ИМИТАЦИОННОЕ МОДЕЛИРОВАНИЕ ПРИ КОНСТРУИРОВАНИИ РЕШЕНИЙ И ОБУЧЕНИИ

Г.А. Реймаров (Обнинск)

Введение

Моделирование и анализ свойств множества реальных объектов не укладываются в прокрустово ложе типовых программ. Реальные задачи намного богаче примеров, кочующих из одной библиотеки программ в другую (например, расчет дискриминантной функции для разных сортов ирисов). В связи с этим, процесс моделирования сложных объектов выражается в конструировании нестандартных, самостоятельных решений и имитационных исследований свойств этих решений. В докладе обобщается опыт решения практических задач и обучения с использованием языка APL.

Инструментарий

Чтобы не увязнуть в мелочах программной реализации нестандартных алгоритмов, целесообразно использовать языки сверхвысокого уровня. Автор предпочитает «компьютерное каратэ» — язык APL. Мощность этого языка видна на примере расчета всех простых чисел в диапазоне от 1 до R:

R←100 (задаем R=100). Программа и результат:

```
      (~R∈R∘.×R)/R←1↓⍳R
2 3 5 7 11 13 17 19 23 29 31 37 41 43 47 53 59 61 67 71 73 79
83 89 97
```

Реализация бутстрепа:

```
  XB←X[?NρN←ρX]
```

Множество матричных операций (как обычных, вроде обращения и транспонирования матриц, так и оригинальных) являются командами этого языка.

В связи с этим, APL является удобным средством программирования и имитационных испытаний оригинальных алгоритмов анализа данных, а также эффективной средой обучения [1-6]. Отметим некоторые программные продукты нашей разработки:

1. Комплекс программ селекции тестов и решающих правил профотбора СТРЕП. Возник в результате расширенного анализа данных психофизиологических обследований персонала атомных электростанций и экипажей атомных подводных лодок. Позволяет формировать эффективные батареи тестов и рассчитывать решающие правила профотбора. Включает оригинальные программы регрессионного и дискриминантного анализа, рассчитанные на обработку многомерных массивов данных с большим числом регрессоров (дискриминаторов). Основная особенность этих программ заключается в том, что пользователь может управлять последовательностью исключения статистически незначимых регрессоров. В итоге вместо одного «оптимального» решения (например, посредством расчета по программе STATISTIKA) может быть выявлено несколько уравнений, имеющих равные права на существование. Пользователь отбирает наилучшее решение исходя из физических соображений. В программе дискриминантного анализа благодаря применению бутстрепа удается снизить примерно в 3 раза ошибки

классификации, связанные с неточностью оценки порога решения.

2. Система комплексной оценки руководителей и специалистов «Персона» (первая версия). Система «Персона» нашла применение более чем на 60 крупных предприятиях России. Уникальна, поскольку дает возможность оценивать не потенциал работника, а качество его деятельности [4]. Отмечена отечественными специалистами как программный продукт, в котором эффективно используются высокие статистические технологии.

3. Библиотека инструментальных средств анализа данных на языке APL2 (БИС «Аналитик»). В текущий момент времени содержит более 100 программ [5]. Основная часть программ реализует ключевые процедуры обработки данных и может использоваться для «блочного» конструирования специализированных программ, способных обеспечить решение широкого спектра нестандартных задач. Разрабатывая комплекс инструментальных средств статистического анализа для целей обучения, авторы исходили из того, что реальному сектору экономики необходимы специалисты, способные быстро адаптироваться к конкретным условиям и не только сопровождать производственные процессы, но и продуктивно участвовать в совершенствовании систем управления, внедрении инноваций.

4. Комплекс программ RAKO [7]. Изначально был предназначен для прогнозирования кадрового состава предприятий агропромышленного комплекса России. Совершенствование алгоритмов и технологии прогнозирования привело к существенному расширению сферы возможного применения этого комплекса.

Недостатки первых реализаций APL (неудобства ввода команд и отсутствие приличной графики), преодолены в версии APL2 [8]. Язык APL2 оказался намного сложнее и в то же время мощнее, чем APL360. Это стало одной из причин снижение популярности APL и в то же время источником вдохновения его поклонников.

Имитационные исследования алгоритмов регрессионного анализа

Испытания алгоритмов вначале ограничивались решением типовых примеров. Так, среди специалистов по статистическому анализу широкую популярность приобрели примеры Хоэрла [8].

Единственный опыт — расчет типового примера не может служить основанием для выводов о работоспособности или сравнительной эффективности исследуемого алгоритма. Нередко на основании типового примера распространяются ошибочные процедуры, пригодные только для условий примера, либо случайно давших верное решение [1,2]. Нормой должна быть обязательная проверка работоспособности алгоритма в широкой области влияющих факторов, а не в единственной точке пространства условий.

В качестве основной характеристики качества решений регрессионных задач разумно использовать относительную ошибку

$$\delta = \|\hat{\beta} - \beta\|/\|\beta\|, \quad (1)$$

где $\|\cdot\|$ — знак евклидовой нормы;

$\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_m)^T$ и $\beta = (\beta_1, \beta_2, \dots, \beta_m)^T$ — соответственно оценка и истинное значение регрессионных коэффициентов.

Учтем, что вместо уравнения МНК-оценки

$$R\beta = T, \quad (2)$$

где $R = X^T X;\ T = X^T Y$,
фактически решается система

$$R_1\hat{\beta} = T_1, \quad (3)$$

где R_1=$R + \varphi;\ T_1 = T + \tau;\ \varphi, \tau$ — ошибки (статистические отклонения).
После вычитания (2) из (3) получим:

$$R\left(\hat{\beta} - \beta\right) = \tau - \varphi\hat{\beta}. \quad (4)$$

Включение в правую (объясняющую) часть последней формулы оценки $\hat{\beta}$ нежелательно. Избавляемся от нее следующим образом. Обозначим $\Delta = \hat{\beta} - \beta$, соответственно $\hat{\beta} = \Delta + \beta$, и вместо выражения (4) можно записать:

$$(R + \varphi)\Delta = \tau - \varphi\beta$$

или

$$\Delta = (R + \varphi)^{-1}(\tau - \varphi\beta),$$

и вместо (1) получаем:

$$\delta = \|(R + \varphi)^{-1}(\tau - \varphi\beta)\|/\|\beta\|. \quad (5)$$

Формула (5) объясняет механизм взаимной компенсации отклонений, а также стабилизирующее (регуляризующее) действие малых ошибок измерения регрессоров (φ) при сильной мультиколлинеарности.

К сожалению, ошибку оценки регрессионных коэффициентов δ невозможно оценить в практических расчетах; в имитационных исследованиях она незаменима. Вынужденной мерой адекватности уравнений является *F*-статистика. Сравним, как ведут себя обе эти меры, δ и *F,* с усилением коррелированности регрессоров.

С помощью средств Библиотеки «Аналитик» имитируем матрицу регрессоров *X* и вектор отклика *Y*. При этом:

- регрессоры имеют стандартное нормальное распределение;
- корреляция всех пар регрессоров одинакова, нормированная корреляционная матрица регрессоров имеет вид:

$$\rho = \begin{bmatrix} 1 & r & r & r \\ r & 1 & r & r \\ r & r & 1 & r \\ r & r & r & 1 \end{bmatrix}.$$

- В качестве примера взята модифицированная модель Хоэрла:

$$Y = 10 + 2x_1 + 3x_2 + 5x_3 + 0x_4 + S, \quad (6)$$

где *S* — случайная ошибка, распределенная по нормальному закону,

$$f(s) = N(0, \sigma).$$

Задавая разные условия имитации опытных данных и получая соответствующие регрессионные уравнения, несложно оценить влияние параметров на точность расчетов. Например:

- Как меняется ошибка оценивания β с изменением числа опытов ($10<n<100$);
- Как влияет на β ошибка измерений ($0.2<\sigma<10$);
- Как меняется структура вектора регрессионных коэффициентов $\hat{\beta}$ при усилении коррелированности регрессоров ($0.5< r \leq 0.99$).

С увеличением *r* растет модуль вектора $\hat{\beta}$, регрессоры обмениваются долей своего влияния на отклик. При этом коэффициенты $\beta_1, \beta_2, \beta_3$ могут быть незначимыми или отрицательными; значимым нередко оказывается «посторонний» регрессор x_4.

На следующих двух рисунках показаны графики, полученные для 10 значений *r*: 0.09 0.19 …0.99. Для каждого из фиксированных значений *r* генерировалось 1000

выборок *X,Y*; из результатов расчетов 1000 уравнений регрессии находились медианы *F* и δ.

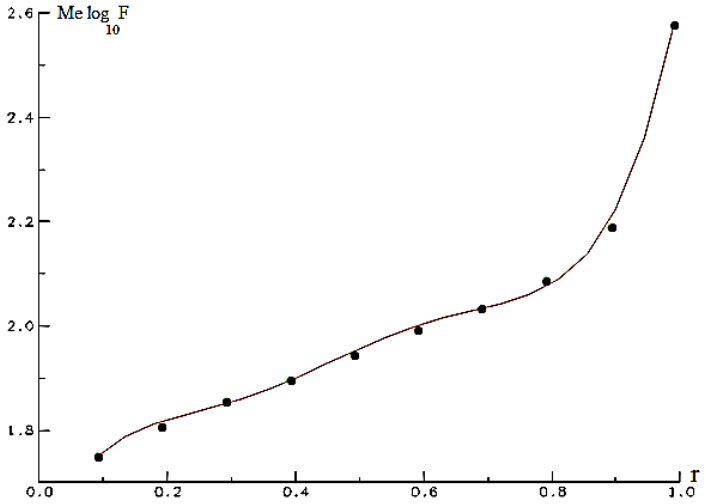

Рис. 1. Влияние корреляции регрессоров на F-статистику

Улучшение F-статистики с увеличением коррелированности регрессоров *r* объясняется изменением числа степеней свободы вследствие исключения значимых регрессоров.

Изменение качества моделей лучше всего характеризует относительная ошибка δ (рис. 2).

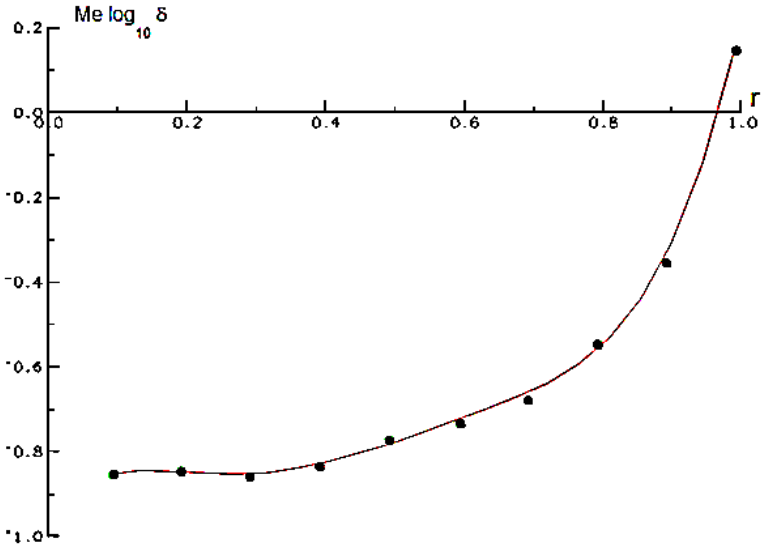

Рис. 2. Влияние корреляции регрессоров на относительную ошибку δ.

Вывод:

В условиях мультиколлинеарности МНК может обеспечить эффективную аппроксимацию поверхности отклика, но не оценку реального вклада отдельных

регрессоров в изменение значений отклика. Использовать результаты многомерного регрессионного анализа для изучения механизма явлений и оптимизации управляемых процессов нельзя.

При обучении слушателей Центрального института повышения квалификации (ЦИПК) Минатома нами использовалась усложненная схема исследований. Слушатели рассчитывали «модель моделей» (6) для двух переменных: ошибки δ и показателя относительной адекватности уравнений $f = \sqrt{F/F_T}$. (F_T — табличное значение F – критерия).

Для построения зависимостей указанных переменных от «факторов» r, n, σ использовался план второго порядка Бокса B_3, включающий 8 опытов полного факторного эксперимента ПФЭ 2^3, 6 опытов в середине граней куба и 3 опыта в центре куба.

Нулевые уровни факторов: $r^0 = 0.85;\ \sigma^0 = 1.1;\ n^0 = 30;$

Интервалы варьирования: $\Delta r = 0.14;\ \Delta\sigma = 0.5;\ \Delta n = 20.$

Нет необходимости описывать, насколько поучительной оказывается такая форма обучения.

Обратим внимание на информационную сторону оценивания регрессии: если в исходных данных недостаточно информации об удельном вкладе регрессоров в изменение отклика, то никакой алгоритм не поможет. Не поможет и увеличение точности вычислений.

В статье [2] автором предложена мера количества информации об удельном влиянии регрессоров

$$I_x \cong 0.7\frac{n(1-\bar{r})F}{m-1},$$

где $\bar{r}$ — среднее значение коэффициентов парной корреляции регрессоров.

Результаты имитационных испытаний алгоритмов [1, 2] подтвердили наличие тесной связи между I_x и ошибкой оценки коэффициентов δ.

Основные результаты испытаний:

1) Использование ортогональных координат (главных компонент). Применение малого числа главных компонент для оценивания регрессионных коэффициентов в общем случае недопустимо; число учитываемых собственных направлений матрицы X^TX должно быть не менее трех;

2) Ридж-оценки. При $I_x < 1$ ридж-оценки точнее МНК-оценок. Увеличение I_x приводит к незначительному проигрышу в точности ридж-оценок относительно МНК-оценок.

Вычислительный подход к обучению статистике

У. Гренандер и В. Фрайбергер, основоположники научного направления, получившего название «вычислительной вероятности и статистики» [10], предложили совместить математический и вычислительный подход, включить в обучение реальные задачи из разных отраслей промышленности. И изложили свою методологию на эффектных примерах реализации алгоритмов на языке APL, который рекомендовали как наиболее полезный для поставленных целей.

Освоение технологии вычислительной статистики (computational statistics) позволяет трезво оценивать возможности статистического анализа и моделирования, отчетливо представлять опасности характерных ошибок:

1. Неверная постановка задачи, связанная с односторонностью

представления об анализируемом объекте, с выделением частных проблем и локализацией решений без учета их удельного веса и системных связей.

2. Неумеренная типизация, использование стандартных программ и методик в оригинальных и сложных ситуациях.

3. Необоснованное (когда известны простые методы) или ненужное (когда задача этого не ст**о**ит) усложнение решений.

4. Попытки исправить отсутствие информации изощренной обработкой данных, преувеличение возможностей преобразования данных.

5. Фетишизация результата, представление его абсолютно надежным, единственным, безупречно строгим, необоснованное распространение выводов на далекие от исходных данных условия.

В настоящее время обучение статистике перегружено множеством теорем и доказательств. При освоении пакетов прикладных программ (SAS, SPSS, STATISTICA, и пр.) обучаемый формально отрабатывает требуемую последовательность действий, не вдаваясь в тонкости, чему способствует развитый сервис ППП. Об учете границ применимости используемых процедур и грамотной интерпретации результатов не может быть речи. Он может в основном верно и резво «стрелять с бедра», решая типовые примеры, но не способен ставить задачи. В лучшем случае это исполнитель, который исправно выполняет расчеты по заданным алгоритмам.

Необходимо учитывать следующее:

- Библиотеки (пакеты) программ предназначены для решения типовых задач и не предполагают отступлений от стандартов и предпочтений разработчиков этих библиотек. В то же время существует и разрабатывается множество алгоритмов, которые предназначены для решений в ситуациях, отличающихся от типовых.
- Стремление разработчиков типовых программ сделать их доступными для непрофессионалов приводит к тому, что отдельные пользователи пользуются ими неуместно, не раздумывая над постановкой задач, над корректностью применения отдельных процедур для реальных условий и ограничений, для разнородных и «засоренных» данных.
- Объекты исследований не всегда втискиваются в готовые формы. В таких случаях встает вопрос о самостоятельном конструировании решений. Веским аргументом в пользу независимых решений является тот факт, что кроме испытанных, универсальных средств анализа данных на рынке программных продуктов предлагаются «оригинальные» поделки, изобилующих грубыми ошибками.
- Однозначное «натаскивание» на стандарты убивает способности к самостоятельности, творчеству, поиску.
- Самостоятельное конструирование решений открывает возможности творческого использования разнообразных (в том числе новых) алгоритмов и их комбинаций, оперативной правки и тонкой настройки процесса расчетов. Создавая собственный программный продукт для определенного класса задач, разработчики имеют больше свободы и возможностей для его последующего развития и совершенствования.
- Разумеется, оригинальность не может быть самоцелью; цель — эффективный, практически полезный результат. Результат, недостижимый (или не вполне достижимый) стандартными средствами, и в то же время требующий от разработчиков соответствующей квалификации и опыта.

Вычислительный подход не только способствует более глубокому пониманию теории, но также развитию навыков творческого, осознанного применения процедур обработки данных и имитационного моделирования, которые пользователь может применить при разработке решений достаточно сложных задач.

Анализ зарубежных публикаций показывает, что до сих пор доминирует тенденция использования «грубых инструментов и наивных методов» для решения задач, требующих «гибкой стратегии решения» [11]. В то же время усиливается роль вычислительной статистики. Например, в учебнике [12] в излагаемый материал включено более 300 примеров реальных статистических исследований. В статье [13] помимо использования реальных примеров при обучении, предлагается отвести больше времени информационным и вычислительным технологиям, которые, так или иначе, используются в работе современных статистиков. Эти технологии изображены на рис. 3.

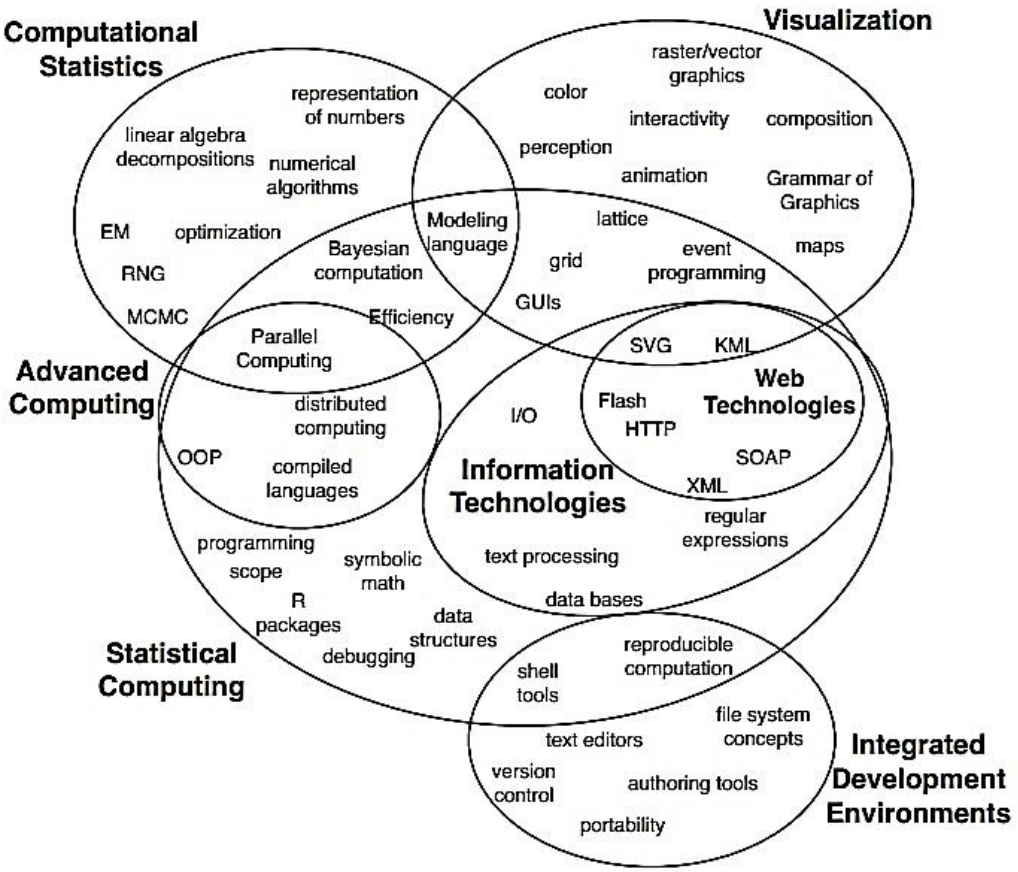

Рис 3. Информационные и вычислительные технологии, имеющие отношение к статистике.

Выводы

В докладе рассмотрены некоторые проблемы, связанные с необходимостью совершенствования технологии анализа данных сложной природы. В стороне остается

вопрос о рациональном соотношении стандартных и нестандартных решений. Следует признать, что нестандартный подход уместен только в том случае, когда объект исследований отклоняется от заданных шаблонов, а исследователь имеет достаточную квалификацию и опыт. Кроме того, такое качество, как умение ставить задачи, сродни изобретательству, которому так же трудно учить. Стандарт дает в среднем верное решение, в то время как «изобредеи» чреваты непредсказуемыми потерями.

Выход один — учиться. А это вновь computational statistics и опыт.

Литература

1. Реймаров Г.А. Имитационные испытания регрессионных алгоритмов и конструирование моделей. // Электронное моделирование, 1988, №1.
2. Меньших Б.И., Реймаров Г.А. Регрессионные модели в задачах анализа и управления. М.: ЦНИИАтоминформ, 1988 г. — 94 с.
3. Реймаров Г.А., Лукша О.П. АПЛ как инструмент конструирования математических моделей. В кн. ДИАЛОГ-87. Материалы конференции. Тбилиси: Мецниереба, 1988. С. 128-130.
4. Реймаров Г.А. Комплексная оценка персонала. Инженерный подход к управлению качеством труда. М.: Издательство ЛКИ, 2010. —424 с.
5. Реймаров Г.А., Лавров А.С. Библиотека инструментальных средств анализа данных на языке APL2 (БИС «Аналитик»). Учебное пособие. Lambert Academic Publishing. Project 116244. ISBN 978-3-659-67833-2. 2015. — 164с.
6. Реймаров Г.А., Лавров А.С. Вычислительный подход к обучению статистике. // Доклады XIII Международной конференции «Безопасность АЭС и подготовка кадров - 2013». 1.10 – 5.10 2013 г. ИАТЭ, С. 161-166.
7. Бахметьев И.И., Реймаров Г.А. Компьютерный анализ динамики кадровой обеспеченности сельскохозяйственных организаций агропромышленного комплекса Российской Федерации. // "Экономика, труд, управление в сельском хозяйстве", 2014, №2.
8. Браун Дж., С. Пейкин, Р. Поливка. APL время пришло. Пер. с англ. М.: «РЕДСтарс», 1995 — 478 с.
9. Hoerl A.E. Application of ridge analysis to regression problems. // Chemical Engineering Progress. — 1962. V. 58, N1. P. 54-59.
10. Гренандер У., Фрайбергер В. Краткий курс вычислительной вероятности и статистики, пер. с англ. М: Наука, 1978. — 192 с.
11. E.N. Brown, R.E. Kass. What is Statistics? // The American Statistician. Том 63. №2. 2009 – с. 105-123.
12. J.M. Utts, R.F. Heckard. Mind on Statistics. — Cengage Lerning, 2010. 717 с.

Deborah Nolan, Duncan Temple Lang. Computing in the Statistics Curricula. // The American Statistician. Том 62. №2. 2010. – С. 97-107.