ПРИЛОЖЕНИЕ 4

Два математика наблюдают за дверью в помещение.
Из этой двери сначала выходят два человека, а потом
туда заходит один человек. Один математик другому:
– Сейчас туда войдет еще человек, и тогда там никого не будет.
Анекдот

Расчет параметров линейного уравнения регрессии методом наименьших квадратов

Уравнение регрессии – это уравнение, описывающее корреляционную зависимость между признаком – результатом Y и признаками факторами (одним или несколькими).

Наиболее часто для описания статистической связи признаков используется линейное уравнение регрессии. Внимание к линейной форме связи объясняется четкой интерпретацией параметров линейного уравнения регрессии, ограниченной вариацией переменных и тем, что в большинстве случаев нелинейные формы связи для выполнения расчетов преобразуют (путем логарифмирования или замены переменных) в линейную форму.

Линейное парное уравнение регрессии имеет вид: $y_i^* = a + bx_i$, i=1,…,n,

где n – объем совокупности (число наблюдений).

Оценки параметров линейной регрессии (a и b) могут быть найдены разными методами, наиболее распространенным является метод наименьших квадратов. Данный метод позволяет получить такие оценки параметров a и b, при которых сумма квадратов отклонений фактических значений результативного признака y_i от расчетных (теоретических) значений y_i^* (рассчитанных по уравнению регрессии) минимальна.

Непосредственно коэффициенты уравнения рассчитываются по представленным справа формулам; черта сверху означает осреднение.	$b = \dfrac{\overline{X \cdot Y} - \overline{X} \cdot \overline{Y}}{\sigma_x^2}, \qquad a = \overline{Y} - b \cdot \overline{X}$

МНК (метод наименьших квадратов) является достаточно точным приемом и позволяет получить вполне надежные результаты. Одновременно он является *интерполяционным* методом, поскольку обеспечивает с определенной вероятностью предсказание любых значений y_i *в интервале* изученных значений x_i.

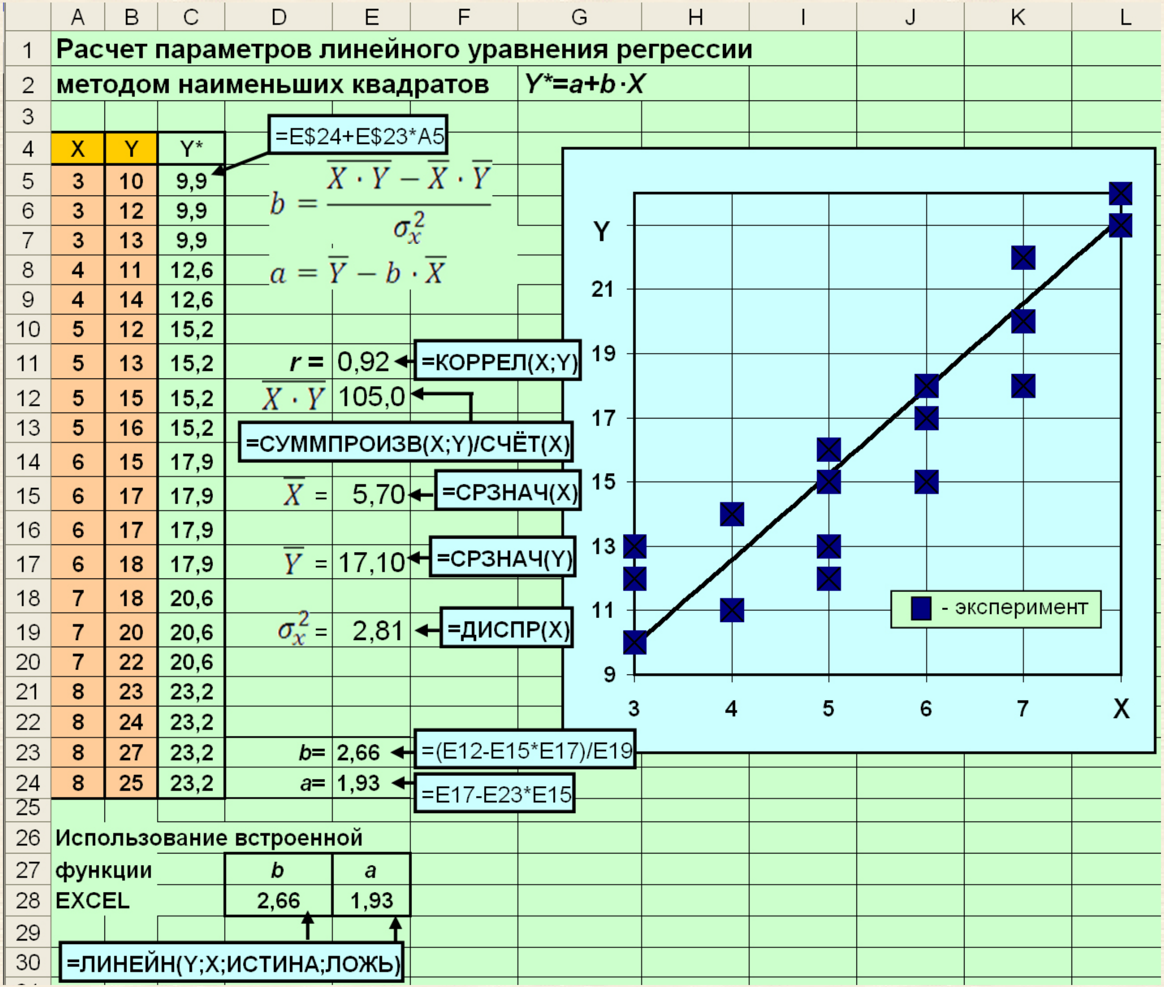
Напомним, что *экстраполяционный* метод (в отличие от интерполяционного) дает возможность предсказывать результаты *за пределами* изученной области.

После того как уравнение регрессии найдено, необходимо определить его статистическую пригодность, т.е. выяснить, насколько оно верно (надежно) предсказывает в интервале x_1; x_2; … x_n экспериментальные результаты для y. Подобную оценку принято называть проверкой на значимость или адекватность.

Пример П4. Пусть заданы массивы Y "функции" и X аргумента выборки эмпирических данных, представленные (и выделенные цветом) на рис. П4.1 соответствующими диапазонами ячеек. Требуется, в случае высокой степени их корреляции, построить уравнение линейной регрессии.

Последовательность вычислений в среде Excel для "ручного" способа вычисления коэффициентов уравнения регрессии следующая.

1. В диапазон A5:A24 заносятся исходные данные X аргумента выборки. Для удобства данный диапазон именуется (выделяет диапазон, далее ВСТАВКА–ИМЯ–ПРИСВОИТЬ) именем X (см.рис. П4.1). Исходные данные Y заносятся в ячейки B5:B24. Диапазон именуется именем Y.
2. Рассчитывается коэффициент корреляции, значение которого позволяет сделать заключение об адекватности строящегося уравнения линейной регрессии. Чем ближе величина коэффициента к 1, тем лучше экспериментальные данные будут описываться уравнением линейной регрессии. В ячейку E11 заносится формула =КОРРЕЛ(X;Y) {=CORREL (X;Y)}.
3. В ячейке E12 рассчитывается вспомогательное значение $\overline{X \cdot Y}$ с помощью формулы =СУММПРОИЗВ(X;Y) /СЧЁТ(X) {=SUMPRODUCT(X;Y) /COUNT(X)}.
4. В ячейках E15,E17 определяются средние значения параметров формулами =СРЗНАЧ(X) {=AVERAGE(X)} и =СРЗНАЧ(Y) {=AVERAGE(Y)}.



Расчет параметров линейного уравнения регрессии методом наименьших квадратов $Y^*=a+b\cdot X$

X	Y	Y*
3	10	9,9
3	12	9,9
3	13	9,9
4	11	12,6
4	14	12,6
5	12	15,2
5	13	15,2
5	15	15,2
5	16	15,2
6	15	17,9
6	17	17,9
6	17	17,9
6	18	17,9
7	18	20,6
7	20	20,6
7	22	20,6
8	23	23,2
8	24	23,2
8	27	23,2
8	25	23,2

=E$24+E$23*A5

$$b=\frac{\overline{X\cdot Y}-\overline{X}\cdot\overline{Y}}{\sigma_x^2}$$

$$a=\overline{Y}-b\cdot\overline{X}$$

$r =$ 0,92 ← =КОРРЕЛ(X;Y)

$\overline{X\cdot Y}$ 105,0 ← =СУММПРОИЗВ(X;Y)/СЧЁТ(X)

$\overline{X} =$ 5,70 ← =СРЗНАЧ(X)

$\overline{Y} =$ 17,10 ← =СРЗНАЧ(Y)

$\sigma_x^2 =$ 2,81 ← =ДИСПР(X)

$b=$ 2,66 ← =(E12-E15*E17)/E19

$a=$ 1,93 ← =E17-E23*E15

Использование встроенной функции EXCEL

b	a
2,66	1,93

=ЛИНЕЙН(Y;X;ИСТИНА;ЛОЖЬ)

Рис. П4.1. Скриншот расчетного листа MS Excel (линейная регрессия)

5. Рассчитываются коэффициенты уравнения регрессии. В ячейке E23 определяется коэффициент b формулой =(E12-E15*E17)/E19, в ячейке E24 коэффициент a формулой =E17-E23*E15.
6. Для проверки и анализа полученных коэффициентов строится столбик расчетных значений y_i^* (C5:C24) вводом в ячейку C5 формулы =E$24 +E$23*A5 и тиражированием ее до адреса C24.

Для вычисления коэффициентов a и b можно использовать встроенную Excel-функцию, которая по правилам работы с массивами (использование F2 затем Ctrl+Shift+Enter) вычисляет оба значения a и b (ячейки E28 и D28) формулой =ЛИНЕЙН (Y;X; ИСТИНА; ЛОЖЬ) {LINEST (Y;X; TRUE(); FALSE()), формула вводится в выделенные ячейки и активируется через Ctrl+Shift+Enter }.