Метод наименьших квадратов. Уравнения линейной и квадратичной регрессий. Построение линейных систем для определения коэффициентов регрессии.

Регрессионный анализ – наиболее распространенный метод обработки данных, который включает в себя метод наименьших квадратов. При регрессионном анализе таблица экспериментальных данных обычно отражается алгебраическими степенными полиномами, которые называют полиномами или уравнениями регрессии. Отсюда термины – задача регрессии, коэффициенты регрессии и т.п. Сам термин регрессия отражает тот факт, что с увеличением степени полинома точность отражения таблицы экспериментальных данных обычно возрастает, а ошибка отражения соответственно уменьшается, регрессирует.

Рассмотрим один из методов, позволяющих проанализировать и обработать данные, полученные в результате эксперимента. Пусть в результате измерений получена таблица зависимости одной величины y от другой x.

Пусть зависимость между двумя переменными x и y выражается в виде таблицы, полученной опытным путем. Это могут быть результаты опыта или наблюдений, статистической обработки материала и т.п.

Таблица 1

x	x_1	x_2	...	x_i	...	x_n
f(x)	y_1	y_2	...	y_i	...	y_n

Требуется наилучшим образом сгладить экспериментальную зависимость между переменными x и y, т.е. по возможности точно отразить общую тенденцию зависимости y от x, исключив при этом случайные отклонения, связанные с неизбежными погрешностями измерений или статистических наблюдений. Такую сглаженную зависимость стремятся представить в виде формулы $y = f(x)$.

Необходимо найти формулу $y = f(x)$, выражающую таблично заданную зависимость аналитически. Применение интерполяции в данном случае нецелесообразно, т.к. значения y_i в узлах получены экспериментально и поэтому являются сомнительными (в ходе эксперимента возникает неустранимая погрешность, обусловленная неточностью измерений). Кроме того, совпадение значений в узлах не означает совпадения характеров поведения исходной и интерполирующей функции. Поэтому необходимо найти такой метод подбора эмпирической формулы, который не только позволяет найти саму формулу, но и оценить погрешность подгонки.

Постановка задачи. Найдем функцию заданного вида $y = f(x)$ которая в точках $x_1, x_2, x_3, ..., x_n$ принимает значения как можно более близкие к табличным значениям $y_1, y_2, y_3, ..., y_n$. Практически вид приближающей функции можно определить визуально: по таблице 1 строится точечный график функции, а затем проводится кривая, по возможности наилучшим образом отражающая характер расположения точек (рис.1).

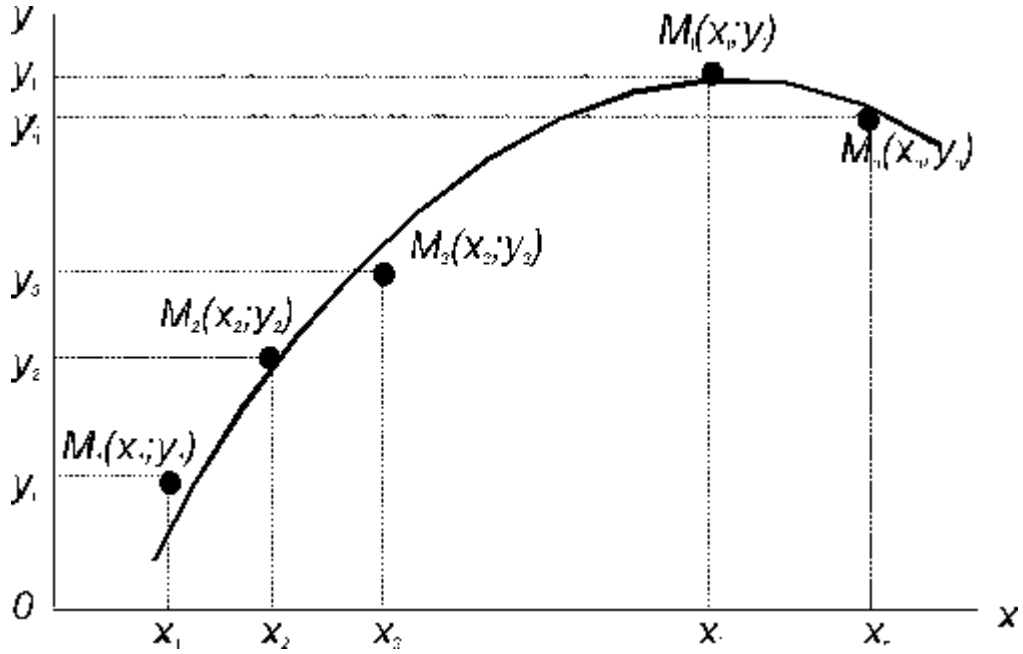


Рис. 1

Формулы, служащие для аналитического представления опытных данных, получили название *эмпирических формул*.

Задача нахождения эмпирических формул разбивается на два этапа. На первом этапе нужно установить **вид зависимости** $y = f(x)$, т.е. решить, является ли она линейной, квадратичной, логарифмической или какой-либо другой.

Предположим, например, что результаты экспериментальных исследований нанесены на плоскость, т.е. паре чисел (x, y) соответствует точка на плоскости с такими же координатами. Разумеется, существует множество кривых, проходящих через эти точки.

Обычно предполагают, что кривая истинной зависимости - это наиболее «гладкая» кривая, согласованная с эмпирическими данными.

Для проверки правильности вывода проводятся дополнительные исследования, т.е. производится еще ряд одновременных измерений величин *x* и *y*. Дополнительные точки наносятся на плоскость. Если они оказываются достаточно близкими к выбранной кривой, то можно считать, что вид кривой установлен. В противном случае, кривую надо скорректировать и вновь провести дополнительные измерения.

Кроме того, для выбора функции $y = f(x)$ привлекаются дополнительные соображения, как правило, не математического характера (теоретические предпосылки, опыт предшествующих исследований и т.п.).

Предположим, первый этап завершен – вид функции $y = f(x)$ установлен. Тогда переходят ко второму этапу – **определению неизвестных параметров этой функции**.

Согласно наиболее распространенному и теоретически обоснованному методу наименьших квадратов в качестве неизвестных параметров функции $f(x)$ выбирают такие значения, чтобы сумма квадратов невязок δ_i, или отклонений *«теоретических» значений* $f(x_i)$, найденных по эмпирической формуле $y = f(x)$, от соответствующих *опытных значений* y_i, т.е.

$$S = \sum_{i=1}^{n} \delta_i^2 = \sum_{i=1}^{n} (f(x_i) - y_i)^2$$

была минимальной.

В качестве величины отклонения S эмпирических точек (x_i, y_i) от точек сглаживающей экспериментальную зависимость кривой $y = f(x)$ в принципе можно было взять обычную сумму невязок $\sum_{i=1}^{n}\delta_i = \sum_{i=1}^{n}(f(x_i)-y_i)$ или сумму их абсолютных величин $\sum_{i=1}^{n}|\delta_i| = \sum_{i=1}^{n}|(f(x_i)-y_i)|$. Но делать это нецелесообразно, так как в первом случае $\sum_{i=1}^{n}\delta_i$ может быть малой или даже равняться нулю при значительном разбросе эмпирических точек, так как положительные отклонения δ_i компенсируются отрицательными.

Во втором случае функция $\sum_{i=1}^{n}|\delta_i|$ лишена этого недостатка, но имеет другой – она не является дифференцируемой, что существенно затрудняет решение задачи.

Пусть в качестве функции $y = f(x)$ взята *линейная функция* $y = a + bx$ и задача сводится к отысканию таких значений параметров a и b, при которых функция

$$S = \sum_{i=1}^{n}(ax_i + b - y_i)^2$$

принимает наименьшее значение.

Заметим, что функция $S = S(a; b)$ есть функция двух *переменных* a и b до тех пор, пока не найдены, а затем не зафиксированы их «наилучшие» (в смысле метода наименьших квадратов) значения, а x_i, y_i – *постоянные* числа, найденные экспериментально.

Таким образом, для нахождения прямой, наилучшим образом согласованной с опытными данными, достаточно решить систему

$$\begin{cases} S'_a = 0, \\ S'_b = 0, \end{cases} \quad \text{или} \quad \begin{cases} \sum_{i=1}^{n} 2(ax_i + b - y_i)x_i = 0, \\ \sum_{i=1}^{n} 2(ax_i + b - y_i) = 0. \end{cases}$$

После алгебраических преобразований эта система принимает вид:

$$\begin{cases} (\sum_{i=1}^{n} x_i^2)a + (\sum_{i=1}^{n} x_i)b = \sum_{i=1}^{n} x_i y_i, \\ (\sum_{i=1}^{n} x_i)a + nb = \sum_{i=1}^{n} y_i. \end{cases}$$

Эта система называется *системой нормальных уравнений*. Она имеет единственное решение, так как ее определитель

$$|A| = \begin{vmatrix} \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & n \end{vmatrix} = n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2 \neq 0$$

(а точнее $|A| > 0$, что можно доказать методом математической индукции при $n \geq 2$).

Убедимся, что найденные из системы нормальных уравнений значения дают минимум функции $S = S(a; b)$. Найдем частные производные

$$S''_{aa} = 2\sum_{i=1}^{n} x_i^2 = A; \quad S''_{ab} = 2\sum_{i=1}^{n} x_i = B; \quad S''_{bb} = 2n = C.$$

Выражение $\Delta AB - C^2 = 4(n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2) > 0$ в силу изложенного выше и $A = 2\sum_{i=1}^{n} x_i^2 > 0$, следовательно, согласно достаточному условию функция имеет единственную точку

минимума, определяемую из системы нормальных уравнений. Заметим, что в этой точке функция $S = S(a; b)$ имеет не просто локальный минимум, но наименьшее значение (глобальный минимум).

Пример. Имеются следующие данные о цене на нефть x (ден. ед.) и индексе акций нефтяных компаний y (усл. ед.).

x	17,28	17,05	18,30	18,80	19,20	18,50
y	537	534	550	555	560	552

Предполагая, что между переменными x и y существует линейная зависимость, найти эмпирическую формулу вида $y = ax + b$, используя метод наименьших квадратов.

Р е ш е н и е. Найдем необходимые для расчетов суммы $\sum_{i=1}^{n} x_i$, $\sum_{i=1}^{n} y_i$, $\sum_{i=1}^{n} x_i y_i$, $\sum_{i=1}^{n} x_i^2$. Промежуточные вычисления оформим в виде вспомогательной таблицы.

x_i	y_i	$x_i\, y_i$	x_i^2
17,28	537	9279,36	298,5984
17,05	534	9104,70	290,7025
18,30	550	10065,00	334,8900
18,80	555	10434,00	353,4400
19,20	560	10752,00	368,6400
18,50	552	10212,00	342,2400
$\sum$ 109,13	3288	59847,06	1988,5209

Система нормальных уравнений имеет вид

$$\begin{cases} 1988{,}5209a + 109{,}13b = 59847{,}06, \\ 109{,}13a + 6b = 3288. \end{cases}$$

Ее решение $a = 12{,}078$, $b = 328{,}32$ дает искомую зависимость: $y = 12{,}078x + 328{,}32$. Таким образом, с увеличением цены нефти на 1 ден. ед. индекс акций нефтяных компаний в среднем растет на 12,08 ед.

Пример. Динамика численности населения.

Вся история развития человечества неразрывно связана с изменениями динамики численности и воспроизводства населения. Из-за отсутствия достоверных данных трудно однозначно оценить динамику численности мирового населения практически вплоть до начала XIX века, когда во многих европейских странах стали проводиться переписи населения в их современном понимании.

Тем не менее, основываясь на приблизительных данных учета мирового населения, о котором упоминается еще в Библии, где приводится численность «сынов Израилевых» – более 600 тысяч человек, можно говорить о постоянном, хотя и очень медленном, росте населения мира.

Динамика численности населения Америки в целом со времени Рождества Христова (млн. человек) (По А.Я. Кваша, В.А. Ионцевой, 1995):

Начало нашей эры	1000	1200	1500	1750	1900
3	13	23	41	15	144

Приближенное вычисление интегралов. Формули прямокутників і трапецій.

Нехай треба обчислити значення визначеного інтегралу $\int_a^b f(x)dx$, де $f(x)$ є деяка задана на проміжку $[a,b]$ неперервна функція. Існує багато прикладів обчислення подібних інтегралів, або за допомогою первісної, якщо вона виражається в скінченному вигляді, або ж – минаючи первісну – за допомогою різних прийомів, як правило, штучних. Потрібно відмітити, однак, що всім цим вичерпується вузький клас інтегралів; за його межами зазвичай вдаються до різних методів наближеного обчислення.

В даній роботі можна ознайомитися з основними із цих методів, в яких наближені формули для інтегралів складаються по деякому числу значень підінтегральної функції, обчислених для ряду (зазвичай рівновіддалених) значень незалежної змінної.

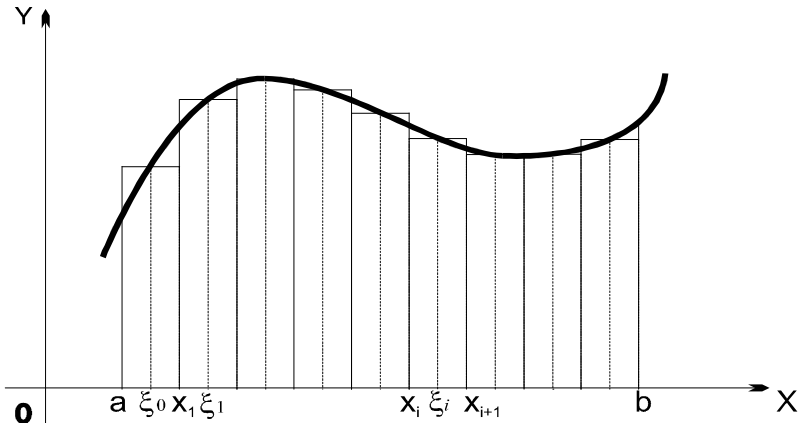
Перші формули, які сюди відносяться, простіші всього отримуються із геометричних міркувань. Витлумачуючи визначений інтеграл $\int_a^b f(x)dx$ як площу деякої фігури, яка обмежена кривою $y=f(x)$, ми і ставимо перед собою задачу знаходження цієї площі.

Перш за все, вдруге використовуючи ту думку, яка привела нас до самого поняття о визначеним інтегралі, можна розбити усю фігуру (мал. 1) на смуги, скажемо однієї і той же ширини $\Delta_{x_i}=\frac{b-a}{n}$, а потім кожну смугу наближено замінити прямокутником, за висоту якого прийнята будь-яка із його ординат. Це приводе нас до формули

$$\int_a^b f(x)dx=\frac{b-a}{n}\left[f(\xi_1)+f(\xi_2)+\ldots+f(\xi_{n-1})\right],$$

де $x_i\le\xi_i\le x_{i+1}$ $(i=0,1,\ldots,n-1)$. Тут шукана площа криволінійної фігури замінюється площею деякої ступінчатої фігури, яка складається із прямокутників (або ж, можна сказати, що визначений інтеграл замінюється інтегральною сумою). Ця наближена формула і називається формулою прямокутників.

На практиці зазвичай беруть $\xi_i=\frac{x_i+x_{i+1}}{2}=x_{i+1/2}$; якщо відповідну середню



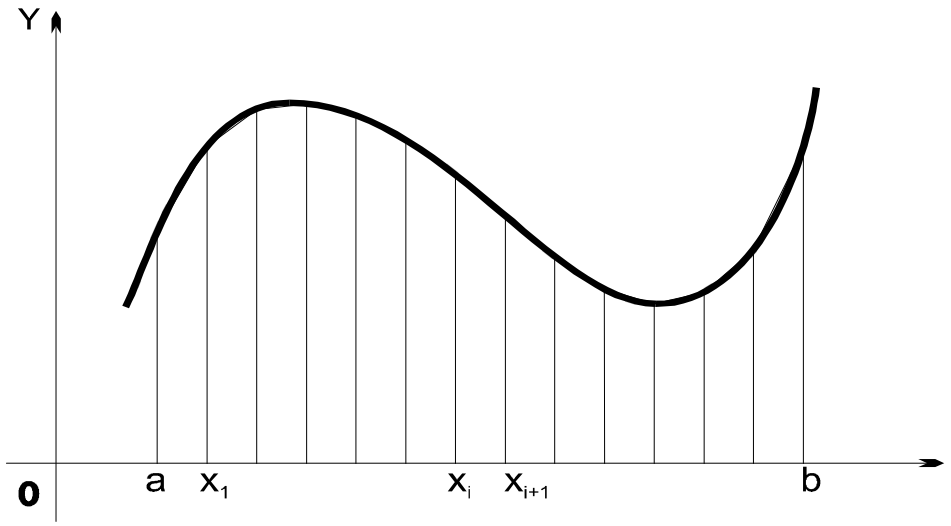
ординату $f(\xi_i) = f\left(x_{i+1/2}\right)$ позначити через $y_{i+1/2}$, то формула перепишеться у вигляді

$$\int_a^b f(x)dx = \frac{b-a}{n}\left(y_{1/2} + y_{3/2} + \ldots + y_{n-1/2}\right). \quad (1)$$

Надалі, кажучи про формулу прямокутників, ми будемо мати на увазі якраз цю формулу.

Геометричні міркування природно приводять і до другої, часто використовуваємій наближеній формулі. Замінивши дану криву вписаною в неї ламаною, з вершинами у точках (x_i, y_i), где $y_i = f(x_i)$ $(i = 0,1,\ldots,n-1)$. Тоді наша криволінійна фігура заміниться іншою, яка складається із ряду трапецій (рис2.). Якщо, як і раніш рахувати, що проміжок $[a,b]$ розбитий на рівні частини, то площі цих трапецій будуть

$$\frac{b-a}{n}\frac{y_0 + y_1}{2}, \frac{b-a}{n}\frac{y_1 + y_2}{2}, \ldots, \frac{b-a}{n}\frac{y_{n-1} + y_n}{2}.$$



Мал. 2

Додаючи, прийдемо до нової наближеної формули

$$\int_a^b f(x)dx = \frac{b-a}{n}\left(\frac{y_0 + y_n}{2} + y_1 + y_2 + \ldots + y_{n-1}\right). \quad (2)$$

Це так звана формула трапецій.

Можна показати, що при зростанні n до нескінченності похибка формули прямокутників і формули трапецій нескінченно зменшується. Таким чином, при достатньо великому n обидві ці формули відтворюють шукане значення з довільним рівнем точності.

Параболічне інтерполювання.

Для наближеного обчислення інтеграла $\int_a^b f(x)dx$ можно спробувати замінити функцію $f(x)$ близьким до неї багаточленом

$$y = P_k(x) = a_0 x^k + a_1 x^{k-1} + \ldots + a_{k-1}x + a_k \qquad (3)$$

і покласти

$$\int_a^b f(x)dx = \int P_k(x)dx$$

Можна сказати, що тут – при обрахуванні площі – дана крива $y = f(x)$ замінюється на параболу k - го порядку (3), в зв′язку з чим цей процес отримав назву параболічного интерполювання.

Сам вибір інтерполюючого багаточлена $P_k(x)$ частіше всього виконують наступним чином. У проміжку $[a,b]$ беруть $k+1$ значень незалежної змінної $\xi_0, \xi_1, \ldots, \xi_k$ і підбирають многочлен $P_k(x)$ так, щоб при усіх взятих значеннях x його значення співпадало зі значенням функції $f(x)$. Цією умовою, як ми знаємо, многочлен $P_k(x)$ визначається однозначно, і його вираз дається інтерполяційною формулою Лагранжа:

$$P_k(x) = \frac{(x-\xi_1)(x-\xi_2)\ldots(x-\xi_k)}{(\xi_0-\xi_1)(\xi_0-\xi_2)\ldots(\xi_0-\xi_k)} f(\xi_0) + \frac{(x-\xi_0)(x-\xi_2)\ldots(x-\xi_k)}{(\xi_1-\xi_0)(\xi_1-\xi_2)\ldots(\xi_1-\xi_k)} f(\xi_1) + \ldots$$
$$\ldots + \frac{(x-\xi_0)(x-\xi_1)\ldots(x-\xi_{k-1})}{(\xi_k-\xi_0)(\xi_k-\xi_1)\ldots(\xi_k-\xi_{k-1})} f(\xi_k)$$

При інтерполюванні виходить лінійний, відносно значень $f(\xi_0), \ldots, f(\xi_k)$ вираз, коефіцієнти якого вже не залежать від цих значень. Вирахувавши коефіцієнти раз і назавжди, можна їх використовувати для будь-якої функції $f(x)$ в даному проміжку $[a,b]$.

В найпростішому випадку, при $k=0$, функція $f(x)$ просто замінюється сталою $f(\xi_0)$, де ξ_0 – будь-яка точка у проміжку $[a,b]$, скажемо, середня: $\xi_0 = \frac{a+b}{2}$. Тоді наближено

$$\int_a^b f(x)dx = (b-a) f\left(\frac{a+b}{2}\right) \qquad (4)$$

Геометрично – площа криволінійної фігури замінюється тут площадью прямокутника з висотою, яка рівна середній її ординаті.

При $k=1$ функція $f(x)$ замінюється лінійною функцією $P_1(x)$, яка має однакові з нею значення при $x=\xi_0$ и $x=\xi_1$. Якщо взяти $\xi_0=a$, $\xi_1=b$, то

$$P_1(x)=\frac{x-b}{a-b}f(x)+\frac{x-a}{b-a}f(b) \quad (5)$$

і, як легко обчислити,

$$\int_a^b P_1(x)dx=(b-a)\frac{f(a)}{2}.$$

Таким чином, тут ми наближено вважаємо

$$\int_a^b f(x)dx=(b-a)\frac{f(a)+f(b)}{2}$$

На цей раз площа криволінійної фігури замінюється площею трапеції: замість кривої береться хорда, яка сполучає її кінці.

Менш тривіальний результат отримаємо взявши $k=2$. Якщо покласти $\xi_0=a_0$, $\xi_1=\frac{a+b}{2}$, $\xi_2=b$, то інтерполяційний многочлен $P_2(x)$ буде мати вигляд

$$P_2(x)=\frac{\left(x-\frac{a+b}{2}\right)(x-b)}{\left(a-\frac{a+b}{2}\right)(a-b)}f(a)+$$

$$+\frac{(x-a)(x-b)}{\left(\frac{a+b}{2}-a\right)\left(\frac{a+b}{2}-b\right)}f\left(\frac{a+b}{2}\right)+\frac{(x-a)\left(x-\frac{a+b}{2}\right)}{(b-a)\left(b-\frac{a+b}{2}\right)}f(b). \quad (7)$$

За допомогою легкого обчислення вираховуємо

$$\int_a^b\frac{\left(x-\frac{a+b}{2}\right)(x-b)}{\left(a-\frac{a+b}{2}\right)(a-b)}dx=\frac{2}{(b-a)^2}\int_a^b\left[(x-b)+\frac{b-a}{2}\right](x-b)=$$

$$=\frac{2}{(b-a)^2}\left[\frac{(x-b)^3}{3}+\frac{b-a}{2}\frac{(x-b)^2}{2}\right]_a^b=\frac{b-a}{6}$$

і, аналогічно

$$\int_a^b \frac{(x-a)(x-b)}{\left(\frac{a+b}{2}-a\right)\left(\frac{a+b}{2}-b\right)}dx = 4\frac{b-a}{6},$$

$$\int_a^b \frac{(x-a)\left(x-\frac{a+b}{2}\right)}{(b-a)\left(b-\frac{a+b}{2}\right)}dx = \frac{b-a}{6}.$$

Таким чином, приходимо до наближеної формули

$$\int_a^b f(x)dx = \frac{b-a}{2}\left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right].$$

Тут площа фігури під даною кривою замінюється площею фігури, яка обмежена звичайною параболою (з вертикальною віссю), що проходить через крайні і середню точки кривої.

При збільшенні степені k інтерполяційного поліному, тобто якщо побудувати параболу (3) через все більше число даної кривої, можна розраховувати отримати більшу точність. Но більш практичним виявляється інший шлях, якій ґрунтується на поєднанні ідеї параболічного інтерполювання із ідеєю дроблення.