УДК 550.34:519.246

А.М.МУРАЛИЕВ, Б.С.АЛЫМБАЕВА

E-mail: ksucta@elcat.kg.

ЛИНЕЙНАЯ АППРОКСИМАЦИЯ ЭМПИРИЧЕСКИХ ДАННЫХ МЕТОДОМ НАИМЕНЬШИХ КВАДРАТОВ

Бул статьяда эң кичине квадраттар методу менен тажрыйбалык маалыматтарды сызыктуу функция менен алмаштыруу каралды.

В статье рассматривается научный вопрос, связанный с подбором линии при аппроксимации эмпирических данных. Вычисление параметров регрессионной модели используется для изучения графика повторяемости землетрясений в одной из сейсмоактивных областей Кыргызстана. Оценка наклона графика повторяемости и его уровень играют большую роль при построении математической модели сейсмического процесса.

In this article examining scientific question; connecting with selection lines in the presence of approcsimation empirical facts. Calculating parameters regret ion model uses for study graphic repeating earthquake in seismoactive province in Kyrgyzstan. Estimation of inclination in graphic repenting and level its plays lit role in learning role in learning the character of seismic process.

Уравнение прямой может быть полезно во многих ситуациях для обобщения наблюдаемой зависимости одной переменной от другой. Мы покажем, как такое уравнение можно получить методом наименьших квадратов, когда имеются данные наблюдений. Обозначим через Y количество землетрясений в логарифмическом масштабе ($\lg N$); X – энергетический класс землетрясения (K). Выборка сейсмических событий произведена в одной из сейсмоактивных областей на территории Кыргызстана.

Предположим, что линия регрессии переменной, которую мы обозначим Y, от переменной (X) имеет вид $\beta_0 + \beta_1 X$ /1/. Тогда можно записать линейную модель:

$$Y = \beta_0 + \beta_1 X + \varepsilon, \quad (1)$$

так что для данного X соответствующее значение Y состоит из величины $\beta_0 + \beta_1 X$ плюс добавка ε, при учете которой любой индивидуальный Y получает возможность не попасть на линию регрессии.

Уравнение (1) – это модель, которую мы предлагаем. Предположение о математической модели процесса необходимо с многих статистических точек зрения. Следует отметить, что то, что мы обычно делаем, есть постулирование модели либо предварительное допущение о ее правильности. Величины β_0 и β_1 называют параметрами модели. Например,

$$Y = \beta_0 + \beta_1 X + \beta_{11} X^2 + \varepsilon$$

есть регрессионная модель второго порядка (по X) и линейная (по β). Если только специально не оговаривается, что модель нелинейна, а это может быть сделано, то имеется в виду линейная по параметрам модель, а слово «линейная» обычно опускается. Порядок модели может быть любым. Обозначение вида β_{11} часто используется в полиномиальных моделях, где параметр β_1 соотносится с X, в то время как β_{11} соотносится с $X^2 = X \cdot X$.

Итак, в уравнении (1) величины β_0 , β_1 и ε неизвестны, причем величину ε на самом деле будет трудно исследовать, поскольку она меняется от наблюдения к наблюдению. Однако β_0 и β_1 остаются постоянными, и, хотя мы не умеем находить их точно без изучения всех возможных сочетаний Y и X, мы используем информацию, содержащуюся в табл.1, для получения оценок b_0 и b_1 параметров β_0 и β_1. Запишем это в таком виде:

$$\hat{Y} = b_0 + b_1 X,$$

(2)

где $\hat{Y}$ обозначает предсказанное значение Y для данного X, когда b_0 и b_1 определены. Уравнение (2) можно использовать как предсказывающее уравнение; подстановка в него значения X позволяет предсказать «истинное» среднее значение Y для этого X.

Общепринято обозначение оценок параметров маленькими латинскими буквами, а самих параметров – греческими: b_0 и b_1 и β_0 и β_1 соответственно. Нашей процедурой оценивания будет метод наименьших квадратов.

Пусть мы имеем множество из n наблюдений $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$. Тогда уравнение (1) можно записать в виде

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,$$

(3)

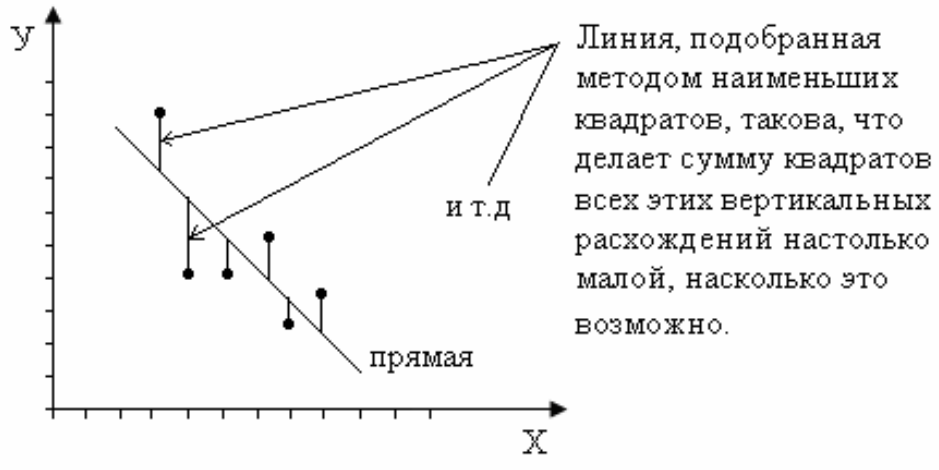

Рис.1. Вертикальные отклонения, минимицирующие сумму
квадратов в методе наименьших квадратов.

где $i = 1,2,\ldots,n$. Следовательно, сумма квадратов отклонений от «истинной» линии есть

$$\mathrm{S} = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 \quad .$$

(4)

Будем подбирать значения оценок b_0 и b_1 так, чтобы их подстановка вместо β_0 и β_1 в уравнение (4) давала наименьшее возможное (минимальное) значение S (рис. 1). Заметим, что X_i, Y_i – это фиксированные числа, которые нам известны. Мы можем определить b_0 и b_1, дифференцируя уравнение (4) сначала по β_0, затем по β_1 и приравнивая результаты к нулю. Тогда

$$\frac{\partial S}{\partial \beta_0} = -2\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i),$$
$$\frac{\partial S}{\partial \beta_1} = -2\sum_{i=1}^{n} X_i (Y_i - \beta_0 - \beta_1 X_i),$$

(5)

так что для оценок b_0 и b_1 имеем

$$\sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i) = 0,$$
$$\sum_{i=1}^{n} X_i (Y_i - b_0 - b_1 X_i) = 0,$$

(6)

где при приравнивании выражений (5) к нулю мы подставили (b_0 , b_1) вместо (β_0 , β_1).

Из (6) имеем:

$$\sum_{i=1}^{n} Y_i - n b_0 - b_1 \sum_{i=1}^{n} X_i = 0,$$
$$\sum_{i=1}^{n} X_i Y_i - b_0 \sum_{i=1}^{n} X_i - b_1 \sum_{i=1}^{n} X_i^2 = 0$$

(7)

или

$$b_0 n + b_1 \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} Y_i ,$$
$$b_0 \sum_{i=1}^{n} X_i + b_1 \sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} X_i Y_i$$

(8)

Эти уравнения называют нормальными.

Решение уравнений (8) относительно угла наклона прямой - b_1 дает

$$b_1 = \frac{\sum X_i Y_i - \left[\left(\sum X_i\right)\left(\sum Y_i\right)\right]/n}{\sum X_i^2 - \left(\sum X_i\right)^2/n} = \frac{\sum \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum \left(X_i - \overline{X}\right)^2},$$

(9)

где суммирование всегда ведется от $i = 1$ до n, а два выражения для b_1 – это обе правильные, но несколько различные формы одной и той же величины. Так как по определению

$$\overline{X} = \left(X_1 + X_2 + \cdots + X_n\right)/n = \sum X_i / n,$$
$$\overline{Y} = \left(Y_1 + Y_2 + \cdots + Y_n\right)/n = \sum Y_i / n,$$

имеем:

$$\sum (X_i - \overline{X})(Y_i - \overline{Y}) = \sum X_i Y_i - \overline{X}\sum Y_i - \overline{Y}\sum X_i + n\,\overline{X}\,\overline{Y} =$$
$$= \sum X_i Y_i - n\,\overline{X}\,\overline{Y} = \sum X_i Y_i - \frac{\left(\sum X_i\right)\left(\sum Y_i\right)}{n}.$$

Отсюда следует эквивалентность числителей в (9), а заодно, при замене Y на X, эквивалентность знаменателей. Величина $\sum X_i^2$ называется некорректированной суммой квадратов X-в, а $\frac{(\sum X_i)^2}{n}$ – коррекцией на среднее значение X-в. Разность между ними называется скорректированной суммой квадратов X-ов. Аналогично $\sum X_i Y_i$ называется некорректированной суммой смешанных (парных) произведений, а $\frac{\left(\sum X_i\right)\left(\sum Y_i\right)}{n}$ – коррекцией на среднее. Разность между ними называется скорректированной суммой произведений X и Y.

Первая форма уравнения (9) обычно используется для вычисления b_1 на калькуляторе, поскольку с ней гораздо легче работать и нет нужды в громоздких подсчетах для каждого X_i и Y_i выражений $(X_i - \overline{X})$ и $(Y_i - \overline{Y})$ соответственно. Полезно иметь в виду, что для уменьшения ошибок округления лучше всего сохранять в процессе счета столько знаков после запятой, сколько возможно.

Здесь и далее возьмем удобные обозначения и запишем:

$$S_{XY} = \sum (X_i - \overline{X})(Y_i - \overline{Y}) = \sum (X_i - \overline{X})Y_i = \sum X_i\,(Y_i - \overline{Y}) =$$
$$= \sum X_i Y_i - \frac{\left(\sum X_i\right)\left(\sum Y_i\right)}{n} = \sum X_i Y_i - n\,\overline{X}\,\overline{Y}.$$

Заметим, что все эти выражения эквивалентны. Аналогично можно записать:

$$S_{XX} = \sum (X_i - \overline{X})^2 = \sum (X_i - \overline{X}) X_i = \sum X_i^2 - \frac{\left(\sum X_i\right)^2}{n} = \sum X_i^2 - n\overline{X}^2;$$

$$S_{YY} = \sum (Y_i - \overline{Y})^2 = \sum (Y_i - \overline{Y}) Y_i = \sum Y_i^2 - \frac{\left(\sum Y_i\right)^2}{n} = \sum Y_i^2 - n\overline{Y}^2.$$

Легко запоминающуюся формулу для b_1 запишем в следующем виде:

$$b_1 = \frac{S_{XY}}{S_{XX}}.$$

(9а)

Решение уравнения (8) относительно свободного члена (отрезка на оси ординат при X=0) b_0 дает

$$b_0 = \overline{Y} - b_n \overline{X}.$$

(10)

С помощью подстановки уравнения (10) в уравнение (2) можно получить оцениваемое уравнение регрессии:

$$\hat{Y}_i = \overline{Y} + b_1 (X_i - \overline{X}), \qquad (11)$$

где b_1 определяется уравнением (9).

Отметим, что если в (11) положить $X_i = \overline{X}$, то окажется, что $\hat{Y}_i = \overline{Y}$. А это означает, что точка $(\overline{Y}, \overline{X})$ лежит на подобранной прямой. Выполним теперь эти вычисления для изучения распределения землетрясений по энергии в одной из сейсмоактивной области Кыргызстана.

Выборка произведена в одной из сейсмоактивных областей Кыргызстана.

Таблица 1

Десять наблюдений переменных X и Y [2]

Номер наблюдения	Номер переменной		Номер наблюдения	Номер переменной	
	Y (lgN)	X (K)		Y (lgN)	X (K)
1	1570 (3,1959)	8	6	18 (1,2553)	13
2	949 (2,9773)	9	7	13 (1,1139)	14
3	559 (2,7474)	10	8	5 (0,6990)	15
4	219 (2,3404)	11	9	5 (0,6990)	16
5	95 (1,9777)	12	10	1 (0.0000)	17

Отметим, что если в (11) положить $X_i = \overline{X}$, то окажется, что $\hat{Y}_i = \overline{Y}$. А это означает, что точка $(\overline{Y}, \overline{X})$ лежит на подобранной прямой. Выполним теперь эти вычисления, пользуясь данными табл.1.

$$n = 10$$

$$\sum_{i=1}^{7} Y_i = 3{,}20 + 2{,}98 + \cdots + 0{,}70 = 17{,}01$$

$$\sum_{i=1}^{7} X_i = 8 + 9 + 10 + \cdots + 17 = 125$$

$$\sum_{i=1}^{7} X_i Y_i = 3{,}20 \cdot 8 + 2{,}98 \cdot 9 + \cdots + 1 \cdot 17 = 182{,}91$$

$$\sum_{i=1}^{7} X_i^2 = 8^2 + 9^2 + 10^2 + \cdots + 17^2 = 1645$$

$$b_1 = \frac{\sum X_i Y_i - \left[\left(\sum X_i\right)\left(\sum Y_i\right)\right]/n}{\sum X_i^2 - \left(\sum X_i\right)^2/n} = \frac{182{,}91 - 212{,}62}{1645 - 1562{,}5} = \frac{-29{,}71}{82{,}5} = -0{,}3601 \approx -0{,}4\,.$$

По этому подобранное уравнение есть $\hat{Y}_i = \overline{Y} + b_1(X_i - \overline{X})$,

$$\hat{Y}_i = 1{,}71 - 0{,}3601(X_i - 1{,}25)$$

$$\hat{Y}_i = 6{,}2112 - 0{,}3601X_i$$

Построенная линия регрессии нанесена на рис.2.

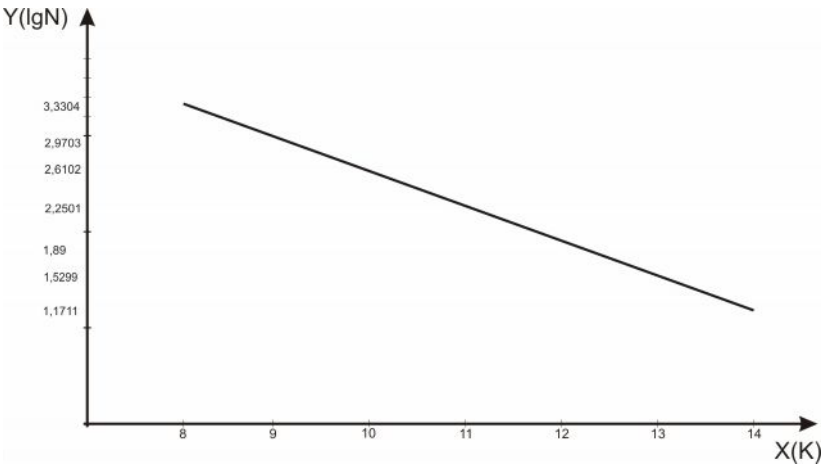

Рис. 2

СПИСОК ЛИТЕРАТУРЫ

1. Себер Дж. Линейный регрессионный анализ / Пер. с англ. под ред. М.Б.Малютова. – М.: Мир, 1980. – 456 с.

2. Муралиев А.М. Сейсмичность и сейсмотектоническая деформация территории Юго- Западной Киргизии и сопредельных областей. – Фрунзе: Илим, 1989. – 106 с.