КОМПЛЕКСНОЕ СТАТИСТИЧЕСКОЕ И ИМИТАЦИОННОЕ МОДЕЛИРОВАНИЕ ПРИ СЕЛЕКЦИИ РЕГРЕССИОННЫХ МОДЕЛЕЙ

Цуканов А.В. (Севастополь)

Одна из существенных проблем в статистическом моделировании состоит в выборе регрессионной модели, которая должна отвечать поставленным целям. В качестве целей построения модели обычно рассматриваются следующие: прогнозирование определенных характеристик объекта либо в заданной области управляемых переменных либо в будущем; построение модели, по которой строится в дальнейшем система управления; оптимизация, то есть поиск значений управляемых переменных, при которых заданные характеристики объекта отвечают условиям оптимальности; классификация или распознавание образов и т.д.

Первой проблемой, которая исторически решалась в прикладной статистике в этом направлении была проблема построения метода оценивания параметров модели по имеющимся статистическим данным. Решением этой проблемы стали такие классические методы оценивания параметров, как метод наименьших квадратов, взвешенный метод наименьших квадратов, ридж-регрессия и множество других. Затем оказалось, что если данные собраны плохо, то исправить ситуацию изощренными методами оценивания параметров практически невозможно. Эта проблема привела к развитию теории и практики планирования эксперимента. Затем выяснилось, что если модель выбрана неправильно, то ни хороший план эксперимента ни метод оценивания не спасают ситуацию. Появилось направление в статистике связанное с селекцией моделей и были предложены методы выборы моделей, такие как метод группового учета аргументов [1], метод скользящего контроля [3] и много других.

Оказалось, что результат статистического моделирования зависит от количества информации, имеющейся у исследователя об изучаемом объекте. Если вид модели и ее параметры известны (например, из законов физики) то статистическое моделирование и не нужно. Если ничего не известно, кроме эмпирических данных, то здесь всегда есть вероятность ошибочного решения. При этом вероятность ошибки зависит от объема статистического материала. (Правда не всегда. Есть задачи, когда результат не обладает свойством состоятельности и с увеличением данных не сходится к истинному значению).

Шаг в направлении использования имитационных методов в прикладной статистике был сделан при создании метода *bootstrap* [3], компьютерного метода исследования распределения статистик вероятностных распределений, основанного на многократной генерации выборок методом Монте-Карло на базе имеющейся выборки.

В связи с ростом использования аналитических методов в управлении предприятиями вновь возрос интерес к построению оптимальных статистических моделей [2], которые позволяют предприятиям получать конкурентные преимущества.

В данной работе рассматривается возможность использования методов имитационного моделирования для оптимального конструирования процесса построения статистической целеориентированной модели объекта исследования.

Пусть истинная модель объекта имеет аддитивную зависимость между детерминированной и случайной частями уравнения

$$y = \eta(x) + \varepsilon \ , \qquad (1)$$

где y – наблюдаемая зависимая переменная, x – вектор независимых переменных (факторов), $\eta(x)$ – истинная, но неизвестная зависимость от x и ε – случайная переменная с нулевым средним и дисперсией σ^2.

Пусть также рассматривается случай, когда возможные модели носят вложенный характер $\eta_j(x) \in S_j$, *(j=1,2,..,q).* Здесь $S_1 \subset S_2 \subset \ldots S_q$, S_j – множество всех возможных моделей для класса j *(j=1,2,..,q),* q – число классов.

В частности для линейных по параметрам моделей можно записать

$$\eta_j(x, \alpha_j) = f_j'(x)\alpha_j, \text{ (j=1,2,..,q)}, \quad (2)$$

где $f_j'(x)$ – вектор известных функций в точке x и α_j – вектор неизвестных параметров размерности n_j.

Будем предполагать, что модель используется для целей прогноза величины y в заданной области W_1 значений величины x, тогда в качестве критерия оптимальности регрессионной модели можно рассмотреть величину ошибки прогноза, полученной по этой модели, которую для модели j можно будет записать в виде

$$L(j) = \int_{W_1} E(y - \eta_j(x, \alpha_j))^2 dx / \int_{W_1} dx, \quad \text{(j=1,2,….q).} \quad (3)$$

Здесь вычисляется среднеквадратическая ошибка прогноза для каждой модели j области W_1. Если дисперсия σ^2 является константой, то функцию потерь для модели j можно записать виде:

$$L(j) = \sigma^2 + \int_{W_1} (\eta(x) - \eta_j(x, \alpha_j))^2 dx / \int_{W_1} dx \quad (4)$$

Предположим также, что в качестве критерия селекции моделей используется классический критерий Маллоуса [6] $C_j = RSS_j + 2n_j\hat{\sigma}^2$. Здесь $\hat{\sigma}^2$ - оценка дисперсии σ^2 для модели самого высокого порядка q.

Для того, чтобы выбрать аппроксимирующую модель, необходимо иметь значения экспериментальных данных в каждой точке $x_i^T = (x_{1i}, x_{2i}, \ldots x_{ki})$ (i=1,2,….N). Пусть матрица $F_j^T = [f_j(x_1), f_j(x_2), \ldots f_j(x_N)]$ - матрица значений вектора f_j в N экспериментальных точках. Предполагается, что матрица F_j имеет полный ранг.

$Y^T = [y_1, y_2, \ldots y_N]$ - вектор N наблюдений над функцией отклика.

В результате процесса статистического моделирования выбирается модель из класса S_j и оцениваются ее параметры α_j . В связи с неопределенностью модели и наличием случайной составляющей, число j является случайным числом и для оценки качества процедуры моделирования можно использовать критерий математического ожидания среднеквадратической ошибки прогноза (AMSEP) [4,5]

$$AMSEP = \sum_{j=1}^{q} \nu_j L_j, \quad (5)$$

здесь ν_j – вероятность выбора модели с номером j.

Проблема заключается в том, что данный критерий зависит от неизвестных параметров модели. Для оценки возможных значений данного критерия предлагается воспользоваться методом имитационного моделирования, то есть промоделировать процесс построения статистической модели при всех возможных значениях параметров. При этом важное значение приобретает априорная информация об области допустимых значений переменных и параметров. Имея эту информацию и результаты имитационного моделирования можно настроить процедуру статистического моделирования, которая будет минимизировать критерий (5).

Рассмотрим пример с выбором метода оценивания. Пусть необходимо выбрать один из двух методов. Например, метод наименьших квадратов или метод ридж-регрессии.

Если модель линейная по параметрам и используется метод наименьших квадратов, тогда оценка вектора параметров для модели из класса S_j

$$\hat{\alpha}_{j(MNK)} = (F_j^T F_j)^{-1} F_j^T Y, \quad (6)$$

Для той же самой модели оценки ридж-регрессии оцениваются по формуле

$$\hat{\alpha}_{j(RIDG)} = (F_j^T F_j + rI_j)^{-1} F_j^T Y, \quad (7)$$

где r - ридж-параметр регуляризации.

Для иллюстрации имитационного моделирования этих двух процедур рассмотрим простой пример идентификации линейной регрессии.

Рассматривались следующие вложенные модели:

недоопределенная модель : $y_1 = \alpha_0 + a_1 x_1 + \varepsilon$;

истинная модель: $y_2 = \alpha_0 + a_1 x_1 + \alpha_2 x_2 + \varepsilon$;

переопределенная модель: $y_3 = \alpha_0 + a_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \varepsilon$.

Матрица экспериментальных данных содержала 12 сильно коррелированных наборов значений переменных. имитация статистической процедуры проводилась с помощью системы MatLab. Вычисления повторялись 1000 раз для дисперсии $\sigma^2 = 1$, значения коэффициентов в тестовых моделях принимались $\alpha_0 = \alpha_1 = 1$, $\alpha 2$, $\alpha 3$ изменялись от 0 до 1 с шагом 0,1. На рисунке 1 показаны результаты имитационного моделирования.

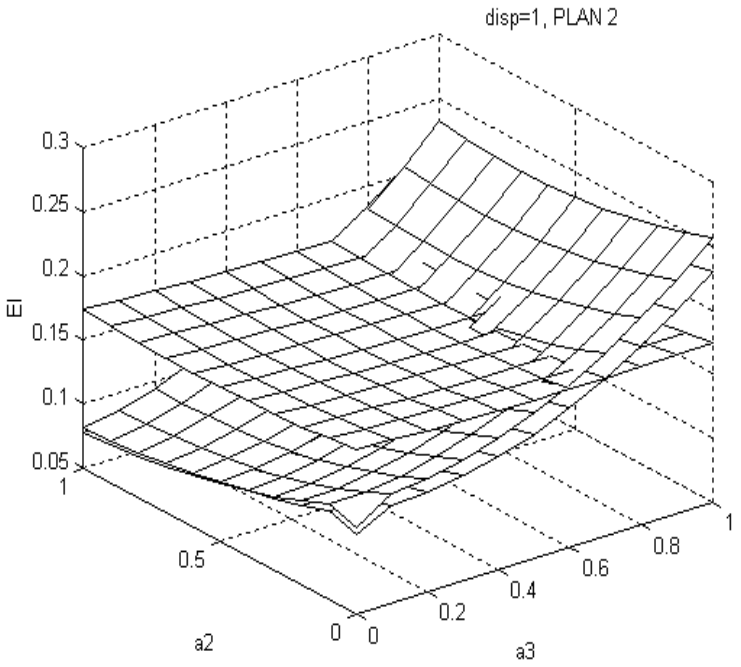

Рис. 1. Результаты имитационного эксперимента по выбору метода оценивания

На основании данного имитационного эксперимента можно определить области параметров, для которых эффективнее использовать метод наименьших квадратов, а для которых метод ридж-регрессии.

В качестве второго примера рассмотрим сравнение последовательной процедуры проведения эксперимента и статической процедуры [8]. Случай двух шагового эксперимента был исследован в работе [9].

Пусть имеется только одна независимая переменная x и области планирования эксперимента и прогноза совпадают $-1 \le x \le +1$.

Пусть рассматриваются три класса полиномиальных моделей:

Модель 1: $y_1 = \alpha_0 + \alpha_1 x + \varepsilon$;

Модель2: $y_2 = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \varepsilon$; (8)

Модель 3: $y_3 = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \alpha_3 x^3 + \varepsilon$.

Рассматривается следующая последовательная процедура последовательной идентификации модели.

Выбирается начальная матрица плана эксперимента X_1 $(N_1 \times k)$. Эта матрица должна позволять оценивать модели из всех рассматриваемых классов. По результатам первой стадии эксперимента методом наименьших квадратов можно получить оценки векторов α_j для всех моделей. Далее выбирается наилучшая модель с помощью критерия C_p. Пусть эта модель имеет номер t. На следующей стадии эксперимента одна экспериментальная точка добавляется в план эксперимента X_1 из возможных точек оптимального плана для модели t . После получения результата эксперимента в этой точке, снова проводится селекция по критерию C_p . Пусть это модель с номером t_2 . Следующая точка выбирается из оптимального плана для модели t_2. Эти действия проводятся до тех пор, пока все экспериментальны ресурсы не будут исчерпаны.

Для сравнения был выбран статистический g-оптимальный план, состоящий из $N=24$ экспериментальных точек.

Случайная составляющая модели ε генерировалась с нулевым средним и с единичной дисперсией. Число имитационных повторов было равным 1000. Критерий AMSEP вычислялся для окончательно полученного набора экспериментальных точек. Имитационное моделирование выполнялось вновь с использованием системы MatLab.

На рисунке 2 показано изменение критерия в связи с изменением коэффициента α2. Мы видим, что зависимость имеет максимум и затем монотонно падает с увеличением параметра α_2. Это падение носит монотонный характер. При больших значениях исследованного параметра обе процедуры дают одинаковое значение критерия. Это объясняется тем фактом, что в этом случае однозначно выбирается истинная модель.

Результаты имитационного эксперимента показывают, что выбор стратегии эксперимента оказывает значительное влияние на критерий AMSEP и при $\alpha_2 > 0.8$ последовательная процедура лучше одношаговой.

Выводы

Предложенная в настоящей работе методика комплексного применения статистических методов моделирования объектов исследования и методов имитационного моделирования позволяет осуществить выбор эффективной процедуры построения эмпирических моделей сложных объектов исследования. Дальнейшие исследования в этом направлении могут быть продолжены по разработке комплексных процедур по применению методов селекции моделей, оценивания статистических моделей, планирования эксперимента и других частных статистических методов.

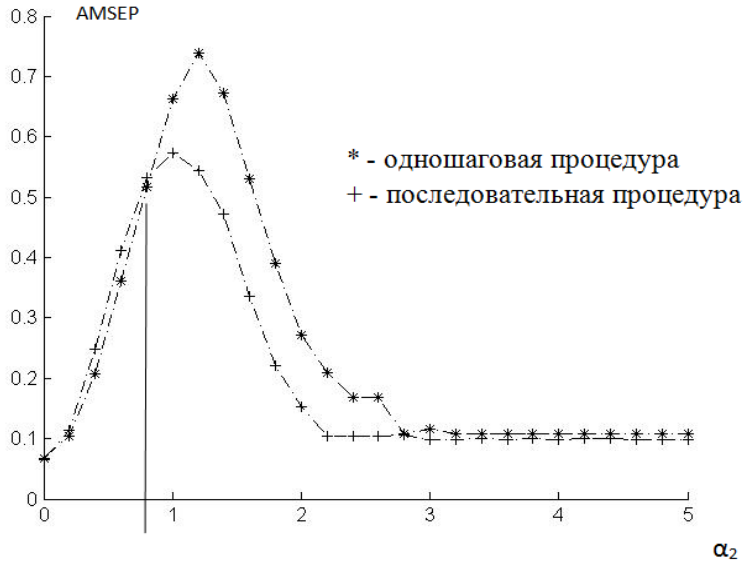

Рис. 2. Результаты имитационного эксперимента по исследованию последовательной процедуры

Литература

1. Ивахненко А.Г. Индуктивный метод самоорганизации моделей сложных систем. – Киев: Наук. Думка, 1982. – 296 с.
2. Цуканов А.В. Особенности планирования экспериментов при моделировании бизнес-процессов// «Менеджмент малого и среднего бизнеса: реинжиниринг»/ Труды одиннадцатой международной науч.-практ. конф.- Севастополь: СевНТУ. – 2014. С. 50-53.
3. Эфрон Б. Нетрадиционные методы многомерного статистического анализа. – М.: Финансы и статистика, 1988. – 263 с.
4. Herzberg A.M., Tsukanov A.V. The design of Experiment for Model selection: Minimization of the Expected Mean-Squared Error // Utilitas Mathematica. – 1995. – Vol. 47. P. 85 – 96.
5. Herzberg A.M., Tsukanov A.V. A Note on the Choice of the Best Selection Criterion for the Optimal Regression Model // Utilitas Mathematica. – 1999. – Vol. 55. P. 243 – 254.
6. Mallows C. L. Some Comments on Cp. // Technometrics. 1973. – Vol. 15, № 4. — P. 661— 676.
7. Potanina M.V., Tsukanov A.V. The comparison analysis of the efficiency of the selection method of variable and ridge regression// Systems control information Methodologies & application. – AMSE, China 1995. – P. 279-283.
8. Tsukanov A.V. Sequential procedure for design of experiment and selection of regression models.- 4th International Conference in Inductive Modeling ICIM' 2013: Proceedings, Kyiv, Ukraine, September 16-20, 2013. - Kyiv: International Research & Training Center for Information Technologies &Systems, 2013. - P. 124-127.
9. Tsukanov A.V. Tsepkova N.A. The Monte-Carlo comparison of the two strategies of the model identification// Advances in modeling & analysis.- 1995.- Ser.B.- V.32.- N3.- P.55-63.