Осень 2022 г.

Специальный практикум (V курс)

Метод взвешенной ортогональной регрессии

И. И. Никифоров

1. Регрессионный анализ

Регрессией (от лат. regressio «обратное движение») называют зависимость среднего значения какой-либо величины от некоторой другой величины или от нескольких величин. В отличии от функциональной зависимости $y = f(x)$, когда каждому значению независимой переменной x соответствует одно определенное значение величины y, при *регрессионной связи* одному и тому же значению x могут соответствовать в зависимости от случая различные значения величины y.

Если уравнение

$$y = y(x), \tag{1}$$

описывает регрессионную связь, то оно называется *уравнением регрессии*, соответствующий график — *линией регрессии*, а переменная x — *регрессионной переменной* или *регрессором*.

Наиболее важное значение имеет *полиномиальная (параболическая) регрессия*:

$$y = a_0 + a_1 x + \ \ldots \ + a_j x^j + \ \ldots \ + a_n x^n \tag{2}$$

и ее частный случай — *линейная регрессия*

$$y = a_0 + a_1 x = a + bx, \tag{3}$$

которым мы и будем заниматься на практикуме.

Методы исследования регрессионной зависимости называются *регрессионным анализом*. Одной из главных его задач является нахождение уравнения регрессии, т.е. в случаях (2) и (3) — определение коэффициентов a_j по некоторому количеству пар измерений

$$(x_i, y_i), \quad i = 1, \ \ldots, \ N. \tag{4}$$

Найденные a_j называются *выборочными коэффициентами регрессии*, а линия регрессии — *эмпирической линии регрессии*.

Наиболее просто последняя задача решается, если величины x_i известны точно — они могут быть заданы (выбраны). Тогда отклонения точек (x_i, y_i) от регрессии обусловлено только ошибками измерений и/или влиянием на среднее значение y других факторов (регрессоров). В этом случае составляется избыточная система уравнений

$$y_i = a + bx_i, \quad i = 1, \ \ldots \ , N, \tag{5}$$

которая решается обычным методом наименьших квадратов (МНК). Последний изучался в курсе «Теория вероятностей и математическая обработка наблюдений». Такое

решение задачи оптимально в смысле, что оно дает регрессию, минимизирующую сумму квадратов отклонений

$$\min \sum_{i=1}^{N} [y_i - y(x_i)]^2. \tag{6}$$

Такая регрессия называется *средней квадратической*.

Однако в реальности часто бывает так, что величины x_i также содержат ошибки измерений. Тогда формальное применение к (5) обычного МНК в общем случае дает искаженные (систематически смещенные) выборочные коэффициенты регрессии a и b и смещенную линию регрессии.

Как же найти более корректную регрессию в этом случае?

Иногда для грубой прикидки обычным МНК по одним и тем же данным определяют две формальные средние квадратические регрессии

$$y = a + bx, \tag{7}$$

$$x = a' + b'y, \tag{8}$$

а затем находят для них биссектрису и ее принимают за решение. Но не во всех случаях этот вообще улучшает эмпирическую регрессию (при «взвешивании» истинная регрессия может и не лежать между этими двумя средними квадратическими). Мы не будем пользоваться этим приемом.

2. Ортогональная регрессия

В случае, когда ошибки измерений присутствуют для обеих координат, более строгим является обобщение обычного МНК на два измерения — т.е. поиск такой регрессии, которая минимизирует сумму квадратов перпендикулярных расстояний точек (x_i, y_i) от линии регрессии

$$\min \sum_{i=1}^{N} \rho_i^2, \tag{9}$$

где

$$\rho_i = \frac{y_i - a - bx_i}{\sqrt{1 + b^2}}, \qquad \rho \in (-\infty, +\infty). \tag{10}$$

Здесь a и b — коэффициенты регрессии

$$y = a + bx. \tag{11}$$

Такая регрессия называется *ортогональной*. Она была предложена R. J. Adcock (1878).

Для линейной регрессии (11) получено аналитическое решение при оптимизации (9) и более общей оптимизации, в которой учитывается различие средних ошибок x_i и y_i. Однако формулы существенно разные в случаях: 1) одинаковых дисперсий всех измерений x_i и одинаковых дисперсий всех измерений y_i, 2) в общем случае произвольных весов измерений x_i и y_i. В первом случае можно в явном виде выписать выражения для выборочных коэффициентов и для их дисперсий. Во втором — аналитические выражения являются лишь приближенными и требуется применять их итеративно для получения окончательного результата. Рассмотрим общий случай, который называется *взвешенной ортогональной регрессией*. Формулы для нее приведены в работе D. York (Canadian Journal of Physics, 1966, 44, 1079–1086).

Взвешенная ортогональная регрессия. Пусть (X_i, Y_i), $i = 1, \ldots, N$, — измерения (наблюдения), $\omega(X_i)$, $\omega(Y_i)$ — соответствующие веса; если известны средние ошибки $\sigma(X_i)$, $\sigma(Y_i)$, то $\omega(X_i) = 1/\sigma^2(X_i)$, $\omega(Y_i) = 1/\sigma^2(Y_i)$. Предполагается, что величины X_i и Y_i не коррелируют друг с другом. Требуется определить коэффициенты a и b линейной регрессии $y = a + bx$, для которой сумма взвешенных квадратов расстояний от точек (X_i, Y_i) до их (неортогональных) проекций на линию регрессии (x_i, y_i) минимальна:

$$\min \sum_{i=1}^{N} \left[\omega(X_i)(x_i - X_i)^2 + \omega(Y_i)(y_i - Y_i)^2\right]. \tag{12}$$

Для получения решения на каждой (k-ой) итерации для коэффициента b требуется начальное приближение $b_{(k)}$, $k = 0, 1, 2, \ldots$ После задания $b_{(k)}$ для каждой пары измерений $\sigma(X_i)$, $\sigma(Y_i)$ находится ее вес

$$W_i = \frac{\omega(X_i)\omega(Y_i)}{b_{(k)}^2\omega(Y_i) + \omega(X_i)} = [b_{(k)}^2\sigma^2(X_i) + \sigma^2(Y_i)], \qquad i = 1, \ldots, N. \tag{13}$$

Затем определяется «гравитационный центр» (средневзвешенная точка) для всей совокупности данных,

$$\langle X\rangle = \frac{\sum\limits_{i=1}^{N} W_i X_i}{\sum\limits_{i=1}^{N} W_i}, \quad \langle Y\rangle = \frac{\sum\limits_{i=1}^{N} W_i Y_i}{\sum\limits_{i=1}^{N} W_i}, \tag{14}$$

и массивы отклонений отдельных измерений от средневзвешенных величин

$$U_i = X_i - \langle X\rangle, \quad V_i = Y_i - \langle Y\rangle, \qquad i = 1, \ldots, N. \tag{15}$$

Затем определяются коэффициенты

$$\alpha = \frac{2\sum\limits_{i=1}^{N} \frac{W_i^2}{\omega(X_i)} U_i V_i}{3\sum\limits_{i=1}^{N} \frac{W_i^2 U_i^2}{\omega(X_i)}}, \quad \beta = \frac{\sum\limits_{i=1}^{N} \frac{W_i^2 V_i^2}{\omega(X_i)} - \sum\limits_{i=1}^{N} W_i U_i^2}{3\sum\limits_{i=1}^{N} \frac{W_i^2 U_i^2}{\omega(X_i)}}, \quad \gamma = -\frac{\sum\limits_{i=1}^{N} W_i U_i V_i}{\sum\limits_{i=1}^{N} \frac{W_i^2 U_i^2}{\omega(X_i)}} \tag{16}$$

кубического уравнения для b

$$b^3 - 3\alpha b^2 + 3\beta b - \gamma = 0. \tag{17}$$

Корни уравнения (17) определяются выражениями

$$b_{j+1} = \alpha + 2\sqrt{\alpha^2 - \beta}\cos\frac{\varphi + 2\pi j}{3}, \quad \cos\varphi = \frac{\alpha^3 - \frac{3}{2}\alpha\beta + \frac{1}{2}\gamma}{(\alpha^2 - \beta)^{3/2}}, \quad j = 0, 1, 2. \tag{18}$$

Решение задачи дает корень b_3 ($j = 2$). Он задает новое приближение для $(k+1)$-ой итерации:

$$b_{(k+1)} = b_3, \qquad a_{(k+1)} = \langle Y\rangle - b_{(k+1)}\langle X\rangle. \tag{19}$$

Итерации продолжаются до достижения нужной точности результата.

Для окончательных оценок a и b находятся их средние ошибки:

$$x_i - X_i = -b\,\frac{W_i}{\omega(X_i)}\,(a + bX_i - Y_i), \qquad y_i - Y_i = \frac{W_i}{\omega(Y_i)}\,(a + bX_i - Y_i); \tag{20}$$

$$\sigma_b^2 = \frac{1}{N-2}\,\frac{\sum\limits_{i=1}^{N} W_i(bU_i - V_i)^2}{\sum\limits_{i=1}^{N} W_i U_i^2}, \qquad \sigma_a^2 = \sigma_b^2\,\frac{\sum\limits_{i=1}^{N} W_i X_i^2}{\sum\limits_{i=1}^{N} W_i}. \tag{21}$$

Примечание. Для получения очередного приближения b можно вместо кубического уравнения использовать линейное (D. York, Earth and Planetary Science Letters, 1969, 5, 320–324):

$$b_{(k+1)} = \frac{\sum\limits_{i=1}^{N} W_i^2 V_i \left(\dfrac{U_i}{\omega(Y_i)} + \dfrac{b_{(k)} V_i}{\omega(X_i)}\right)}{\sum\limits_{i=1}^{N} W_i^2 U_i \left(\dfrac{U_i}{\omega(Y_i)} + \dfrac{b_{(k)} V_i}{\omega(X_i)}\right)}. \tag{22}$$

Проверить, так ли это!

3. Данные

X_i	0.0	0.9	1.8	2.6	3.3	4.4	5.2	6.1	6.5	7.4
$\omega(X_i)$	1000	1000	500	800	200	80	60	20	1.8	1.0
Y_i	5.9	5.4	4.4	4.6	3.5	3.7	2.8	2.8	2.4	1.5
$\omega(Y_i)$	1.0	1.8	4.0	8.0	20	20	70	70	100	500

4. Решение методом Йорка

Получить результаты при помощи кубического и линейного уравнений для данных, приведенных выше в таблице. Привести начальное приближение b_0, таблицу зависимости b и a от номера итерации, рисунок (X, Y), на котором привести точки (X_i, Y_i) с барами ошибок $\sigma(X_i) = 1/\sqrt{\omega(X_i)}$, $\sigma(Y_i) = 1/\sqrt{\omega(Y_i)}$, полученную регрессию и проекции точек (X_i, Y_i) на линейную модель.

5. Решение способом численной минимизации целевой функции

Минимизацию целевой функции в (12), дающую решение задачи, можно интерпретировать как минимизацию непостоянной части логарифмической функции правдоподобия:

$$\min \mathcal{L}^{(1)}(a, b), \tag{23}$$

$$\mathcal{L}^{(1)}(a, b) = \frac{1}{2}\sum_{i=1}^{N} \min_{x_i} \left[\omega(Y_i)(a + bx_i - Y_i)^2 + \omega(X_i)(x_i - X_i)^2\right]. \tag{24}$$

Действительно

$$\frac{1}{2}\min\sum_{i=1}^{N}\left[\omega(X_i)(x_i-X_i)^2+\omega(Y_i)(y_i-Y_i)^2\right]=$$

$$\min\frac{1}{2}\sum_{i=1}^{N}\left[\omega(Y_i)(a+bx_i-Y_i)^2+\omega(X_i)(x_i-X_i)^2\right]=$$

$$=\min_{a,b}\frac{1}{2}\sum_{i=1}^{N}\min_{x_i}\left[\omega(Y_i)(a+bx_i-Y_i)^2+\omega(X_i)(x_i-X_i)^2\right]= \tag{25}$$

$$=\min\mathcal{L}^{(1)}(a,b).$$

Для тех же данных найти решение — параметры a и b — при помощи численной минимизации (23) целевой функции (24), т.е. при помощи минимизации (25). Определить доверительные интервалы (величины ошибок) для оценок параметров способом, известным для метода наибольшего правдоподобия.

Сравнить результаты применения двух методов и их эффективность (трудовые затраты на программирование, скорость работы программ).