Архипова С.А., канд. техн. наук, доцент
Национальный технический университет Украины
"Киевский политехнический институт им. Игоря Сикорского", г. Киев
Факультет социологии и права, кафедра философии, доцент

ПОСТРОЕНИЕ РЕГРЕССИОННЫХ МОДЕЛЕЙ ПРИ НЕПОЛНОЙ ИНФОРМАЦИИ О ПОГРЕШНОСТЯХ ИЗМЕРЕНИЙ

Рассматривается задача построения наилучшей модели $y = \mu(x_1, x_2, ...)$ в классе линейных регрессионных моделей по экспериментально полученным данным: матрице плана $[x_{ij}]$ и вектору $[z_i]=[y_i]+[e_i]$ результатов измерения значений зависимой переменной *Y*, наблюдаемой в присутствии аддитивного случайного шума *E*. Исследования выполняются на двухфакторном x_1, x_2 тестовом примере, объем выборки *n*=200 значений.

Последовательность значений $x_{i1}, i = \overline{1,n}$, задается программным генератором псевдослучайных чисел. Вектор точных значений переменной рассчитывается в соответствии с истинным уравнением регрессии

$$y = x_1 + x_2 + 0{,}04x_1^2 + 0{,}05x_2^2 + 0{,}003x_1x_2^2. \qquad (1)$$

Для селекции наилучшей модели применим один из вариантов процедуры регрессионного анализа, описанный в [1]. Оценивание параметров моделей производится методом наименьших квадратов (МНК). Процедуры селекции [3] сводятся к следующему:

(1) Модели регрессии, в зависимости от количества m_j входящих в них элементов, разбиваются на девять классов: А, Б, В, Г, Д, Е, Ж, З, И. Класс А представлен моделью среднего $y = a_0$. Класс Б включает модели, состоящие из одного регрессора, класс В – из двух регрессоров и т.д. Всего в девяти классах было отобрано и исследовано 29 моделей.

(2) Для каждой из 29-ти моделей рассчитывается средний квадрат ошибки

аппроксимации $S^2 = 1/n \sum_{i=1}^{n} (z_i - \tilde{y}_i)^2$, где $\tilde{y}_i$ – модельное значение зависимой переменной.

(3) В каждом классе отбираются одна или несколько моделей, имеющих минимальные для этого класса значения S_j^2.

(4) Производится сопоставление моделей, проверка адекватности моделей исходным данным, проверка необходимости усложнения этих моделей. Конечная цель этого анализа – выбор структуры аппроксимативной модели.

В результате получен ряд минимальных в классах Б–И значений S_j^2:

257,41; 191,96; 142,92; 132,79; 121,49; 121,06; 120,24; 120,19.

Очевидно, с введением новых переменных в модель скорость уменьшения показателя S_j^2 замедляется. С этих позиций (согласно [1,2]) можно полагать адекватными модели в классах Г с тремя регрессорами ($S_Г^2$ =142,92) или Д с четырьмя регрессорами ($S_Д^2$ =132,79).

Эмпиризм подобного подхода обычно побуждает исследователя к поиску других способов селекции модели. При этом прибегают к априорному постулированию нормальности распределения погрешностей измерения E значения зависимой переменной Y, что позволяет сразу решить две проблемы: обосновать оптимальность применения МНК для оценивания параметров модели и использовать наработки классического регрессионного анализа по оценке качества линейной регрессии для селекции структуры модели.

Для проверки значимости уравнений регрессии применяют статистику

$$\tilde{F}_j = \frac{S_1^2/(n-1)}{S_j^2/(n-m_j)}, \qquad (2)$$

которая определяет во сколько раз исследуемая j-тая модель лучше предсказывает результаты эксперимента, чем среднее $\bar{y}$, т.е. модель $y(1) = a_0$ из класса А .

Расчет [3] значений $\tilde{F}_j$ для всех 29-ти моделей определяет максимальное

значение $\widetilde{F}_{27}$=4,58, что соответствует семиэлементной модели μ_{27} из класса З, что подтверждает только статистическую значимость этой модели по отношению к модели среднего, а не ее оптимальность.

Дальнейший подбор модели [1] основан на применении более сложного попарного сопоставительного анализа моделей j и r при поэлементном введении регрессоров в структуру модели по статистике

$$\widetilde{F}_{j\to r} = \frac{S_j{}^2 - S_r{}^2}{S_r{}^2} \frac{n - m_r}{m_r - m_j} \ . \qquad (3)$$

Если теоретическое значение меньше $\widetilde{F}_{j\to r}$ следует признать, что усложнение модели приводит к значимому росту ее качества.

По результатам исследований [3] можно предположить, что наиболее адекватны исходным данным шестиэлементные модели класса Ж.

Если сопоставить итог селекции моделей по показателю $\widetilde{F}_{j\to r}$ с моделями, отобранными по показателю S^2, очевидно наличие существенных различий в сложности отобранных групп моделей: область возможных решений перекрывает четыре класса Г, Д, Е, Ж. Признать удовлетворительным подобный результат нельзя из-за содержащейся в нем значительной неопределенности. Необходимо проведение дополнительных исследований, позволяющих радикально сузить эту область.

Один из возможных способов уточнения вида модели состоит в проверке значимости найденных оценок коэффициентов моделей [2,4], целиком базирующейся на гипотезе нормальности погрешности.

Поэтому актуальным представляется вопрос о возможности проверки гипотезы нормальности распределения погрешности значений зависимой переменной. В частности, для этого в [2] рекомендуется проверка нормальности остатков (невязок) $\varepsilon_i = z_i - \widetilde{y}_i$. Однако, как показали исследования, описанные в статье [3], если для параметрической идентификации параметров модели (1) использовался МНК, в подавляющем

большинстве случаев, независимо от вида исходного закона распределения погрешности e_i, распределения невязки ε_i как правило оказывается нормальным. Другими словами, применение МНК обуславливает эффект нормализации невязки.

Таким образом, если следовать рекомендациям [2], то практически всегда будет принята гипотеза нормального распределения погрешностей в значениях зависимой переменной, из чего очевидно будет следовать нормальность распределения оценок параметров исследуемой модели.

Результаты исследований позволяют сделать следующие выводы:

– процедура классического регрессионного анализа не содержит способов надежной селекции структуры модели при отсутствии априорных сведений о нормальности погрешности *E*;

– апостериорное утверждение о принятии гипотезы нормальности остатков регрессии не гарантирует нормальности исходной погрешности *E* и, следовательно, не может служить обоснованием принятия положений регрессионного анализа, опирающихся на нормальность распределения *E*, в частности, гипотезы о нормальности распределения оценок параметров и базирующихся на ней методов проверки значимости оценок и модели регрессии;

– при отсутствии априорной информации о виде распределения погрешности измерений следует применять методы структурной идентификации свободные от вида распределения.

Литература

1.Себер Дж. Линейный регрессионный анализ. - М.: Мир, 1980.- 456с.

2.Львовский Е.Н. Статистические методы построения эмпирических формул. - М.: Высш. школа, 1982. - 224с.

3.Архипов А.Е., Архипова С.А. О некорректности параметризации распределений оценок при решении задачи идентификации. //Адаптивні системи автоматичного управління. Межвідом. науково-техн. зб. - Дніпропетровськ: Системні технології, 1999. - Вип.2(22). - с. 114-121.