УДК 519.233.5

Н.Н. Щелканов

НОВЫЙ МЕТОД НАХОЖДЕНИЯ КОЭФФИЦИЕНТОВ ЛИНЕЙНОЙ РЕГРЕССИИ МЕЖДУ ДВУМЯ ФИЗИЧЕСКИМИ ВЕЛИЧИНАМИ

Представлена обобщенная формула, позволяющая находить коэффициенты регрессии линейного уравнения $Y = K_0 + K_1 X$ для общего случая, когда разброс точек в корреляционной связи величин X и Y обусловлен как их случайными погрешностями измерений, так и неконтролируемыми физическими факторами. Все известные выражения для коэффициентов регрессии оказались частными случаями полученной формулы.

Ключевые слова: *линейная регрессия*, *случайные погрешности.*

При работе с разными массивами данных часто возникает необходимость нахождения коэффициентов линейной регрессии между двумя случайными физическими величинами. В большинстве случаев коэффициенты регрессии имеют конкретный физический смысл и для корректной интерпретации полученных результатов очень важно найти их значения наилучшим образом. Существует несколько формул для нахождения коэффициентов регрессии [1 – 3], но не для всех есть общее понимание, в каких случаях их следует использовать. В настоящее время отсутствует единый подход к нахождению коэффициентов линейной регрессии для общего случая, т.е. когда разброс точек в корреляционной связи между двумя величинами обусловлен как их случайными погрешностями измерений, так и неконтролируемыми физическими факторами.

Цель настоящей работы заключается в том, чтобы представить обобщенную формулу для вычисления коэффициентов линейной регрессии.

1. Постановка задачи

Рассмотрим две случайные физические величины X_0 и Y_0, между которыми существует статистическая корреляционная связь. Предположим, что эта связь может быть описана линейной зависимостью

$$Y_0 = K_0 + K_1 X_0, \tag{1}$$

и требуется найти коэффициенты регрессии K_0 и K_1, которые наилучшим образом отражают физическую взаимосвязь между ними.

Так как X_0 и Y_0 измеряются со случайными погрешностями, то на практике мы имеем дело с величинами X и Y, для которых уравнение регрессии запишется в виде

$$Y = K_0 + K_1 X. \tag{2}$$

Запись уравнений (1) и (2) с одинаковыми коэффициентами регрессии говорит о том, что последние не должны зависеть от случайных погрешностей измерен-

ных величин X и Y. В дальнейшем будем говорить о нахождении только коэффициента регрессии K_1, так как K_0 вычисляется после нахождения K_1 по известной формуле

$$K_0 = \overline{Y} - K_1 \cdot \overline{X}\,, \tag{3}$$

где $\overline{X}$ и $\overline{Y}$ – средние значения X и Y.

2. Новый подход

Новый подход к нахождению коэффициента регрессии K_1 заключается в следующих двух моментах:

1. Предлагается случайные величины X и Y нормировать соответственно на значения $\sqrt{\delta_X^2 + \delta_{X_0}^2}$ и $\sqrt{\delta_Y^2 + \delta_{Y_0}^2}$. Здесь δ_X и δ_Y – случайные среднеквадратические погрешности измерения X и Y для рассматриваемого массива данных; δ_{X_0} и δ_{Y_0} – некоторые величины, характеризующие разброс точек в корреляционной связи физических величин X_0 и Y_0 за счет неконтролируемых физических параметров.

2. При нахождении коэффициента регрессии K_1 используется ортогональная среднеквадратическая регрессия, т.е. минимизируется сумма квадратов отклонений, перпендикулярных искомой прямой.

Тогда уравнение линейной регрессии запишется в виде

$$\frac{Y}{\sqrt{\delta_Y^2 + \delta_{Y_0}^2}} = K_0{}' + K_1{}' \cdot \frac{X}{\sqrt{\delta_X^2 + \delta_{X_0}^2}}\,. \tag{4}$$

Здесь величины δ_{X_0} и δ_{Y_0} находятся из решения системы двух уравнений.

Первое уравнение имеет вид

$$\left|\rho_{X_0Y_0}\right| \cdot \sigma_{X_0} \cdot \sigma_{Y_0} = \sqrt{\sigma_{X_0}^2 - \delta_{X_0}^2} \cdot \sqrt{\sigma_{Y_0}^2 - \delta_{Y_0}^2}\,, \tag{5}$$

где $\sigma_{X_0} = \sqrt{\sigma_X^2 - \delta_X^2}$ и $\sigma_{Y_0} = \sqrt{\sigma_Y^2 - \delta_Y^2}$ – среднеквадратические отклонения величин X_0 и Y_0; σ_X и σ_Y – среднеквадратические отклонения величин X и Y; $\rho_{X_0Y_0}$ – коэффициент корреляции между X_0 и Y_0. Коэффициент корреляции $\rho_{X_0Y_0}$ находится из известного уравнения [1]:

$$\rho_{XY}\sigma_X\sigma_Y = \rho_{X_0Y_0}\sigma_{X_0}\sigma_{Y_0}, \tag{6}$$

где ρ_{XY} – коэффициент корреляции между X и Y. Заметим, что из уравнения (6) следует уравнение (5).

Второе уравнение запишем в виде

$$\frac{\delta_{X_0}}{\sigma_{X_0}} = \frac{\delta_{Y_0}}{\sigma_{Y_0}} \tag{7}$$

и назовем условием пропорциональности величин δ_{X_0}, δ_{Y_0} и σ_{X_0}, σ_{Y_0}. Введение величин δ_{X_0}, δ_{Y_0} и запись условия (7) являются ключевыми моментами в данной работе, так как это позволило получить обобщенное решение для коэффициентов линейной регрессии уравнения (2).

3. Результаты

После решения системы уравнений (5) и (7) получим

$$\delta_{X_0} = \sigma_X \cdot \sqrt{\left(1-\frac{\delta_X^2}{\sigma_X^2}\right)\cdot\left(1-\frac{|\rho_{XY}|}{\sqrt{\left(1-\delta_X^2/\sigma_X^2\right)\cdot\left(1-\delta_Y^2/\sigma_Y^2\right)}}\right)}, \tag{8}$$

$$\delta_{Y_0} = \sigma_Y \cdot \sqrt{\left(1-\frac{\delta_Y^2}{\sigma_Y^2}\right)\cdot\left(1-\frac{|\rho_{XY}|}{\sqrt{\left(1-\delta_X^2/\sigma_X^2\right)\cdot\left(1-\delta_Y^2/\sigma_Y^2\right)}}\right)}. \tag{9}$$

С учетом (8) и (9) найдем значения $\sqrt{\delta_X^2+\delta_{X_0}^2}$ и $\sqrt{\delta_Y^2+\delta_{Y_0}^2}$ в следующем виде:

$$\sqrt{\delta_X^2+\delta_{X_0}^2} = \sigma_X \cdot A\,; \tag{10}$$

$$\sqrt{\delta_Y^2+\delta_{Y_0}^2} = \sigma_Y \cdot B\,, \tag{11}$$

где

$$A = \sqrt{1-\left|\rho_{X_0Y_0}\right|\cdot\left(1-\frac{\delta_X^2}{\sigma_X^2}\right)} = \sqrt{1-|\rho_{XY}|\cdot\sqrt{\frac{1-\delta_X^2/\sigma_X^2}{1-\delta_Y^2/\sigma_Y^2}}}\,; \tag{12}$$

$$B = \sqrt{1-\left|\rho_{X_0Y_0}\right|\cdot\left(1-\frac{\delta_Y^2}{\sigma_Y^2}\right)} = \sqrt{1-|\rho_{XY}|\cdot\sqrt{\frac{1-\delta_Y^2/\sigma_Y^2}{1-\delta_X^2/\sigma_X^2}}}\,. \tag{13}$$

С учетом (10) и (11) уравнение линейной регрессии (4) запишется в виде

$$\frac{Y}{\sigma_Y \cdot B} = K_0' + K_1' \cdot \frac{X}{\sigma_X \cdot A}\,. \tag{14}$$

Уравнение (14) легко привести к виду (2):

$$Y = K_0' \cdot \sigma_Y \cdot B + K_1' \cdot \frac{\sigma_Y \cdot B}{\sigma_X \cdot A} \cdot X = K_0 + K_1 \cdot X\,, \tag{15}$$

где

$$K_0 = K_0' \cdot A \cdot \sigma_Y \cdot B\,; \tag{16}$$

$$K_1 = K_1' \cdot \frac{\sigma_Y \cdot B}{\sigma_X \cdot A}\,. \tag{17}$$

Применяя ортогональную среднеквадратическую регрессию к уравнению (14) и используя соотношение (17), получим выражение для искомого коэффициента регрессии:

$$K_1 = \frac{\sigma_Y}{\sigma_X} \cdot \frac{B}{A} \cdot \frac{1}{2\cdot\rho_{XY}} \cdot \left\{\left(\frac{A}{B}-\frac{B}{A}\right)+\sqrt{\left(\frac{A}{B}-\frac{B}{A}\right)^2+4\cdot\rho_{XY}^2}\right\}, \tag{18}$$

где A и B определяются выражениями (12) и (13). Впервые формула (18) была представлена в [4] и подробно описана в [5].

4. Анализ

Выражение (18) позволяет устанавливать однозначную связь между величинами X и Y и определять условия использования известных типов линейной регрессии.

Покажем, что все известные аналитические выражения для коэффициента регрессии K_1 уравнения (2) являются частными случаями формулы (18).

4.1. Так, для случая, когда разброс точек в корреляционной связи X и Y обусловлен только их случайными погрешностями, т.е. $\rho_{X_0Y_0} = 1$, получим известное выражение для коэффициента регрессии K_1, приведенное в [1]:

$$K_1 = \frac{\delta_Y}{\delta_X} \cdot \frac{1}{2 \cdot \rho_{XY}} \cdot \left\{ \left(\frac{\sigma_Y}{\sigma_X} \cdot \frac{\delta_X}{\delta_Y} - \frac{\sigma_X}{\sigma_Y} \cdot \frac{\delta_Y}{\delta_X} \right) + \sqrt{\left(\frac{\sigma_Y}{\sigma_X} \cdot \frac{\delta_X}{\delta_Y} - \frac{\sigma_X}{\sigma_Y} \cdot \frac{\delta_Y}{\delta_X} \right)^2 + 4 \cdot \rho_{XY}^2} \right\}. \quad (19)$$

4.1.1. При $\rho_{X_0Y_0} = 1$, $\delta_X = 0$ и $\delta_Y \neq 0$ имеем

$$K_1 = \lim_{\delta x \to 0} \frac{\delta_Y}{\delta_X} \cdot \frac{1}{2 \cdot \rho_{XY}} \cdot \left\{ \left(\frac{\sigma_Y}{\sigma_X} \cdot \frac{\delta_X}{\delta_Y} - \frac{\sigma_X}{\sigma_Y} \cdot \frac{\delta_Y}{\delta_X} \right) + \left(-\frac{\sigma_Y}{\sigma_X} \cdot \frac{\delta_X}{\delta_Y} + \frac{\sigma_X}{\sigma_Y} \cdot \frac{\delta_Y}{\delta_X} \right) \sqrt{1 + 4 \cdot \rho_{XY}^2 \cdot \left(\frac{\sigma_Y}{\sigma_X} \cdot \frac{\delta_X}{\delta_Y} \right)^2} \right\}.$$

Разлагая выражение под квадратным корнем в ряд Маклорена [6] и оставляя первые два члена, получим

$$K_1 = \lim_{\delta x \to 0} \frac{\delta_Y}{\delta_X} \cdot \frac{1}{2 \cdot \rho_{XY}} \left\{ \left(\frac{\sigma_Y}{\sigma_X} \cdot \frac{\delta_X}{\delta_Y} - \frac{\sigma_X}{\sigma_Y} \cdot \frac{\delta_Y}{\delta_X} \right) + \left(-\frac{\sigma_Y}{\sigma_X} \cdot \frac{\delta_X}{\delta_Y} + \frac{\sigma_X}{\sigma_Y} \cdot \frac{\delta_Y}{\delta_X} \right) \cdot \left[1 + 2 \cdot \rho_{XY}^2 \left(\frac{\sigma_Y}{\sigma_X} \cdot \frac{\delta_X}{\delta_Y} \right)^2 \right] \right\} = \frac{\sigma_Y}{\sigma_X} \cdot \rho_{XY}. \quad (20)$$

Это известная формула для коэффициента K_1 уравнения прямой регрессии $Y = K_0 + K_1 X$, которая находится путем минимизации суммы квадратов отклонений вдоль оси Y от искомой прямой [2].

4.1.2. При $\rho_{X_0Y_0} = 1$, $\delta_Y = 0$ и $\delta_X \neq 0$ имеем

$$K_1 = \lim_{\delta Y \to 0} \frac{\delta_Y}{\delta_X} \cdot \frac{1}{2 \cdot \rho_{XY}} \cdot \left\{ \left(\frac{\sigma_Y}{\sigma_X} \cdot \frac{\delta_X}{\delta_Y} - \frac{\sigma_X}{\sigma_Y} \cdot \frac{\delta_Y}{\delta_X} \right) + \left(\frac{\sigma_Y}{\sigma_X} \cdot \frac{\delta_X}{\delta_Y} - \frac{\sigma_X}{\sigma_Y} \cdot \frac{\delta_Y}{\delta_X} \right) \sqrt{1 + 4 \cdot \rho_{XY}^2 \cdot \left(\frac{\sigma_X}{\sigma_Y} \cdot \frac{\delta_Y}{\delta_X} \right)^2} \right\}.$$

Проведя процедуру разложения выражения под квадратным корнем в ряд Маклорена [6] и оставляя первые два члена, получим

$$K_1 = \lim_{\delta Y \to 0} \frac{\delta_Y}{\delta_X} \cdot \frac{1}{2 \cdot \rho_{XY}} \left\{ \left(\frac{\sigma_Y}{\sigma_X} \cdot \frac{\delta_X}{\delta_Y} - \frac{\sigma_X}{\sigma_Y} \cdot \frac{\delta_Y}{\delta_X} \right) + \left(\frac{\sigma_Y}{\sigma_X} \cdot \frac{\delta_X}{\delta_Y} - \frac{\sigma_X}{\sigma_Y} \cdot \frac{\delta_Y}{\delta_X} \right) \cdot \left[1 + 2 \cdot \rho_{XY}^2 \left(\frac{\sigma_X}{\sigma_Y} \cdot \frac{\delta_Y}{\delta_X} \right)^2 \right] \right\} = \frac{\sigma_Y}{\sigma_X} \cdot \frac{1}{\rho_{XY}}. \quad (21)$$

Формула (21) – также известная формула для коэффициента $1/K_1^*$ уравнения обратной регрессии $X = K_0^* + K_1^* \cdot Y$, которая получается путем минимизации суммы квадратов отклонений вдоль оси X от искомой прямой [2].

4.1.3. При $\rho_{X_0Y_0} = 1$ и $\delta_X = \delta_Y \neq 0$ получим известную формулу

$$K_1 = \frac{1}{2 \cdot \rho_{XY}} \cdot \left\{ \left(\frac{\sigma_Y}{\sigma_X} - \frac{\sigma_X}{\sigma_Y} \right) + \sqrt{\left(\frac{\sigma_Y}{\sigma_X} - \frac{\sigma_X}{\sigma_Y} \right)^2 + 4 \cdot \rho_{XY}^2} \right\} \tag{22}$$

для коэффициента K_1 уравнения ортогональной регрессии $Y = K_0 + K_1 X$, которая находится путем минимизации суммы квадратов отклонений, перпендикулярных искомой прямой [3].

4.2. Если для массива данных выполняется соотношение $\frac{\delta_X}{\sigma_X} = \frac{\delta_Y}{\sigma_Y}$, то из выражения (18) вытекает простая формула для коэффициента регрессии

$$K_1 = \frac{\sigma_Y}{\sigma_X}. \tag{23}$$

Так как соотношение $\frac{\delta_X}{\sigma_X} = \frac{\delta_Y}{\sigma_Y}$ выполняется для большинства экспериментальных данных, то формулу (23) можно рекомендовать к использованию при отсутствии информации о величинах случайных погрешностей X и Y. Заметим, что формула (23) представляет собой среднее геометрическое формул (20) и (21).

5. Диапазон изменчивости коэффициента регрессии

Для случая, когда разброс точек в корреляционной связи величин X и Y обусловлен только их случайными погрешностями, т.е. $\rho_{X_0 Y_0} = 1$, коэффициент регрессии будет изменяться в следующих пределах:

$$\frac{\sigma_Y}{\sigma_X} \cdot |\rho_{XY}| \le |K_1| \le \frac{\sigma_Y}{\sigma_X} \cdot \frac{1}{|\rho_{XY}|}, \tag{24}$$

а при $\rho_{X_0 Y_0} < 1$

$$\frac{\sigma_Y}{\sigma_X} \cdot |\rho_{XY}| < |K_1| < \frac{\sigma_Y}{\sigma_X} \cdot \frac{1}{|\rho_{XY}|}. \tag{25}$$

Как видно из выражений (24), (25), коэффициенты для прямой и обратной регрессий принимают соответственно минимальное и максимальное значения.

Заключение

Основные результаты:

1. Получена обобщенная формула, позволяющая находить коэффициенты регрессии линейного уравнения $Y = K_0 + K_1 X$ для общего случая, когда разброс точек в корреляционной связи случайных величин X и Y обусловлен как их случайными погрешностями измерений, так и неконтролируемыми физическими факторами.

2. Все известные выражения для коэффициентов регрессии являются частными случаями полученной формулы. Определены условия использования известных выражений.

Обобщенная формула позволяет получать устойчивые и физически-корректные коэффициенты регрессии. Формула представляет интерес для специалистов, занимающихся обработкой разных массивов данных, и может быть использована для их корректной физической интерпретации, независимо от области знания.

ЛИТЕРАТУРА

1. *Кендалл М.*, *Стьюарт А.* Статистические выводы и связи. М.: Наука, 1973. Т. 2. 900 с.
2. *Зайдель А.Н.* Погрешности измерений физических величин. Л.: Наука, 1985. 112 с.
3. *Крамер Г.* Математические методы статистики. М.: Мир, 1975. 648 с.
4. *Щелканов Н.Н.* Построение регрессионной зависимости между аэрозольными оптическими толщами атмосферы с учетом их случайных погрешностей // II заседание рабочей группы проекта «Аэрозоли Сибири»: тезисы докладов. Томск. Изд-во ИОА СО РАН, 1995. С. 16.
5. *Щелканов Н.Н.* Обобщенный метод построения линейной регрессии и его применение для построения однопараметрических моделей аэрозольного ослабления // Оптика атмосферы и океана. 2005. Т.18. № 1–2. С. 86–90.
6. *Кудрявцев В.А.*, *Демидович Б.П.* Краткий курс высшей математики. М.: Наука, 1975. 624 с.

Щелканов Николай Николаевич
Институт оптики атмосферы им. В.Е. Зуева СО РАН (г. Томск)
E-mail: snn@iao.ru

Поступила в редакцию 6 октября 2010 г.