адаптивного оценивания тоже обладает свойством устойчивости и его можно рекомендовать для оценивания при наличии в выборке выбросов. Также из табл.1 видно, что применение предложенного метода

адаптивного оценивания при асимметричном распределении ошибок приводит к более точным результатам по сравнению с МНК и методом LTS. Наиболее точные результаты в данном случае дает метод LMS. Однако, несмотря на незначительный проигрыш в точности, оценки, полученные адаптивным методом, обладают свойствами асимптотической эффективности. Другие результаты проведенных исследований представлены в работах [6,7].

В разделе 3 описаны подходы к построению планов в рамках активного и пассивного эксперимента.

В п.3.1 показано, что наличие выбросов оказывает влияние не только на оценки неизвестных параметров, но и на информационную матрицу Фишера. Это означает, что если рассматривать оценочное подмножество как самостоятельный план эксперимента, то появляется возможность совместного использования вычислительной схемы LTS-оценивания и методов планирования эксперимента. Действительно, процесс формирования оценочных подмножеств можно проводить не только исходя из условия минимальных остатков схемы LTS, но и с учетом выбранного критерия оптимальности плана эксперимента. Полученный в результате алгоритм управления выборкой позволяет максимально приблизить фактически используемый план эксперимента (соответствующий текущему оценочному подмножеству) к оптимальному. Эффект использования таких алгоритмов будет более ощутимым если процесс сбора исходных данных соответствовал оптимальному плану.

Дело в том, что как только появляются выбросы, оценки дисперсии случайной ошибки в каждой точке спектра плана становятся достаточно большими и различными, в результате классические условия оптимальности не выполняются. В связи с этим для оценивания степени отклонения текущего плана в схеме LTS-оценивания от оптимального предложены следующие меры близости для критериев D- и A-оптимальности соответственно:

- $\varphi^{D}(\xi) = \max\limits_{x \in \xi} \lambda(x) d(x, \xi) - \min\limits_{x \in \xi} \lambda(x) d(x, \xi)$;

- $\varphi^A(\xi) = \max\limits_{x\in\xi} \lambda(x) f^T(x) M^{-2}(\xi) f(x) - \min\limits_{x\in\xi} \lambda(x) f^T(x) M^{-2}(\xi) f(x)$.

Очевидно, что если план ξ является оптимальным, то $\varphi(\xi) = 0$.

Для того чтобы максимально возможно сохранить оптимальные свойства исходного плана эксперимента автором предложен алгоритм построения оценочного подмножества, близкого к оптимальному плану, представленный ниже.

- Для заранее заданного h выполняется метод LTS до сходимости.
- В каждой точке спектра начального плана ξ_0 производится сортировка наблюдений в порядке возрастания остатков.
- Вычисляется количество точек, входящих в оценочное подмножество, которым соответствуют минимальные остатки:

$$n_1 = \gamma + \eta h,\ \eta \in [0,1],\ \gamma = \begin{cases} 0, & \eta \neq 0 \\ a, & \eta = 0 \end{cases},$$

где a – минимальное число точек в оценочном подмножестве, например, если используется линейное регрессионное уравнение, то $a = 2$, квадратическое – $a = 3$ и так далее; конкретное значение η выбирается заранее. Такое представление позволяет учитывать два граничных случая. При $\eta = 1$ в оценочное подмножество входят только те наблюдения, которые соответствуют минимальным остаткам; при $\eta = 0$ оценочное подмножество формируется, используя только алгоритмы планирования эксперимента.

- В новое оценочное подмножество для каждой точки спектра плана записываются по n_1 / s наблюдений, соответствующие минимальным остаткам в этих точках. В результате получается равновесный план, содержащий s точек

$$\xi_k = \begin{Bmatrix} x_1 & x_2 & \dots & x_s \\ 1/s & 1/s & \dots & 1/s \end{Bmatrix}.$$

Значение счетчика k устанавливается равным n_1.

- Пока $k < h$ выполняется следующая последовательность действий:
  - находится точка, для которой выполняется $x^* = Arg \max\limits_{x\in\xi_0} \phi(x, \xi)$,

где ξ - план, состоящий из $k+1$ точек, в который входят k точек плана ξ_k и точка x из плана ξ_0, которая не вошла в ξ_k;

- x^* добавляется в оценочное подмножество $\xi_{k+1} = \left(1 - \frac{1}{k}\right)\xi_k + \frac{1}{k}\xi(x^*)$;
- увеличиваем значение счетчика на единицу $k = k+1$.

Функционал $\phi(x,\xi)$ для критерия A-оптимальности имеет вид $\phi(x,\xi) = f^T(x)M^{-2}(\xi)f(x)$, для D-оптимальности: $\phi(x,\xi) = f^T(x)M^{-1}(\xi)f(x)$, где $f(x) = \left(f_1(x),...,f_m(x)\right)^T$.

В качестве исследуемой была взята модель (3), где значения входных факторов x_{ij} выбирались в соответствие с являющимся одновременно A- и D-оптимальным планом $\xi = \begin{Bmatrix} (-1,-1) & (-1,1) & (1,-1) & (1,1) \\ 1/4 & 1/4 & 1/4 & 1/4 \end{Bmatrix}$. Случайные ошибки ε_i, $i = 1,...,n$ моделировались независимыми и одинаково распределенными с функцией распределения вида (4). В качестве показателя точности оценивания параметров регрессии было взято соотношение (5). В качестве итоговых результатов представлены усредненные по 150 экспериментам значения.

На рис. 4 представлены результаты, полученные для случая с 10% выбросов ($\mu = 0.1$), уровнями шума $\rho_1 = 5\%$ и $\rho_2 = 50\%$.

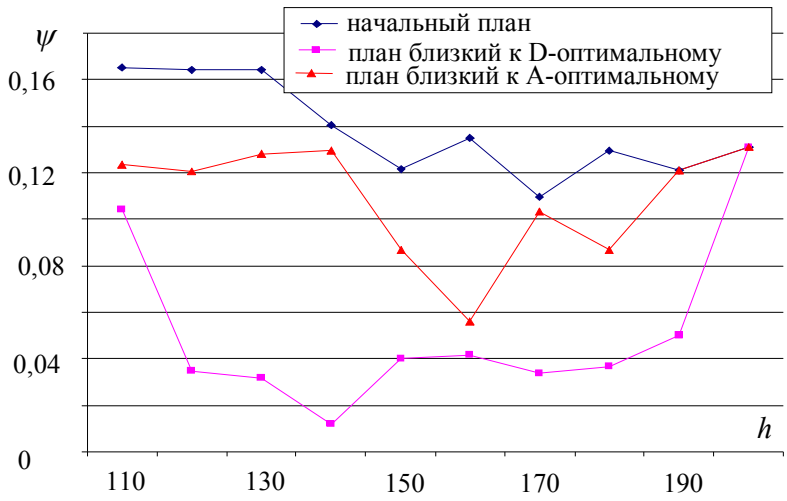


Рис.4. Зависимость точности оценивания параметров регрессии от размера оценочного подмножества

По рис.4 видно, что при построении оценочного подмножества в соответствии с предложенным алгоритмом точность оценивания параметров выше. Наиболее точными получаются оценки, полученные на плане близкому к D-оптимальному. Другие результаты исследования работы алгоритма построения оценочного подмножества для метода LTS, представлены в работах [2,4,5].

В рамках активного эксперимента в п.3.2 предложен способ вычисления элементов информационной матрицы Фишера, который позволяет учитывать на области планирования как неоднородность дисперсии, так и неоднородность формы распределения. Для этого были сформулированы следующие утверждение и следствие [9,10], доказательства которых приведены в [10].

Утверждение. *Для регрессионной модели (1) с независимыми и имеющими одинаковое GL–распределение ошибками* ε_i, $i = 1,...,n$, *элементы информационной матрицы вычисляются по следующей формуле:*

$$M_{jk} = -\sum_{i=1}^{n} f_j(x_{ij}) f_k(x_{ik}) \int_0^1 g''_{\varepsilon_i}(z_i) g_{\varepsilon_i}(z_i) du, \; j = 1...m, \; k = 1...m.$$

Следствие. *В условиях использования нормированного плана эксперимента элементы информационной матрицы вычисляются по формуле:*

$$M = \sum_{i=1}^{s} \lambda(x_i) p_i f(x_i) f^T(x_i)$$

где соотношение для вычисления функции эффективности имеет вид:

$$\lambda(x_i) = -\int_0^1 g''_{\varepsilon_i}(z_i) g_{\varepsilon_i}(z_i) du \quad , \; i = 1,...,s \, . \qquad (6)$$

На основе утверждения и следствия автором был разработан и реализован алгоритм планирования эксперимента, который обобщает классический алгоритм, предложенный В.В.Федоровым, на случай, когда ошибки имеют GL-распределение. В предложенном алгоритме вычисление функции эффективности проводится по формуле (6) с использованием метода трапеций, поскольку интеграл (6) не выражается в элементарных функциях.

В п.3.4 приведены результаты построения оптимальных планов с использованием разработанного алгоритма. В качестве истинной зависимости использовалась следующая модель:

$$y_i = \theta_0 + \theta_1 x_{i1} + \varepsilon_i,\ i = 1,...,n, \qquad (7)$$

где количество регрессоров $m = 2$, область планирования $[-1,1]$, $\theta^{ucm} = (25,25)^T$, случайные ошибки ε_i, $i = 1,...,n$ независимые и имеют GL–распределение. Задача состоит в построении оптимального плана ξ.

В случае, когда ошибки имеют нормальное распределение и имеет место лишь неоднородность дисперсий на области планирования, результаты синтеза оптимального плана с использованием классического и предложенного алгоритмов совпадают, что подтверждает корректность работы последнего.

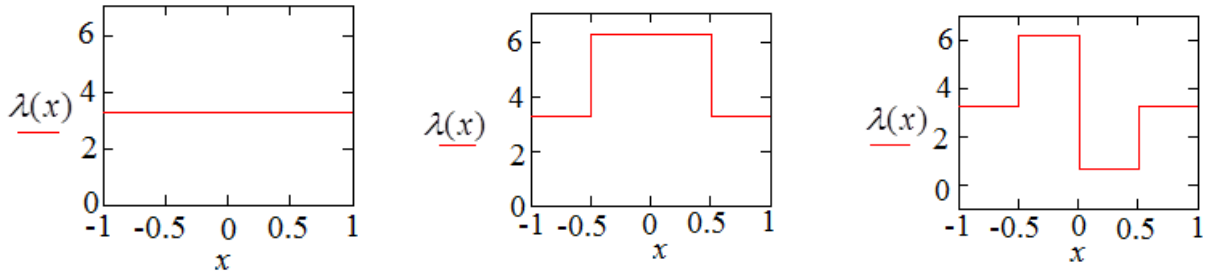
Для случая, когда распределение ошибок на области планирования различно были получены следующие результаты. В качестве исследуемых были взяты следующие GL–распределения: несимметричное с левой асимметрией $GLD_1(0,1,0.002,0.5)$, несимметричное с правой асимметрией $GLD_2(0,1,0.5,0.002)$, симметричное $GLD_3(0,1,0.5,0.5)$ и близкое к распределению Вейбулла $GLD_4(0,1,0.04,0.3)$. Распределения ошибок ε_i, $i = 1,...,n$ на области планирования имеет вид:

I. GLD_1, при $x \in [-1,0)$ и GLD_2, при $x \in [0,1]$;

II. GLD_1, при $x \in [-1,-0.5)$, GLD_3, при $x \in [-0.5,0.5]$, GLD_2, при $x \in (0.5,1]$;

III. GLD_1, при $x \in [-1,-0.5)$, GLD_3, при $x \in [-0.5,0)$, GLD_4, при $x \in [0,0.5]$ и GLD_2, при $x \in (0.5,1]$.

На рис.6 представлены графики функции эффективности, вычисленной при помощи соотношения (6). Следует отметить, что ее значения характеризуют неоднородность формы распределения на области планирования.



а) б) в)

Рис.6. Функции эффективности для случаев а) – I; б) – II; в) – III

В табл.2 представлены результаты применения обобщенного алгоритма планирования эксперимента. В случае I построенный план совпадает с классическим планом, что объясняется постоянной функцией эффективности (см. рис.6 а)). В случаях II и III построенные оптимальные планы отличаются от классических, что является следствием неоднородности формы распределения (см. рис.6 б)-в)). Также следует отметить, что для всех трех случаев выполняется условие D-оптимальности плана эксперимента. Это свидетельствует о том, что построенные планы являются оптимальными.

Таблица 2

Результаты построения оптимальных планов и точность оценивания параметров при отличном от нормального распределении ошибок наблюдений

Распределение ошибок		Случай I	Случай II	Случай III
Оптимальный план		$\xi^*=\begin{Bmatrix}-1 & 1\\ 0.5 & 0.5\end{Bmatrix}$	$\xi^*=\begin{Bmatrix}-1 & -0.5 & 0.5 & 1\\ 0.30 & 0.20 & 0.19 & 0.31\end{Bmatrix}$	$\xi^*=\begin{Bmatrix}-1 & -0.5 & 1\\ 0.12 & 0.39 & 0.49\end{Bmatrix}$
$\lambda(x)d(x,\xi^*),\ x\in\xi^*$		2.001	2.002	2.002
МНК	ψ	2.654E-02	9.650E-02	1.213E-01
	$\det M^{-1}$	2.851E-01	5.188E+00	3.560E+00
Адаптивный метод	ψ	2.654E-02	2.272E-02	6.798E-02
	$\det M^{-1}$	4.564E-02	8.439E-03	8.589E-02
φ		0.4001	0.400	0.289

1 В табл.2 показано, что использование построенных планов позволяет повысить качество оценивания. Для регрессионной модели (7), при $n=1000$ приведены результаты исследования качества оценивания параметров, значения

определителя дисперсионной матрицы и эффективности плана. При этом использовались методы наименьших квадратов и адаптивного оценивания на основе GL-распределения. Измерения проводились в соответствии с построенными оптимальными планами (см. табл.2). Оценка эффективности плана определялась следующим образом:

$$\varphi = \sqrt[s]{\left|M^{-1}(\xi^*)\right| \Big/ \left|M^{-1}(\xi_0)\right|}\,.$$

В качестве итоговых показателей точности ψ, эффективности плана φ и определителя дисперсионной матрицы использовались усредненные по 100 экспериментам значения. По табл.2 видно, что построенные планы являются эффективными и применение на этих планах адаптивного метода оценивания дает более точные результаты. Также результаты исследования обобщенного алгоритма представлены в работе [10].

Также в п.3.4 представлены результаты синтеза планов для других моделей, в том числе на двумерной области планирования.

В разделе 4 приведено описание разработанной программной системы оценивания параметров регрессионных моделей и планирования эксперимента.

В п.4.1 описаны задачи, которые решаются с ее использованием, перечислены режимы работы, показаны взаимосвязи между ними. Также представлено алгоритмическое наполнение программной системы.

В п. 4.2 и в работах [3,8,15] приведено подробное описание каждого из режимов работы программной системы.

В разделе 5 показано применение методов LTS, его модификаций и LMS для реальной технической задачи оценки параметров кривых провисания троса, приведены результаты идентификации GL-распределения остатков, применен алгоритм планирования эксперимента на основе GL-распределения для нахождения координат максимально информативных точек [13,14]. Также применены методы планирования к задаче прогнозирования покупательского спроса, построены D-оптимальные планы.

В заключении сформулированы основные результаты работы, которые сводятся к следующему.

1. Сформулировано и доказано утверждение о вычислении элементов информационной матрицы Фишера с использованием GL–распределения, на его основе предложен обобщенный алгоритм планирования эксперимента.
2. Предложены, реализованы и исследованы схемы формирования оценочного подмножества для метода LTS на основе расстояний Кука, Велша-Куха, Махаланобиса и робастного расстояния и модификация рангового метода на основе расстояния Махаланобиса.
3. Разработан и исследован алгоритм адаптивного оценивания неизвестных параметров регрессионного уравнения, основанный на идентификации распределения остатков с использованием универсального GL-распределения.
4. Предложен, реализован и исследован алгоритм формирования оценочного подмножества на основе критериев A- и D- оптимальности для схемы LTS-оценивания.
5. Разработанные алгоритмы включены в программную систему устойчивого и адаптивного оценивания параметров регрессионных моделей и планирования эксперимента.
6. С помощью разработанных алгоритмов решены техническая и экономическая практические задачи.

СПИСОК ПУБЛИКАЦИЙ ПО ТЕМЕ ДИССЕРТАЦИИ

1. Вострецова Е. А. Устойчивое оценивание параметров регрессионных моделей с использованием идей метода наименьших квадратов / В.С. Тимофеев, Е.А. Вострецова // Научн. вестн. НГТУ. – Новосибирск: Изд-во НГТУ, 2007. – N2(27). – С.57-67.
2. Вострецова Е.А. Устойчивое оценивание параметров регрессии при использовании оптимальных планов эксперимента /Е.А. Вострецова // Материалы всерос. научной конф. молодых ученых «Наука. Технологии. Инновации» в 7 частях. – Новосибирск: Изд-во НГТУ, 2007. – Ч.1. – С.27-28.

3. Вострецова Е.А. Программная система планирования эксперимента и устойчивого оценивания параметров регрессионных моделей / Е.А. Вострецова // Материалы всерос. научной конф. молодых ученых «Наука. Технологии. Инновации» в 7 частях. – Новосибирск: Изд-во НГТУ, 2008. – Ч.1. – С.7-9.

4. Вострецова Е.А. Адаптация алгоритмов метода наименьших взвешенных квадратов к использованию на оптимальных планах эксперимента / В. С. Тимофеев, Е.А. Вострецова // Высокие технологии, фундаментальные и прикладные исследования, образование. Сб. трудов пятой междунар. научн.-практич. конф. СПб., 28-30апр.2008г. – СПб.: Изд-во Политех. Ун-та, 2008. – Т.12. – С.120-121.

5. Вострецова Е.А. Использование алгоритмов планирования эксперимента в схеме LTS-оценивания / В.С. Тимофеев, Е.А. Вострецова // Научн. вестн. НГТУ. – Новосибирск: Изд-во НГТУ, 2009. – N1(34). – С.95-105.

6. Хайленко Е.А. Исследование распределений остатков при устойчивом оценивании с использованием обобщенного лямбда-распределения / Е.А. Хайленко // Материалы всерос. научной конф. молодых ученых «Наука. Технологии. Инновации» в 6 частях. – Новосибирск: Изд-во НГТУ, 2009.–Ч.1. –С.55-56.

7. Хайленко Е.А. Адаптивное оценивание параметров регрессионных моделей с использованием обобщенного лямбда – распределения / В.С. Тимофеев, Е. А. Хайленко//Доклады академии наук высшей школы РФ. – Новосибирск: Изд-во НГТУ, 2010. – N2 (15). – С.25-36.

8. Тимофеев В. С. Программная система устойчивого и адаптивного оценивания параметров регрессии и планирования эксперимента / В.С. Тимофеев, Е.А. Хайленко // Актуальные проблемы электронного приборостроения АПЭП-2010: Материалы X междунар. конф., Новосибирск, 22-24 сент. 2010г. – Новосибирск: Изд-во НГТУ, 2010. – Т.6. – С.73-79.

9. Хайленко Е.А. Построение информационной матрицы для регрессионных моделей с использованием обобщенного лямбда-распределения / Е.А. Хайленко // Материалы всерос. научной конф. молодых ученых «Наука. Технологии. Инновации» в 6 частях. – Новосибирск: Изд-во НГТУ, 2010.–Ч.1.– С.39-40.

10. Хайленко Е.А. Оптимальное планирование эксперимента для регрессионных моделей с обобщенным лямбда-распределением ошибок / В.С. Тимофеев, Е.А. Хайленко // Научн. вестн. НГТУ, 2011. – N1(42). – С.27–37.

11. Хайленко Е.А. Модификации метода LTS для устойчивого оценивания параметров регрессионных моделей / Е.А. Хайленко // Сборник научных трудов НГТУ, 2011. – N1(63). – С.75-82.

12. Хайленко Е.А. Модификации рангового метода для устойчивого оценивания параметров регрессионных моделей / Е.А. Хайленко // Материалы всерос. научной конф. молодых ученых «Наука. Технологии. Инновации» в 6 частях. – Новосибирск: Изд-во НГТУ, 2011. – Ч.1. – С.133-134.

13. Хайленко Е.А. Планирование уточняющих наблюдений при контроллинге воздушных линий по данным лазерного сканирования / В.И. Денисов, В.С. Тимофеев, Е.А. Хайленко // Сибирский журнал индустриальной математики. – Новосибирск: СО РАН, 2012. – Т.XV. – № 2(50). – С.75-85.

14. Оценивание уравнений кривых провисания воздушных линий устойчивыми методами / В.С. Тимофеев, В.Ю. Щеколдин, Е.А. Хайленко, Д.В. Харьковский // Прикладная информатика, 2012. – N3(39). – С.33-42

15. Свидетельство на программу для ЭВМ 2011614692 Российская Федерация. Программная система устойчивого и адаптивного оценивания параметров регрессионных моделей и планирования эксперимента / В.И. Денисов, В.С. Тимофеев, Е.А. Хайленко; правообладатель НГТУ. – 2011613035; заявл. 28.04.11; зарегистрировано 15.06.11. – 1с. – Тип ЭВМ: IBM PC – совместимый с ПК; язык: C++; ОС: Microsoft Windows 9X/NT/2000/2003/XP; объем: 1,56 Мб.

Подписано в печать 15.04.2013 г. Формат 60 × 84 × 1/16

Бумага офсетная. Тираж 100 экз. Печ. л. 1.5.

Заказ №

Отпечатано в типографии

Новосибирского государственного технического университета

630073, г. Новосибирск, пр-т К. Маркса, 20