На правах рукописи

Лери Марина Муксумовна

ИССЛЕДОВАНИЕ И РАЗРАБОТКА ПРАВИЛ ВЫБОРА МЕТОДОВ АНАЛИЗА ДАННЫХ ДЛЯ ИНТЕЛЛЕКТУАЛИЗИРОВАННЫХ СИСТЕМ ПРИКЛАДНОЙ СТАТИСТИКИ

05.13.18 — математическое моделирование, численные методы и комплексы программ

АВТОРЕФЕРАТ
диссертации на соискание ученой степени кандидата технических наук

Петрозаводск 2006

Работа выполнена в Институте прикладных математических исследований Карельского научного центра РАН.

Научный руководитель доктор физико-математических наук, профессор Павлов Юрий Леонидович.

Официальные оппоненты:
доктор технических наук, доцент Рогов Александр Александрович,
доктор технических наук, с.н.с. Белашев Борис Залманович.

Ведущая организация Нижегородский государственный университет им. Н.И. Лобачевского.

Защита состоится 3 ноября 2006 г. в 14 часов 00 мин. на заседании диссертационного совета Д 212.190.03 при Петрозаводском государственном университете по адресу: 185910, Петрозаводск, пр. Ленина, 33.

С диссертацией можно ознакомиться в библиотеке Петрозаводского государственного университета.

Автореферат разослан "14" сентября 2006 г.

Ученый секретарь
диссертационного совета В. В. Поляков.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы. Использование методов прикладной статистики за последние десятилетия вышло на качественно новый уровень и приобрело массовый характер. Все более широкое распространение получают программные средства, предназначенные для статистического анализа данных. Однако, до сих пор острой остается проблема некорректного использования статистического программного обеспечения. Это связано с тем, что уровень статистического образования многих пользователей оказывается недостаточным, а наиболее известные зарубежные продукты почти не содержат функций помощи при выборе метода анализа данных и обучения работе с существующими методами. Поэтому, одним из актуальных направлений разработки отечественного статистического программного обеспечения остается работа над его интеллектуализацией, т.е. над созданием программных средств статистического анализа данных, предназначенных не только для решения задач методами прикладной статистики, но и содержащих развитые информационно-справочные и экспертные системы. Такие системы должны включать в себя сведения обо всех используемых в пакете понятиях и методах математической статистики и помогать пользователю при выборе метода решения задачи, режима работы соответствующей программы и при интерпретации полученных результатов. Рекомендации, предлагаемые системой, должны быть обоснованы, однако современная математическая теория на многие возникающие здесь конкретные вопросы ответа не дает, поэтому для разработки экспертных правил необходимо проводить соответствующие исследования.

Автор диссертации является одним из разработчиков системы "Статистик-Консультант" – специализированного методо-ориентированного статистического комплекса программ, исторически первого отечественного пакета такого типа, созданного в среде Windows, и получившего высокую оценку специалистов. В ходе создания экспертной системы пакета возникли вопросы, связанные с выбором методов анализа данных среди нескольких альтернатив.

В диссертации рассматриваются вопросы разработки рекомендаций по выбору метода анализа данных для двух групп методов. Рассмотрены выбор критерия согласия и выбор метода поиска наибо-

лее информативного множества признаков в линейном регрессионном анализе.

Цель исследования. Целью диссертационной работы является построение математических моделей и разработка рекомендаций, используемых при выборе методов анализа данных в интеллектуализированном программном обеспечении прикладной статистики.

Объекты исследования. Объектами исследования были датчики псевдослучайных чисел и две группы методов статистического анализа данных: три критерия согласия и шесть методов поиска наиболее информативного множества признаков (ПНИМП) в линейном регрессионном анализе.

Методы исследования. Исследование методов анализа данных и сравнение получаемых с их помощью результатов аналитическим путем представляется весьма затруднительным, особенно вследствие того, что необходимо учитывать различные условия возникновения данных. Одним из путей решения этой проблемы предлагается метод статистических испытаний (метод Монте-Карло), который приобретает все большую популярность при сравнении методов анализа данных. Этот метод и был выбран в качестве основного метода исследования в диссертации. Кроме того, использовались методы оценивания параметров, методы проверки статистических гипотез, методы линейного регрессионного анализа. Основную сложность при получении результатов, несмотря на существенную автоматизацию процесса, представлял объем вычислительных экспериментов, так, например, трудозатраты на проведение экспериментов составили не менее 2000 чел.-дней.

Научная новизна. Все основные результаты диссертации являются новыми. В частности, впервые получены модели зависимостей мощностей критериев согласия от уровня значимости, объема выборки и значений параметров проверяемых гипотез, разработаны рекомендации пользователям по выбору критериев согласия в статистическом программном обеспечении и по определению некоторых условий проведения экспериментов при их планировании. Также впервые проведено сравнение шести методов поиска наиболее информативного множества признаков в линейном регрессионном анализе по вероятности возникновения эффекта "вздувания" коэффициента детер-

минации. Получены модели взаимной зависимости числа регрессоров, включаемых в начальный набор, параметров методов и числа случаев ошибочной работы методов и разработаны рекомендации по выбору метода поиска наиболее информативного множества признаков в линейном регрессионном анализе.

Основные результаты диссертации, выносимые на защиту: На защиту выносятся:

1. Комплекс программ, реализующий систему датчиков псевдослучайных чисел и процедуры критериев согласия. Эти программы включены в статистический пакет "Статистик-Консультант".

2. Модели зависимостей функций мощностей критериев согласия Пирсона, Колмогорова-Смирнова и пустых ящиков от параметров проверяемых гипотез. На основе этих моделей разработаны рекомендации пользователям по выбору критериев согласия.

3. Модели взаимной зависимости параметров методов ПНИМП, числа регрессоров и вероятности возникновения эффекта "вздувания" коэффициента детерминации. Установлено, что этот эффект является главным лимитирующим фактором при выборе метода поиска наиболее информативного множества признаков. На основе построенных моделей разработаны рекомендации по выбору методов линейного регрессионного анализа, направленные на снижение вероятности возникновения эффекта "вздувания" коэффициента детерминации.

Связь работы с крупными научными программами, темами. Результаты диссертации были получены в рамках трех тем планов научно-исследовательских работ Института прикладных математических исследований Карельского научного центра РАН: "Исследование и разработка методов математической статистики и теории многокритериальных задач с целью их реализации в интеллектуализированных системах" (№ гос. регистрации 01.9.40009930), "Исследование и разработка методов создания интеллектуальных систем статистического анализа данных" (№ гос. регистрации 01.9.80009162) и "Разработка методов исследования случайных структур и их применения при принятии статистических решений" (№ гос. регистрации

01.200.202223). В 1996–1997г. исследования проводились при поддержке Российского Фонда Фундаментальных Исследований (грант 96-01-00162). В 2000г. работа была поддержана грантом конкурса персональных грантов для студентов, аспирантов и молодых ученых проведенного Администрацией Санкт-Петербурга, Министерством образования РФ и РАН при участии ФЦП "Интеграция". В 2001г. был получен грант Российского Фонда Фундаментальных Исследований (грант 01-01-10850) для участия в VI международной конференции "Computer Data Analysis and Modeling: Robustness and Computer Intensive Methods" (Минск, Белоруссия).

Апробация результатов диссертации. Основные результаты докладывались на международной конференции "Computer Data Analysis and Modeling" (Минск, 1995), Шестой научной конференции стран СНГ "Применение многомерного статистического анализа в экономике и оценке качества продукции" (Москва, 1997), Первом Всероссийском симпозиуме по прикладной и промышленной математике (Петрозаводск, 2000), Всероссийской научной школе "Математические методы в экологии" (Петрозаводск, 2001), Шестой международной конференции "Computer Data Analysis and Modeling: Robustness and Computer Intensive Methods" (Минск, 2001), Российско-Скандинавском симпозиуме "Probability Theory and Applied Probability" (Петрозаводск, 2006).

Публикация результатов. Основные результаты диссертации опубликованы в десяти работах, из них свидетельство об официальной регистрации программы для ЭВМ, три статьи в трудах международных конференций, две статьи в сборниках трудов Петрозаводского государственного университета и Института прикладных математических исследований Карельского научного центра РАН и четверо тезисов докладов на международных и всероссийских конференциях.

Структура и объем диссертации. Диссертация состоит из введения, трех глав, заключения, списка литературы и 4 приложений. Объем диссертации без приложений составляет 133 страницы, объем приложений – 81 страница. Список литературы содержит 61 наименование.

СОДЕРЖАНИЕ РАБОТЫ

Во введении дается обоснование актуальности темы диссертации, приводится краткое описание комплекса программ "Статистик-Консультант" и вклада автора в его создание, сформулированы цель работы и основные результаты, выносимые на защиту, дано описание структуры диссертации.

Первая глава посвящена вопросам, связанным с имитационным моделированием, в частности построению системы датчиков псевдослучайных чисел.

Целью исследования, описанного в первой главе, было построение совокупности, датчиков псевдослучайных чисел, позволяющих генерировать выборки, соответствующие 15-ти законам распределения, реализованным в системе "Статистик-Консультант" (10 непрерывных законов: равномерный, нормальный, логнормальный, хи-квадрат, Стьюдента, Фишера, бета, гамма, экспоненциальный, Колмогорова; и 5 дискретных: Бернулли, биномиальный, отрицательно-биномиальный, геометрический, Пуассона).

В **параграфе 1.1** рассматривается один из достаточно простых алгоритмов выработки псевдослучайных последовательностей, часто рекомендуемый в литературе. Это так называемый линейный конгруэнтный метод, где последовательность псевдослучайных чисел, равномерно распределенных на отрезке $[0, 1]$, порождается по формуле:

$$\xi_i = \frac{u_i}{m+1}, \quad u_i = (a \cdot u_{i-1} + c) \bmod m, \tag{1}$$

Этот датчик используется в алгоритмах получения псевдослучайных чисел, имеющих другие распределения. Следуя рекомендациям Д. Кнута, значения a, c и m были выбраны равными: $a = 314159256$, $c = 2718281829$, $m = 2^{31} - 1$.

В **параграфе 1.2** приводятся алгоритмы, реализующие все включенные в систему датчики.

В **параграфе 1.3** описана методика проверки качества рассматриваемой системы датчиков посредством метода статистических испытаний.

Известно, что качество линейного конгруэнтного датчика стандартного равномерного распределения (1), а следовательно и любых

связанных с ним датчиков других законов распределения, существенно зависит не только от выбранных значений a, c и m, но и от выбора начального значения датчика u_0. Автором были проведены исследования с целью нахождения такого начального значения u_0 датчика равномерного закона распределения, при использовании которого качество датчиков всех реализованных законов распределения было бы приемлемым.

В **параграфе 1.4** рассматриваются результаты вычислительных экспериментов по подбору значения u_0 и проверке качества полученной совокупности датчиков псевдослучайных чисел. В результате проведенных вычислительных экспериментов было выбрано значение $u_0 = 0175875379$. Новизна подхода состояла в том, что важный параметр датчика стандартного равномерного распределения u_0 подбирался не для обеспечения качества именно этого датчика, а для построения группы датчиков различных распределений с различными значениями параметров. Разработанная система датчиков использовалась при проведении исследований, описанных в главах 2 и 3 диссертации.

Таким образом, в результате проведенных исследований была разработана система датчиков псевдослучайных чисел, генерирующих последовательности, надежность согласования которых с соответствующими распределениями равномерна по всем 15-ти реализованным законам и достаточна для практического применения.

Вторая глава посвящена методам проверки статистических гипотез о распределении исследуемых случайных величин – критериям согласия. Цель исследования состояла в разработке рекомендаций пользователям статистического программного обеспечения по выбору этих критериев в конкретных ситуациях. Основой для таких рекомендаций послужили построенные модели функций мощности рассмотренных критериев согласия.

В **параграфе 2.1** приведена общая схема критериев согласия и даны описания трех критериев, реализованных в статистическом пакете "Статистик-Консультант": критерия χ^2 Пирсона, критерия Колмогорова-Смирнова и критерия пустых ящиков.

В **параграфе 2.2** приводятся алгоритмы программ проверки гипотез о соответствии выборки 15-ти наиболее употребительным законам распределения по трем, описанным в параграфе 2.1, критериям

согласия.

Параграф 2.3 содержит методику проведения вычислительных экспериментов, цель которых состояла в получении эмпирических оценок функций мощности исследуемых критериев согласия. Выборки, соответствующие альтернативным гипотезам (H_1), генерировались с помощью датчиков псевдослучайных чисел. Далее проверялись нулевые гипотезы (H_0), т.е. согласие распределения каждой из выборок с остальными законами распределения по трем рассматриваемым критериям согласия. На основе полученных результатов для каждой рассмотренной пары гипотез (H_0, H_1) и каждой выборки подсчитывалась доля экспериментов, в которых гипотеза H_0 отвергалась при заданном планом экспериментов значении уровня значимости α_0. Эта доля рассматривалась в качестве эмпирической оценки значения функции мощности критерия согласия для конкретной пары гипотез (H_0, H_1), заданных параметров альтернативного распределения и заданного уровня значимости α_0.

В **параграфе 2.4** с помощью метода ветвей и границ линейного регрессионного анализа построены модели зависимостей мощностей рассмотренных критериев от уровня значимости, объема выборки и значений параметров проверяемых гипотез. Общее число построенных моделей – 149. На основе этих моделей были сформулированы рекомендации пользователям по выбору критериев согласия при использовании статистического программного обеспечения и по определению некоторых условий проведения экспериментов при их планировании.

Например, пусть гипотеза H_0 состоит в том, что распределение выборки соответствует *нормальному* распределению при альтернативной гипотезе H_1 о том, что распределение выборки *равномерно*. План эксперимента содержал 3 стандартных значения уровня значимости α_0 (0.01, 0.05, 0.10; для простоты, в моделях α_0 обозначено как α); 5 значений объема выборки n (10, 50, 100, 500, 1000) и 12 интервалов (a, b), соответствующих плотности равномерного распределения. Были построены следующие модели, оценивающие мощность μ каждого критерия согласия в зависимости от переменных n, a, b, α:

– для критерия *хи-квадрат* ($12 \leqslant n \leqslant 1000$):

$$\mu = -0.07 + 0.0028n + 4\alpha - 0.000002n^2 - 9.8\alpha^2 - 0.0034n\alpha,$$

коэффициент детерминации модели: $R^2 = 0.98$;

– для критерия *Колмогорова-Смирнова* ($87 \leqslant n \leqslant 1000$):

$$\mu = -0.22 + 0.0022n + 4.1\alpha - 0.0000012n^2 - 20\alpha^2,$$

коэффициент детерминации модели: $R^2 = 0.93$;

– для критерия *пустых ящиков* ($27 \leqslant n \leqslant 1000$):

$$\mu = -0.097 + 0.0028n + 2.45\alpha - 0.000002n^2 - 0.002n\alpha,$$

коэффициент детерминации модели: $R^2 = 0.95$.

В круглых скобках указаны дополнительные ограничения на область определения значений, в данном случае, на число элементов выборки n. Эти ограничения были продиктованы особенностями метода регрессионного анализа и конкретных функций распределения. Понятно, что приведенные модели можно использовать только при значениях параметров, указанных для каждой модели. В связи с этим, полученную в главе 2 совокупность моделей можно рассматривать как базовую, которая может пополняться и уточняться.

На основе построенных моделей были разработаны рекомендации пользователям статистического программного обеспечения по выбору критериев согласия в различных условиях. Так, например, в случае проверки на нормальное распределение при альтернативе о равномерности выборки, если $n = 500$ и $\alpha = 0.1$, используя приведенные выше модели получаем: $\mu_{\chi^2} = 0.962$, $\mu_{ks} = 0.79$, $\mu_{eb} = 0.948$, где μ_{χ^2}, μ_{ks}, μ_{eb} – значения функций мощности критериев хи-квадрат, Колмогорова-Смирнова и пустых ящиков соответственно. Нетрудно видеть, что наибольшую мощность в данном случае имеет критерий хи-квадрат, который и следует рекомендовать пользователю. В случае же $n = 1000$ и $\alpha = 0.1$, значения функций мощности будут следующими: $\mu_{\chi^2} = 0.692$, $\mu_{ks} = 0.99$, $\mu_{eb} = 0.748$. В этом случае наибольшую мощность имеет критерий Колмогорова-Смирнова, поэтому в такой ситуации пользователю можно рекомендовать воспользоваться именно этим критерием.

Таким образом, в главе 2 были получены следующие результаты:

1. разработаны программы вычисления значений прямой и обратной функций наиболее употребительных законов распределения и проверки гипотез о соответствии выборки этим законам по трем критериям согласия: χ^2 Пирсона, Колмогорова-Смирнова и пустых ящиков. Все эти программы включены в пакет "Статистик-Консультант";
2. построены модели зависимостей мощностей рассмотренных критериев согласия от уровня значимости, объема выборки и значений параметров проверяемых гипотез;
3. разработаны рекомендации пользователям статистического программного обеспечения по выбору критериев согласия и по определению некоторых условий проведения экспериментов при их планировании.

Третья глава посвящена исследованиям методов поиска наиболее информативного множества признаков (ПНИМП) в линейном регрессионном анализе.

В **параграфе 3.1** дается описание классической задачи регрессионного анализа. Предположим, что случайная величина η имеет некоторое распределение вероятностей при фиксированном значении случайного вектора $\xi = \{\xi_1, \xi_2, \ldots, \xi_m\}$ такое, что $\mathbf{M}(\eta|\xi) = g(\xi, \beta)$, где $\mathbf{M}(\eta|\xi)$ – условное математическое ожидание η при фиксированном ξ, а β – совокупность неизвестных параметров, определяющих функцию $g(\xi, \beta)$. Пусть вектор $\mathbf{Y}' = \{y_1, y_2, \ldots, y_n\}$ содержит результаты n независимых наблюдений величины η, а соответствующие им наблюдения вектора ξ выражены в виде числовой матрицы: $\mathbf{X} = (x_{ij})$, где x_{ij} являются i-ми реализациями величин ξ_j. Требуется по $\mathbf{X}$ и $\mathbf{Y}$ оценить значения параметров β. В практических приложениях решение такой задачи позволяет установить связь между величинами η и ξ в виде математической модели, основанной на упрощенных допущениях: конкретные реализации величин $\xi_1, \ldots, \xi_m$ являются контролируемыми и могут быть заданы, а наблюдаемые значения η представимы в виде

$$y_i = g(x_{i1}, \ldots, x_{im}, \beta) + \varepsilon_i, \qquad i = 1, 2, \ldots, n;$$

где величины ε_i, носящие название ошибок, являются реализациями независимых и одинаково распределенных случайных величин с нулевым математическим ожиданием и постоянной дисперсией (будем считать, что эти случайные величины имеют нормальное распределение). Переменные, являющиеся координатами вектора ξ, принято называть независимыми, признаками, предикторами. Переменную η называют зависимой, откликом.

Одной из основных задач классического регрессионного анализа является выбор модели, то есть вида функции $g(\xi, \beta)$. Наиболее удобной для исследования и оценки и, следовательно, наиболее употребительной, является модель регрессии, линейная относительно параметров $\beta = (\beta_0, \beta_1, \ldots, \beta_k)$, в которой функция $g(\xi, \beta)$ представима в виде линейного уравнения:

$$g(\xi, \beta) = \beta_0 g_0(\xi) + \beta_1 g_1(\xi) + \ldots + \beta_k g_k(\xi), \qquad (2)$$

где $g_0(\xi), \ldots, g_k(\xi)$ – некоторые функции от ξ, не зависящие от β. Вид этих функций обычно выбирается из теоретических соображений или путем подбора. Функции $g_0(\xi), \ldots, g_k(\xi)$ принято называть регрессорами.

Каждая модель вида (2) представляет собой регрессионную зависимость. Поскольку возможности образования регрессоров практически неисчерпаемы, возникает проблема выбора среди различных зависимостей наилучшей в смысле некоторого критерия. Другими словами, требуется из некоторого конечного множества регрессоров $\{g_0(\xi), g_1(\xi), \ldots, g_k(\xi)\}$ выбрать для включения в уравнение подмножество $\{g_{i_1}(\xi), g_{i_2}(\xi), \ldots, g_{i_l}(\xi)\}$, где $l \leqslant k$, обеспечивающее высокое качество модели. Легко видеть, что без ограничения общности мы можем считать, что $g_0(\xi) = 1$, $g_j(\xi) = \xi_j$, $j = 1, \ldots, k$ и называть величины $\xi_1, \ldots, \xi_k$ регрессорами. Таким образом, далее мы будем рассматривать модели, для которых

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i,$$

где $i = 1, 2, \ldots, n$, $k < n - 1$.

Минимальный набор регрессоров, которые можно включить в модель таким образом, чтобы она адекватно (в каком-то определенном

смысле) описывала изучаемое явление, называется наиболее информативным множеством признаков. В регрессионном анализе разработан целый ряд методов поиска такого множества. В проведенном исследовании рассматриваются шесть методов ПНИМП: метод всех возможных регрессий, два вида методов ветвей и границ и три вида пошаговых методов.

Суть каждого из методов ПНИМП состоит в том, что он осуществляет перебор регрессионных моделей, которые можно построить на основе данного начального набора регрессоров, с целью нахождения среди них "наилучшей". Поясним, что в диссертации понимается под "наилучшей" моделью. *Полной моделью* назовем уравнение регрессии, содержащее все регрессоры из начального набора. *Моделью, адекватной полной*, назовем такую ее подмодель, которая не отличается статистически значимо (в смысле некоторого критерия) от полной, но содержит меньше регрессоров. Под "*наилучшей*" будем понимать модель, которая среди всех моделей, адекватных полной, содержит наименьшее число регрессоров. Включение модели в множество моделей, адекватных полной, производится с помощью критериев качества уравнения регрессии. Одним из показателей качества модели является коэффициент детерминации R^2, где R – выборочный коэффициент множественной корреляции.

В системе "Статистик-Консультант" оценка качества регрессионной модели производится по результатам проверки гипотезы о значимости отличия коэффициента детерминации рассматриваемой модели от коэффициента детерминации полной модели.

Параграф 3.2 посвящен проблеме, впервые поднятой А. Н. Колмогоровым – эффекту "вздувания" коэффициента множественной корреляции. Этот эффект заключается в том, что модель существенно (со статистической точки зрения) преувеличивает реально существующую зависимость между исследуемыми переменными, что выражается в слишком большом значении коэффициента множественной корреляции.

Каждый из методов ПНИМП осуществляет проверку гипотезы о том, что полученная им регрессионная модель является адекватной начальным данным. Еще в 1933 г. А. Н. Колмогоров обратил внимание на то, что при осуществлении многократных расчетов коэффициента

множественной корреляции для регрессионных моделей, полученных на основе одного и того же начального набора регрессоров, может происходить так называемое "вздувание" коэффициента множественной корреляции. При этом, хотя в уравнение регрессии обычно не вводится более 5 – 7 переменных, запас переменных, из которых они могут быть выбраны, может быть очень велик.

Построенные таким образом модели могут хорошо описывать имеющиеся данные, но не быть пригодными для прогнозирования. Предложение А. Н. Колмогорова по уменьшению данного эффекта вздувания коэффициента множественной корреляции состоит в том, чтобы ограничить число переменных, входящих в начальный набор регрессоров. Но остается неясным вопрос о конкретном числе переменных, которые можно включить в начальный набор регрессоров, для каждого метода ПНИМП, позволяющем найти модель так, чтобы вероятность вздувания коэффициента детерминации была бы допустимой.

В **параграфе 3.3** подробно рассмотрена методика исследования описанных в параграфе 3.1 методов ПНИМП с точки зрения возникновения эффекта вздувания коэффициента детерминации.

Идея исследования заключается в генерировании с помощью датчиков псевдослучайных чисел, описанных в главе 1, данных по заранее заданным моделям, включая случай отсутствия зависимости отклика от регрессоров, и сравнении построенных с помощью методов ПНИМП моделей с заданными. Если регрессионный метод находит модель, близкую к заданной, то результат такого эксперимента считался успехом, в противном случае – "неудачей". По результатам исследования были построены модели, отражающие взаимозависимость числа "неудач" метода, числа регрессоров, включаемых в начальный набор и величины параметра метода (для метода всех возможных регрессий и методов ветвей и границ – это уровень значимости регрессионной модели (в случае отсутствия зависимости отклика от регрессоров) или уровень значимости для проверки гипотезы о включении конкретного регрессора в модель (в случае существования зависимости); для пошаговых методов – это F-статистика включения/исключения). Кроме того, проводилось сравнение методов ПНИМП между собой с точки зрения их отличия друг от друга по вероятности возникновения эффекта "вздувания" коэффициента

детерминации в зависимости от числа регрессоров, включаемых в начальный набор и значений параметров методов.

Параграф 3.4 посвящен результатам вычислительных экспериментов по определению для каждого метода ПНИМП максимального числа регрессоров, при котором вероятность неудачи (см. выше) была бы допустимой.

Так, например, для метода всех возможных регрессий в разных условиях формирования отклика и предикторов были получены следующие модели зависимости параметра метода p от числа регрессоров начального набора r и числа "неудач" метода f (понятия параметра метода и "неудачи" введены выше):

- в случае отсутствия зависимости отклика от регрессоров, когда распределение отклика и регрессоров было нормальным:
$$p = 4.6828 - 0.2541r + 0.0937f;$$
- в случае отсутствия зависимости отклика от регрессоров, когда распределение отклика и регрессоров неизвестно:
$$p = 0.1732 - 0.005r + 0.0027f;$$
- в случае существования зависимости отклика от одного регрессора:
$$p = 2.5898 - 0.1485r + 0.0638f;$$
- в случае существования зависимости отклика от набора из шести регрессоров:
$$p = 3.1126 - 0.1622r + 0.0701f.$$

Результаты попарного сравнения методов ПНИМП и построенные модели были использованы для разработки рекомендаций пользователям статистического программного обеспечения при работе с методами ПНИМП. Приведем пример формирования рекомендации по выбору параметра метода, основанной на приведенных выше моделях.

Пусть у пользователя имеется выборка наблюдений, соответствующих некоторой зависимой переменной, 15 независимых переменных, а заданная им допустимая вероятность возникновения эффекта вздувания коэффициента детерминации равна 5%. Необходимо найти регрессионную модель зависимости отклика от каких-то регрессоров из имеющегося начального набора.

Основываясь на полученных результатах сравнения методов, система рекомендует воспользоваться методом всех возможных регрессий. Далее, для формирования рекомендации по выбору параметра метода, используя приведенные выше модели, получаем следующее:

- если отклик не зависит от набора регрессоров, и если распределение отклика и регрессоров нормально, то $p = 1.34\%$,
- если отклик не зависит от набора регрессоров, и если распределение отклика и регрессоров неизвестно, то $p = 0.11\%$.

Очевидно, что для того, чтобы обеспечить решение поставленной задачи, из приведенных значений p нужно выбрать наименьшее. Далее:

- если отклик зависит от одного регрессора, то $p = 0.68\%$,
- если отклик зависит от шести регрессоров, то $p = 1.03\%$.

Таким образом, в данном примере можно рекомендовать воспользоваться методом всех возможных регрессий, причем для того, чтобы вероятность возникновения эффекта вздувания коэффициента детерминации была бы не более 5%, можно рекомендовать значение уровня значимости регрессионной модели 0.11%, а значение уровня значимости отдельного регрессора – 0.68%. Это значит, что модель, содержащую не менее одного регрессора, следует считать значимой, если гипотеза о равенстве нулю коэффициента детерминации модели отвергнута с уровнем значимости 0.11%, а гипотезы о равенстве нулю каждого коэффициента модели (кроме свободного члена) отвергнуты с уровнем значимости 0.68%.

До сих пор считалось, что основным лимитирующим фактором при выборе метода ПНИМП является число включаемых в модель регрессоров, поскольку методы, осуществляющие просмотр и анализ значительного числа регрессионных уравнений, требуют очень большого времени счета, зависящего от числа регрессоров. В ходе проведенных исследований оказалось, что эффект Колмогорова появляется значительно раньше, чем исчерпываются возможности вычислительной техники. Поэтому главным лимитирующим фактором должен считаться этот эффект и в большинстве случаев достаточно использовать метод всех возможных регрессий. Разумеется, это не значит, что следует отказаться от других методов. При необходимости рассмотрения большого числа регрессоров их использование должно проходить в

значительно более жестких условиях, чем обычно принято, в первую очередь это относится к методам оценки значимости коэффициентов модели.

Проведенное в главе 3 исследование позволило получить следующие результаты:

1. показано, что главным лимитирующим фактором при выборе метода ПНИМП в регрессионном анализе должно быть не время вычислений, как это считалось ранее, а вероятность возникновения эффекта "вздувания" коэффициента детерминации;
2. выявлены отличия методов ПНИМП друг от друга по вероятности возникновения эффекта "вздувания" коэффициента детерминации в зависимости от числа регрессоров, включаемых в начальный набор и выбранных параметров методов;
3. построены модели взаимной зависимости числа регрессоров, включаемых в начальный набор регрессоров, параметров методов и числа случаев ошибочной работы методов (общее число моделей – 72);
4. сформулированы рекомендации по выбору методов ПНИМП и их параметров, направленные на снижение вероятности возникновения эффекта "вздувания" коэффициента детерминации. В частности, показано, что при применении пошаговых методов значение F-статистики включения/исключения следует брать не менее 8, а не 4, как это часто рекомендуется в статистической литературе;

В заключении приводятся основные результаты диссертации и указываются возможности их применения и развития.

СПИСОК ОПУБЛИКОВАННЫХ РАБОТ ПО ТЕМЕ ДИССЕРТАЦИИ

1. Свидетельство об официальной регистрации программы для ЭВМ № 950298. – М.: РосАПО, 1995.

2. Pavlov Y. L., Leri M. M., Spector E. N., Stafeev S. V., Heninen A. J. Some problems of intellectualization of statistical packages //

Proceedings of the International Conference "Computer data analysis and modeling", v. 1, Minsk, 1995. – P. 116–120.

3. Лери М. М. Эмпирическая оценка мощности критериев согласия для базы знаний экспертной системы // Труды Петр. ГУ, серия "Прикладная математика и информатика", вып. 6, 1997. – С. 187–192.

4. Лери М. М. О выборе "наилучшей" регрессии // Труды Института ПМИ, вып. 1. – Петрозаводск: КарНЦ РАН, 1999. – С. 21–28.

5. Leri M. M. On one problem of A. N. Kolmogorov // Proceedings of the Fifth International Petrozavodsk conference "Probabilistic methods in discrete mathematics", VSP, Utrecht, 2001. – P. 219–225.

6. Leri M. M., Pavlov Y. L. On methods of searching for regression patterns // Proceedings of the Sixth International Conference "Computer Data Analysis and Modeling: Robustness and Computer Intensive Methods", v.2, Minsk, 2001. – P. 49–54.

Тезисы докладов

7. Лери М. М., Павлов Ю. Л. Разработка экспертных правил выбора критерия согласия // Тез. докл. VI науч. конф. стран СНГ "Применение многомерного статистического анализа в экономике и оценке качества продукции". – Москва, 1997. – С. 130–131.

8. Лери М. М. Об одной задаче А. Н. Колмогорова // Обозрение прикладной и промышленной математики, т. 7, вып. 1, 2000. – С. 190–192.

9. Лери М. М. Об использовании методов регрессионного анализа // Тез. докл. Всероссийской научной школы "Математические методы в экологии". – Петрозаводск: КарНЦ РАН, 2001. – С. 314–316.

10. Leri M. M. On some approaches to the development of intellectualized statistical software // Extended abstracts of Russian-Scandinavian Symposium "Probability Theory and Applied Probability". – Petrozavodsk: KarRC RAS, 2006. – P. 35–37.

Изд. лиц. № 00041 от 30.08.99 г. Сдано в печать 17.08.06.
Формат 60x84 1/16. Гарнитура Times. Уч.-изд. л. 1,0. Усл. печ. л. 1,2.
Тираж 100 экз. Изд. № 52. Заказ № 596.

Карельский научный центр РАН
Редакционно-издательский отдел
Петрозаводск, пр. А. Невского, 50