MSC 62K99

НЕКЛАССИЧЕСКИЙ ВАРИАНТ РЕГРЕССИОННОГО АНАЛИЗА

*М.М. Ошхунов, **З.М. Ошхунова, *М.А. Джанкулаева

*Кабардино-Балкарский государственный университет им. Х.М. Бербекова,
ул. Чернышевского, 173, Нальчик, 360004, Россия, e-mail: muaed@inbox.ru,
madina.dzhan@gmail.com;
**Кабардино-Балкарский филиал ОАО Ростелеком,
ул. Головко, 4, Нальчик, 360000, Россия, e-mail: zalina_oshhunova@mail.ru

Аннотация. Предлагается алгоритм нахождения регрессионных функций, отличный от классического подхода. Суть метода заключается в выборе линий регрессии по минимуму суммы квадратов расстояний от статистических точек. Такой подход приводит к неединственному решению (в отличие от классического алгоритма выбора прямой), что имеет геометрическое объяснение. Даны примеры расчета коэффициента корреляции по двум методам, которые показали бо́льшую эффективность предлагаемого метода по степени отклонения от статистических данных. Предпринята попытка распространить данный алгоритм на задачи многофакторного регрессионного анализа.

Ключевые слова: регрессионный анализ, метод наименьших квадратов, неклассический вариант метод наименьших квадратов, дисперсия.

Как известно |1|, практически вся прикладная классическая статистика основана па методах, разработанных применительно к нормально распределённым величинам. В последнее время появилось большое количество публикаций с нападками па нормальный закон распределения Гаусса. Утверждается, что нормальный закон распределения в экономике встречается весьма редко. В этом случае традиционные методы анализа статистической экономической информации не пригодны. Это заключение весьма непростое, т.к. требуется разработать другие, отличные от классических, методы анализа.

Регрессионный анализ относится к одним из наиболее часто используемых приёмов исследования статистической информации. Он позволяет находить средневзвешенные тренды, которые дают возможность прогнозировать динамические процессы, например, в экономике на основе накопленной ранее информации. Объективность таких прогнозов иногда не очень высокая, но есть несколько требований, когда такой прогноз может быть научно оправдан. Начнём с однофакторного линейного анализа. Требуется найти тесноту связи между двумя признаками в виде линейной зависимости

$$C = kx + b\,. \tag{1}$$

Коэффициенты k, b находят по минимуму суммы квадратов отклонений но оси Oy статистических данных от линейной функции.

Работа выполнена при финансовой поддержке РФФИ в рамках научного проекта №12-07-00624 а, №13-07-01003 а.

Если ввести новые переменные $\tilde{x}_i = x_i - \bar{x}, \quad \tilde{y}_i = y_i - \bar{y}$, где

$$\bar{x} = \left(\sum_{i=1}^{n} x_i\right)/n, \quad \bar{y} = \left(\sum_{i=1}^{n} y_i\right)/n,$$

то оптимальные значения этих коэффициентов вычисляются по формулам

$$k = \frac{\sum_{i=1}^{n} \tilde{x}_i \tilde{y}_i}{\sum_{i=1}^{n} \tilde{x}_i^2} = \frac{\mu_{xy}}{\sigma_x^2}. \tag{2}$$

$$b = \bar{y} - k\bar{x} \tag{3}$$

Формулы (2), (3) определяют параметры линейной регрессии по методу наименьших квадратов.

Уравнение (1) удобно записать в виде:

$$y - \bar{y} = r\frac{\sigma_y}{\sigma_x}(x - \bar{x}) \ , \tag{4}$$

где

$$r = \frac{\mu_{xy}}{\sigma_x \sigma_y} \ , \tag{5}$$

$$\mu_{xy} = \left(\sum_{i=1}^{n} \tilde{x}_i \tilde{y}_i\right) \Big/ n.$$

Коэффициент r в формуле (5), как известно, носит название коэффициента корреляции и обладает свойством $|r| \leq 1$. Если $|r| \approx 0$, то случайные величины y, x почти не связаны и, наоборот, если $|r| \approx 1$, то теснота связи между этими величинами – сильная.

Заметим, что предложенный алгоритм регрессии никак не зависит от того, является ли плотность распределения статистической информации нормальной или нет.

Заменим условие минимума суммы квадратов отклонений по оси Oy на минимизацию суммы квадратов расстояний

$$S(k, b) = \sum_{i=1}^{n} \frac{(kx_i + b - y_i)^2}{k^2 + 1} \to \min .$$

Такая замена, в отличие от классического подхода, приводит к двойственности решения и особенно целесообразна, когда теснота связи между факторами y, x — сильная $|2|$.

Оптимальные значения коэффициента k в случае $\sum_{i=1}^{n} \tilde{x}_i \tilde{y}_i \neq 0$ определяются из решения уравнения

$$k^2 + \frac{\sum_{i=1}^{n} (\tilde{y}_i^2 - \tilde{x}_i^2)}{\sum_{i=1}^{n} \tilde{x}_i \tilde{y}_i} \cdot k - 1 = 0 \, ,$$

или в общепринятых терминах статистики

$$k^2 + \frac{\sigma_y^2 - \sigma_x^2}{\mu_{xy}} k - 1 = 0\,. \tag{6}$$

После нахождения корней k_1, k_2 уравнения (6), параметры прямой b_1, b_2 вычисляются по формулам:

$$b_1 = \bar{y} - k_1\bar{x}\,, \qquad b_2 = \bar{y} - k_2\bar{x}\,,$$

где

$$k_1 = \frac{-\alpha + \sqrt{\alpha^2+4}}{2}\,, \qquad k_2 = \frac{-\alpha - \sqrt{\alpha^2+4}}{2}\,, \qquad \alpha = \frac{\sigma_y^2 - \sigma_x^2}{\mu_{xy}}\,.$$

Случайные величины

$$\frac{(k_1\tilde{x}_i - \tilde{y}_i)}{\sqrt{k_1^2+1}}\,, \qquad \frac{(k_2\tilde{x}_i - \tilde{y}_i)}{\sqrt{k_2^2+1}}\,,$$

которые равны отклонениям по расстоянию от точек с координатами $(x_i; y_i)$ до прямых $y = k_1x + b_1$, $y = k_2x + b_2$ должны быть нормально распределенными с нулевыми математическими ожиданиями. Только в этом случае ими можно пользоваться для прогноза динамических процессов в экономике.

Исследования в экономике с использованием множественного регрессионного анализа, предъявляют к статистическому распределению такие же требования нормальности отклонений, как это описано выше.

Таким образом, для обработки статистической информации в экономике методами регрессионного анализа с последующим использованием полученных зависимостей для прогнозных решений, рекомендуется алгоритм минимизации суммы квадратов расстояний, т.к. он даёт существенно меньше погрешности при прогнозе. Согласно приведенным выше формулам

$$S_1(k) = \sum_{i=1}^{n} (k\tilde{x}_i - \tilde{y}_i)^2 = n\left(\sigma_2^x k^2 - 2\mu_{xy}k + \sigma_y^2\right)\,. \tag{7}$$

Оптимальное значение k из формулы (7) позволяет напрямую подсчитать сумму квадратов отклонений по оси Oy:

$$S_1(k)_{<8=} = n\left(\sigma_y^2 - \frac{\mu_{xy}^2}{\sigma_x^2}\right) = \sum_{i=1}^{n}\tilde{y}_y^2 - \left(\sum_{i=1}^{n}\tilde{x}_i\tilde{y}_i\right)^2 \Big/ \left(\sum_{i=1}^{n}\tilde{x}_i^2\right).$$

В случае, когда регрессионная прямая выбирается из минимума суммы квадратов расстояний, имеем

$$S_2(k) = \sum_{i=1}^{n} \frac{(k\tilde{x}_i - \tilde{y}_i)^2}{k^2+1} = \frac{S_1(k)}{k^2+1}\,. \tag{8}$$

Из формулы (8) следует важный вывод. Независимо от значения параметра k, а, следовательно, и для минимального его значения

$$S_2(k) \le S_1(k)\,.$$

Таким образом, выбор прямой регрессии по минимуму суммы квадратов расстояний разумно и практически более обосновано, чем по классической схеме.

Соотношение (8) означает, что сумма квадратов отклонений различаются более существенно для статистических данных, которые сильно коррелированны, т.е. при $|r| \approx 1$ (в этом случае $|k| \to +\infty$).

Заметим, что из полученных значений $S_2(k_1)$, $S_2(k_2)$ следует выбрать минимальное и такой выбор решает задачу оптимизации регрессионной прямой по минимуму суммы квадратов расстояний.

Предлагаемый алгоритм может быть использован без всякого изменения в задачах оптимальной трассировки линейных участков водо- и газопроводов, автомобильных трасс, минимально удаленных от потребителей и населенных пунктов.

Изложенные выше идеи распространяются на многофакторный регрессионный анализ. Рассмотрим, для простоты, реализацию предлагаемого алгоритма применительно к двухфакторному случаю. Пусть переменная z зависит от двух признаков x, y, т.е.

$$z = \alpha + \beta x + \gamma y\,. \tag{9}$$

Неизвестные параметры α, β, γ определяются по следующему алгоритму.

Введем новые переменные

$$\tilde{x}_i = x_i - \bar{x}, \qquad \tilde{y}_i = y_i - \bar{y}, \qquad \tilde{z}_i = z_i - \bar{z}\,.$$

Тогда формула (9) перепишется в виде

$$\tilde{z}_i = \beta\tilde{x}_i + \gamma\tilde{y}_i\,, \tag{10}$$

$$\alpha = \bar{z} - \beta\bar{x} - \gamma\bar{y}\,, \tag{11}$$

$$\bar{x} = \left(\sum_{i=1}^{n} x_i\right) \Big/ n, \quad \bar{y} = \left(\sum_{i=1}^{n} y_i\right) \Big/ n, \quad \bar{z} = \left(\sum_{i=1}^{n} z_i\right) \Big/ n.$$

Коэффициенты β, γ по классической схеме находят из минимума суммы квадратов отклонений по оси Oy

$$S_1(\beta,\gamma) = \sum_{i=1}^{n} \left(\beta\tilde{x}_i + \gamma\tilde{y}_i - \tilde{z}_i\right)^2 \to \min\,.$$

Для минимизации функции $S(\beta,\gamma)$ необходимо решить систему уравнений

$$\frac{\partial S_1}{\partial\beta} = 0, \quad \frac{\partial S_1}{\partial\gamma} = 0\,. \tag{12}$$

Система (12) записывается в виде

$$\sum_{i=1}^{n} \left(\beta\tilde{x}_i + \gamma\tilde{y}_i - \tilde{z}_i\right)\tilde{x}_i = 0, \quad \sum_{i=1}^{n} \left(\beta\tilde{x}_i + \gamma\tilde{y}_i - \tilde{z}_i\right)\tilde{y}_i = 0\,. \tag{13}$$

Используя стандартные обозначения статистики, (13) перепишем ее в виде системы двух уравнений

$$\begin{cases} \beta\sigma_x^2 + \gamma\mu_{xy} = \mu_{xz}, \\ \beta\mu_{xy} + \gamma\sigma_y^2 = \mu_{yz}. \end{cases} \tag{14}$$

Решением системы (14), очевидно, являются

$$\beta = \frac{\mu_{xz}\sigma_y^2 - \mu_{xy}\mu_{yz}}{\sigma_x^2\sigma_y^2 - \mu_{xy}^2}\,, \qquad \gamma = \frac{\mu_{yz}\sigma_x^2 - \mu_{xy}\mu_{yz}}{\sigma_x^2\sigma_y^2 - \mu_{xy}^2}\,. \tag{15}$$

Формулы (15) дают оптимальные значения параметров β, γ. После их нахождения параметр α вычисляется по формуле (11). Изложенная методика есть классический двухфакторный регрессионный анализ.

Новый подход заключается в нахождении тех же коэффициентов но минимуму суммы квадратов расстояний, что приводит к минимизации функции

$$S_2(\beta,\gamma) = \sum_{i=1}^{n} \frac{\left(\beta\tilde{x}_i + \gamma\tilde{y}_i - \tilde{z}_i\right)^2}{1+\beta^2+\gamma^2}\,.$$

Система уравнений

$$\frac{\partial S_2}{\partial\beta} = 0, \quad \frac{\partial S_2}{\partial\gamma} = 0,$$

имеет вид

$$\begin{cases} \sum\limits_{i=1}^{n} \left[\tilde{x}_i\left(\beta\tilde{x}_i + \gamma\tilde{y}_i - \tilde{z}_i\right)\left(1+\beta^2+\gamma^2\right) - \beta\left(\beta\tilde{x}_i + \gamma\tilde{y}_i - \tilde{z}_i\right)^2\right] = 0, \\ \sum\limits_{i=1}^{n} \left[\tilde{y}_i\left(\beta\tilde{x}_i + \gamma\tilde{y}_i - \tilde{z}_i\right)\left(1+\beta^2+\gamma^2\right) - \gamma\left(\beta\tilde{x}_i + \gamma\tilde{y}_i - \tilde{z}_i\right)^2\right] = 0. \end{cases} \tag{16}$$

Используя общепринятые обозначения прикладной статистики, систему (16) можно записать в виде

$$\begin{cases} \left(1+\beta^2+\gamma^2\right)\left(\beta\sigma_x^2 + \gamma\mu_{xy} - \mu_{xz}\right) - \beta\left(\beta^2\sigma_x^2 + \gamma^2\sigma_y^2 + \right. \\ \qquad\qquad \left. + \sigma_z^2 - 2\beta\gamma\mu_{xy} + 2\beta\mu_{xz} + 2\gamma\mu_{yz}\right) = 0\,, \\ \left(1+\beta^2+\gamma^2\right)\left(\beta\mu_{xy} + \gamma\sigma_y^2 - \mu_{yz}\right) - \gamma\left(\beta^2\sigma_x^2 + \gamma^2\sigma_y^2 + \right. \\ \qquad\qquad \left. + \sigma_z^2 - 2\beta\gamma\mu_{xy} + 2\beta\mu_{xz} + 2\gamma\mu_{yz}\right) = 0\,. \end{cases} \tag{17}$$

Система (17) – нелинейная и может иметь несколько решений. Её можно переписать в виде

$$\begin{cases} \sum\limits_{i=1}^{n} a_i\tilde{x}_i = \dfrac{\beta}{1+\beta^2+\gamma^2}\sum\limits_{i=1}^{n} a_i^2\,, \\ \sum\limits_{i=1}^{n} a_i\tilde{y}_i = \dfrac{\gamma}{1+\beta^2+\gamma^2}\sum\limits_{i=1}^{n} a_i^2\,, \end{cases}$$

где α после нахождения параметров β, γ вычисляется но формуле

$$\alpha = \bar{z} - \beta\bar{x} - \gamma\bar{y}.$$

Число решений, как показывает простой анализ |3|, может быть конечным или бесконечным.

Для двухфакторного корреляционного анализа справедлив вывод, сделанный выше: выбор регрессионной зависимости но минимуму суммы квадратов расстояний даёт меньшую дисперсию, чем классический метод. Дисперсия но двум методам будет различаться более существенно для сильно коррелированных величин. Средняя сумма квадратов расстояний будет меньше в $1 + \beta^2 + \gamma^2$ раз, чем средняя сумма отклонений по оси Oz, т.е. дисперсии при классическом выборе |4|.

Литература

1. Бобров С.П. Экономическая статистика / М.-Л.: Государственное издательство, 1930. – 520 с.
2. Ошхунов М.М. // Изв. КБНЦ РАН. – 2001, №1 (4), С.63-67.
3. Ошхунов М.М. Введение в математическую статистику / Нальчик: КБГУ, 2004/ – 36 с.
4. Дубровский С.А. Прикладной многомерный статистический анализ / М.: Финансы и статистика, 1982. – 216 с.

NON-CLASSICAL VARIANT OF REGRESSIONS ANALYSIS

***M.M. Oshkhunov, **Z.M. Oshkhunova, *M.A. Dzhankulaeva**

*Kabardino-Balkarian State University named H.M. Berbekov,
Chernishevkogo St., 173, Nalchik, 360004, Russia, e-mail: muaed@inbox.ru, madina.dzhan@gmail.com;
**Kabardino-Balkarian department of OJSC Rostelecom,
Golovko St., 4, Nalchik, 360000, Russia, e-mail: zalina_oshhunova@mail.ru

Abstract. proposes The algorithm of finding regression functions finding is proposed. It is differed from the classical approach. The method consists of the choice of regression lines according to the minimum of the sum of squared distances form statistical points. The approach leads to non-uniqueness of the solution (in the opposite to the classical algorithm the lines choice) which has the geometric explanation. Examples are given of calculating the correlation coefficient for two methods which showed the higher efficiency of the method according to the degree of deviation from statistical data. It is made the attempt to extend the algorithm to the problem of multivariate regression analysis. It is provided some practical calculations of correlation parameters in relation to the specific statistical information.

Key words: regression analysis, method of least squares, non-classical method of least squares, dispersion.