УДК 519.246.8

С.Г. РАДЧЕНКО, С.Н. ЛАПАЧ

ОСНОВНЫЕ КОНЦЕПЦИИ МНОЖЕСТВЕННОГО РЕГРЕССИОННОГО АНАЛИЗА

Анотація. *Наведено типові умови отримання статистичних регресійних моделей. Розроблено методи отримання моделей з можливо найкращими статистичними властивостями. Приведено конкретні системні вирішення основних задач регресійного аналізу та напрями подальших досліджень.*
Ключові слова: *регресійний аналіз, планування експерименту, стійке оцінювання моделей,отримання квазіоптимальних планів експериментів.*

Аннотация. *Приведены типичные условия получения статистических регрессионных моделей. Разработаны методы получения моделей с возможно наилучшими статистическими свойствами. Приведены конкретные системные решения основных задач регрессионного анализа и направления дальнейших исследований.*
Ключевые слова: *регрессионный анализ, планирование эксперимента, устойчивое оценивание моделей, получение квазиоптимальных планов экспериментов.*

Abstract. *The typical conditions for statistic regressive models formation were described. The methods for models with the possibly optimal statistical properties were developed. The specific system solutions for the key problems of regression analysis and directions for future research were described.*
Keywords: *regression analysis, design of experiment, stable estimation of models, quasioptimal experimental designs generation.*

1. Вступление. Постановка проблемы

Многофакторные статистические регрессионные модели полиномиального вида, линейные по параметрам и, в общем случае, не линейные по факторам, широко используются в технических, технологических, агробиологических и других исследованиях систем. При построении моделей необходимо восстановить в виде формализованного выражения влияние управляемых факторов и оценить случайную составляющую, которая не несет полезной информации. Источником ее являются неуправляемые и неконтролируемые факторы. Такие задачи получили название обратных. Обратная задача – определение коэффициентов $\boldsymbol{B}$ в уравнении $\boldsymbol{Y=XB+E}$ по измеренному выходному результату $\boldsymbol{Y}$ и условиям наблюдения $\boldsymbol{X}$; $\boldsymbol{E}$ – значение случайной ошибки ε. Многие обратные задачи являются некорректно поставленными задачами.

2. Анализ условий получения регрессионных моделей

Статистические модели получают в следующих типичных условиях.

1. Число возможных опытов ограничено и практически не превышает 32…50(64), что приводит к использованию дробного факторного эксперимента. Полный факторный эксперимент обычно возможен только для 3…4 факторов, если число уровней не превышает трех.

2. Число факторов и число их уровней могут быть такими, что практически трудно найти экономный (по числу опытов) план эксперимента.

3. Для многофакторных регулярных планов экспериментов не известны последовательные планы.

4. Структуры определяемых статистических моделей почти всегда исследователю не известны.

5. Форма факторного пространства при решении реальных прикладных задач может быть произвольной, т.е. не соответствовать многомерному прямоугольному параллелепипеду, сфере, симплексу. В этом случае факторы будут коррелированы друг с другом и необходимо решать некорректно поставленную задачу.

6. Степень влияния неуправляемых и неконтролируемых факторов может весьма изменяться в повторных сериях опытов, и тогда влияние управляемых факторов становится статистически незначимым.

3. Цель публикации

Разработать методы получения регрессионных моделей, которые в вышеприведенных условиях обеспечивают наилучшие возможные их критерии качества.

4. Изложение разработанных методов

Получение многофакторных статистических моделей с наилучшими свойствами возможно только при планировании эксперимента, т.е. в том случае, когда матрица, которая используется для построения модели, конструируется таким образом, чтобы обеспечить требуемые свойства всего процесса моделирования. То есть оптимальность плана должна определяться как ошибкой модели, так и ошибкой определения коэффициентов модели [1].

Это означает, что к построению регрессионных моделей необходимо подходить системно: процесс должен включать построение плана эксперимента, формализованный выбор структуры модели, устойчивое оценивание коэффициентов модели.

Под устойчивым (робастным) планом эксперимента понимается план полного или дробного факторного эксперимента, позволяющий выбрать неизвестные исследователю структуры «истинных» статистических моделей $\hat{y}_w$ полиномиального вида, линейных по параметрам, и получить адекватные модели (w – текущий номер определяемой модели, $1 \le w \le m$, m – общее число определяемых моделей по устойчивому плану эксперимента). План эксперимента не изменяется для получаемых различных структур моделей [2].

Устойчивым робастным планам экспериментов соответствуют полные факторные эксперименты, многофакторные регулярные, не близкие к насыщенным, планы экспериментов, планы на основе $ЛП_\tau$ равномерно распределенных последовательностей [2].

Устойчивая структура многофакторной статистической модели – структура, которая характеризуется неизменностью множества главных эффектов и взаимодействий многофакторной статистической модели полиномиального вида при изменении значений результатов экспериментов (откликов), порождаемых случайными ошибками (погрешностями) результатов наблюдений, измерений, вычислений и неопределенностью искомой структуры модели. Структурные элементы моделей выбираются из множества структурных элементов модели полного факторного эксперимента с ортогональными или слабо коррелированными (коэффициент парной корреляции $\left|r_{ij}\right| < 0,3$) эффектами с использованием устойчивого (робастного) плана эксперимента [2].

Под устойчивостью коэффициентов статистической модели будем понимать минимально возможную изменчивость коэффициентов многофакторной статистической модели полиномиального вида к случайным ошибкам (погрешностям) результатов наблюдений, измерений и вычислений. Для оценки устойчивости коэффициентов используется число обусловленности $cond\,(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})$. Устойчивость наилучшая, если $cond\,(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}) = 1$, хорошая $1 < cond\,(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}) \le 10$, удовлетворительная $10 < cond\,(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}) \le 100$, неудовлетворительная $cond\,(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}) > 100$ [2].

Анализ условий решения прикладных реальных задач, свойств полного факторного эксперимента и многофакторных регулярных планов позволил сформулировать основные требования к структуре статистической модели.

1. Структурная группа коэффициентов многофакторного уравнения регрессии не известна исследователю.

2. Структуры моделей выбираются из структуры модели полного факторного эксперимента с ортогональными или близкими к ортогональным структурными эффектами.

3. Выбор структуры модели должен быть формализованным.

4. Возможность формализованного отображения в математической модели произвольной (но конечной) по сложности реальной действительности при условии правильного выбора степени полинома по каждому фактору.

5. Доступность, простота и надежность фактического получения адекватной структуры при решении задач на потоке.

При переходе от натуральных значений факторов $X_1,...,X_k$ к системе ортогональных полиномов Чебышева (системе ортогональных контрастов) структура математической модели имеет вид

$$(1+x_1^{(1)}+x_1^{(2)}+...+x_1^{(s_1-1)})\times...\times(1+x_k^{(1)}+x_k^{(2)}+...+x_k^{(s_k-1)})\to N_{\Pi},$$

где 1 – значение фиктивного фактора $x_0\equiv 1$;

$x_1^{(1)}$, …, $x_1^{(s_1-1)}$, …, $x_k^{(1)}$, …, $x_k^{(s_k-1)}$ – ортогональные контрасты факторов $X_1,...,X_k$;

$s_1,...,s_k$ – число различных уровней факторов $X_1,...,X_k$;

k – общее число факторов;

(1), (2), …, $(s_1-1),...,(s_k-1)$ – порядок контрастов факторов $X_1,...,X_k$;

N_{Π} – число структурных элементов полного факторного эксперимента, равное числу опытов эксперимента.

Предполагается, что максимальное значение порядка ортогонального контраста $s_i-1\left(1\le i\le k\right)$ достаточно для адекватного описания влияния фактора X_i по всей области факторного пространства. Значение s_i назначается исследователем, исходя из логически профессионального анализа предметной области.

Для полного факторного эксперимента число структурных эффектов (элементов) модели равно числу опытов плана эксперимента N_{Π}, и все эффекты ортогональны друг к другу. Получаемая статистическая модель будет адекватна результатам эксперимента, так как множество структурных элементов необходимо и достаточно для описания результатов опытов.

В случае выбора дробного факторного регулярного плана эксперимента все главные эффекты будут ортогональны друг к другу. Из структуры модели полного факторного эксперимента возможно выделение различных структур статистических моделей $\hat{y}_w$ для дробного факторного эксперимента. Если план эксперимента не выбирать близким к насыщенному, то некоторые взаимодействия будут ортогональны к главным эффектам, введенным в модель, и модель будет адекватна либо близка к адекватной.

Рассмотрим построение структуры модели для полного факторного эксперимента $2^1\times3^1\times4^1//24$: первый фактор X_1 на двух, второй X_2 на трех, третий X_3 на четырех уровнях. Формализованная структура статистической модели будет следующей:

$$(1+x_1^{(1)})(1+x_2^{(1)}+x_2^{(2)})(1+x_3^{(1)}+x_3^{(2)}+x_3^{(3)})\to N_{\Pi}=24,$$

где $x_1^{(1)}$, $x_2^{(1)}$, $x_3^{(1)}$ – линейные контрасты факторов X_1, X_2, X_3;

$x_2^{(2)}$, $x_3^{(2)}$ – квадратичные контрасты факторов X_2, X_3;

$x_3^{(3)}$ – кубический контраст фактора X_3;

$N_{П} = 24$ – число структурных элементов статистической модели, равное числу опытов плана экспериментов.

Общий вид статистической модели будет следующий:

$$\hat{y} = b_0 x_0 + b_1 x_1^{(1)} + b_2 x_2^{(1)} + b_3 x_2^{(2)} + b_4 x_3^{(1)} + b_5 x_3^{(2)} + b_6 x_3^{(3)} + b_7 x_1^{(1)} x_2^{(1)} + b_8 x_1^{(1)} x_2^{(2)} +$$

$$+ b_9 x_1^{(1)} x_3^{(1)} + b_{10} x_1^{(1)} x_3^{(2)} + b_{11} x_1^{(1)} x_3^{(3)} + b_{12} x_2^{(1)} x_3^{(1)} + b_{13} x_2^{(1)} x_3^{(2)} + b_{14} x_2^{(1)} x_3^{(3)} +$$

$$+ b_{15} x_2^{(2)} x_3^{(1)} + b_{16} x_2^{(2)} x_3^{(2)} + b_{17} x_2^{(2)} x_3^{(3)} + b_{18} x_1^{(1)} x_2^{(1)} x_3^{(1)} + b_{19} x_1^{(1)} x_2^{(2)} x_3^{(1)} +$$

$$+ b_{20} x_1^{(1)} x_2^{(1)} x_3^{(3)} + b_{21} x_1^{(1)} x_2^{(2)} x_3^{(1)} + b_{22} x_1^{(1)} x_2^{(2)} x_3^{(2)} + b_{23} x_1^{(1)} x_2^{(2)} x_3^{(3)}.$$

Модель содержит семь главных эффектов, одиннадцать двойных взаимодействий и шесть тройных взаимодействий.

Для обеспечения решения задач прикладного множественного регрессионного анализа в рамках системного подхода разработаны следующие концепции.

Как развитие теории многофакторных регулярных планов:

1) квазиортогональные, квази-*D*-оптимальные, квазирегулярные и квазиравномерные планы экспериментов, для получения которых разработаны алгоритмы RASTA1, RASTA2, RASTA8 [2];

2) генерирование последовательных регулярных планов экспериментов [2].

2. Область факторного пространства в технических и технологических системах часто не соответствует стандартной – многофакторному прямоугольному параллелепипеду. Для преобразования области факторного пространства к стандартной разработан топологический метод устойчивого оценивания регрессионных моделей. Он заключается в установлении взаимно однозначного и взаимно непрерывного соответствия между прообразом факторного пространства, в котором эффекты ортогональны друг к другу или близки к ортогональным и в котором можно оптимально планировать эксперимент и получать статистические модели с наилучшими возможными свойствами, и образом факторного пространства, который задается в предметной области и в котором планирование эксперимента традиционными методами невозможно из-за мультиколлинеарности факторов [3]. Топологический метод устойчивого оценивания регрессионных моделей привел к созданию инвариантно-группового подхода в теории планирования эксперимента. Он имеет следующие модификации метода:

1) Получение функций отображения прообраза факторного пространства в образ [3].

2) Установление собственной кодированной системы координат в области прообраза и области образа [3].

3) Планирование эксперимента с использованием фиктивных факторов [3].

3. Как развитие робастных планов эксперимента предложено использовать регулярные планы и планы на основе $ЛП_\tau$ равномерно распределенных последовательностей [2, 4, 5]. Эти планы наилучшим образом отвечают системным требованиям к процессу построения регрессионной модели. Использование данных планов обеспечивает одновременно оптимальные условия для поиска неизвестной структуры уравнения регрессии и достаточно близкие к оптимальным условия получения устойчивых оценок коэффициентов регрессии. Кроме того, эти планы дополнительно устойчивы к отклонениям от самого плана: пропус-

ки отдельных экспериментов и незначительные отклонения от значения уровней плана. Это свойство, а также возможность использовать планы как последовательные, представляют значительные удобства (и экономический выигрыш) для экспериментатора.

4. Распространением теории эксперимента на ситуации, в которых экспериментатор не может проводить эксперимент по заранее построенному плану, является разработка методов построения из матрицы пассивного эксперимента матрицы, обладающей необходимыми свойствами для получения устойчивой и информативной регрессионной модели [6].

5. Показано, что для обеспечения устойчивости процесса получения оценок коэффициентов регрессии необходимо ортогональное представление эффектов (главных и взаимодействий) в виде ортогональных нормированных контрастов [2, 4].

6. Для формализованного получения устойчивых структур моделей, заранее не известных исследователю, из структурного множества эффектов полного факторного эксперимента разработаны соответствующие алгоритмы (RASTA3 [2]) и программное средство «Планирование, регрессия и анализ моделей» (ПС ПРИАМ) [7].

В алгоритме RASTA3 проводится последовательная проверка статистической значимости главных эффектов и взаимодействий для введения их в получаемую модель. Предполагается использование устойчивого (робастного) плана эксперимента. Условия ввода эффектов: ортогональность или малая коррелированность (коэффициент парной корреляции $\left|r_{ij}\right| \le 0,3...0,4$ с выбранными для введения в модель эффектами); выбор вводимых эффектов проводится из числа эффектов структуры модели полного факторного эксперимента.

ПС ПРИАМ характеризуется следующими возможностями:

1) Реализация специально разработанной технологии решения научных и прикладных задач по построению математических моделей и многокритериальной оптимизации, а не набор стандартных методов и средств прикладной статистики.

2) Ориентация на массового пользователя: ПС содержит все необходимые средства для решения задач от ее постановки до подготовки отчета, обеспечивается получение результатов высокого качества за счет самоадаптирующихся вычислительных схем, настраивающихся на исходные данные и промежуточные результаты. Имеется возможность изменять параметры вычислительных схем и активно вмешиваться в процесс решения задачи на любом этапе:

• достигается высокая надежность и достоверность конечного результата;

• контекстная помощь позволяет в любом месте получить необходимую информацию.

3) Робастное (устойчивое) конструирование эксперимента.

4) Эффективные алгоритмы определения структуры уравнения регрессии.

5) Устойчивое оценивание сильно коррелированных факторов в многофакторном уравнении регрессии.

7. Поправка RASTA для оценивания и исключения в информационном смысле систематических погрешностей от влияния неуправляемых и неконтролируемых факторов в различных сериях повторных опытов. Это позволяет обоснованно определить значимость влияния управляемых факторов и повысить точность получаемых результатов [2].

В исследовании урожайности кормовых бобов среднеквадратическая ошибка результатов экспериментов была уменьшена в 6,8 раза, что позволило обоснованно установить статистическую значимость влияния эффектов всех управляемых факторов на урожайность бобов [8].

Повышение воспроизводимости результатов экспериментальных исследований было использовано также в исследованиях технологического процесса нарезания наружной резьбы винторезными самооткрывающимися головками ЗКА по критерию точности [9]. Дисперсия воспроизводимости критерия качества была уменьшена в 2,6 раза.

5. Выводы и перспективы дальнейших исследований

Практика использования разработанных концепций для решения более ста прикладных задач показала их эффективность и соответствие реальности условий. Отметим, что необходимо использовать системный подход, включающий все приведенные разработки. Это позволяет рассматривать весь процесс построения регрессионных моделей как единую технологию и принимать решения на каждом этапе, исходя из требований получения модели необходимого качества, а не из возможностей применения каких-либо методов.

С разработанными методами решения регрессионных задач и полученными результатами можно ознакомиться в [10, 11].

Дальнейшее возможное развитие методологии регрессионного анализа целесообразно проводить в следующих направлениях:

1. Системный подход в получении многофакторных статистических моделей, включающий 1) устойчивый (робастный) план эксперимента, 2) устойчивую структуру модели, априори не известную исследователю, 3) устойчивое оценивание коэффициентов модели.

2. Распространение концепции ортогональности во множественном регрессионном анализе на нестандартные области факторного пространства.

3. Дальнейшее развитие инвариантно-группового подхода в теории планирования эксперимента.

4. Исследование статистических свойств планов экспериментов на основе использования $ЛП_\tau$ равномерно распределенных последовательностей.

5. Дальнейшее развитие генерирования квазиортогональных квазирегулярных планов экспериментов.

6. Разработка методов выделения из массивов исходных непланированных данных информативного подмножества с наилучшими возможными статистическими свойствами.

СПИСОК ЛИТЕРАТУРЫ

1. Математическая теория планирования эксперимента / Под ред. С.М. Ермакова. – М.: Наука, ГРФМЛ, 1983. – 392 с.
2. Радченко С.Г. Методология регрессионного анализа / Радченко С.Г. – К.: Корнійчук, 2011. – 376 с.
3. Радченко С.Г. Устойчивые методы оценивания статистических моделей / Радченко С.Г. – К.: ПП «Санспарель», 2005. – 504 с.
4. Лапач С.Н. Статистические методы в фармакологии и маркетинге фармацевтического рынка / Лапач С.Н., Пасечник М.Ф., Чубенко А.В. – К.: ЗАТ «Укрспецмонтажпроект», 1999. – 312 с.
5. Лапач С.Н.Статистические методы в медико-биологических исследованиях с использованием Excel / Лапач С.Н., Чубенко А.В., Бабич П.Н. – [2-е изд. перераб. и доп.]. – К.: Морион, 2001. – 408 с.
6. Лапач С.М. Забезпечення необхідних властивостей вибірки для побудови регресійної моделі / С.М. Лапач // Физические и компьютерные технологии. Труды 15-й Междунар. научно-техн. конф., (Харьков, 2–3 декабря 2009 г.). – Харьков: ХНПК «ФЭД», 2009. – С. 179 – 182.
7. Лапач С.Н. Планирование, регрессия и анализ моделей PRIAM (ПРИАМ) / С.Н. Лапач, С.Г. Радченко, П.Н. Бабич // Программные продукты Украины: каталог. – К., 1993. – С. 24 – 27.
8. Статистичні методи планування експериментів та обробки їхніх результатів у рослинництві / В.Ф. Петриченко, С.Г. Радченко, П.М. Бабіч [та ін.] // Вісник аграрної науки. – 2006. – № 11. – С. 25 – 29.
9. Радченко С.Г. Оптимизация технологических условий нарезания наружных резьб винторезными самооткрывающимися головками по критерию точности / С.Г. Радченко, С.С. Добрянский // Вестник машиностроения. – 1986. – № 1. – С. 56 – 59.
10. Лаборатория экспериментально-статистических методов исследований (ЛЭСМИ) [Электронный ресурс]. – Режим доступа: http://www.n-t.org/sp/lesmi.

11. Сайт кафедры «Технология машиностроения» Механико-машиностроительного института Национального технического университета Украины «Киевский политехнический институт» [Электронный ресурс]. – Режим доступа: http://tm-mmi.kpi.ua/index.php/ru/1/ publications/352?task=view.

Стаття надійшла до редакції 14.09.2012