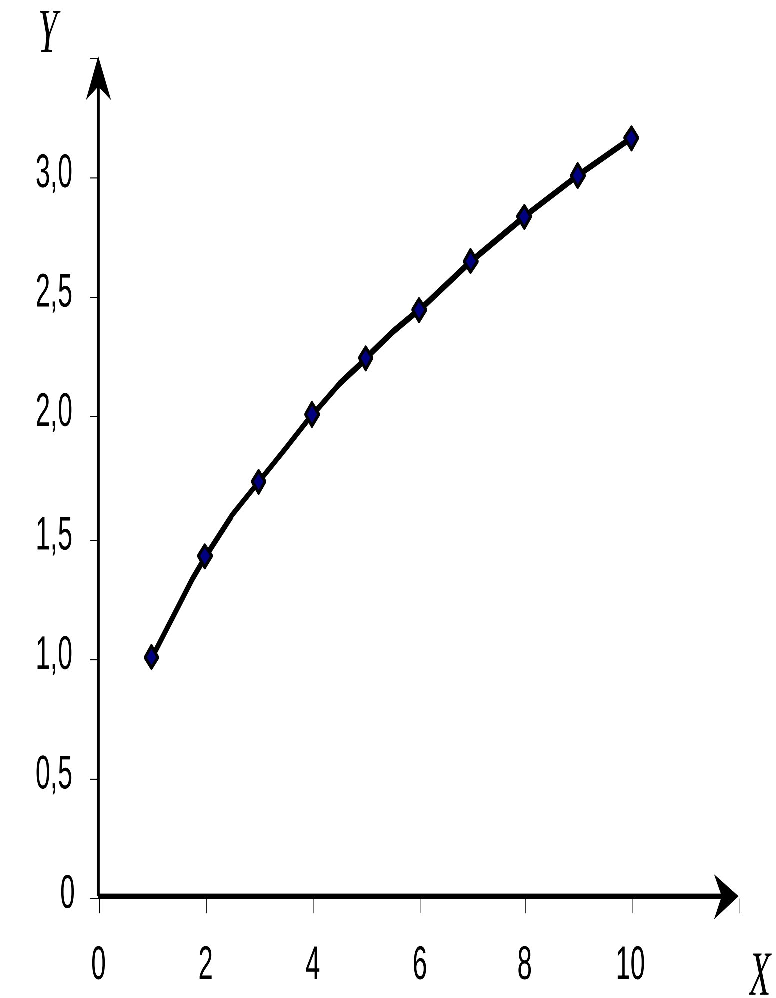
Корреляционный анализ.

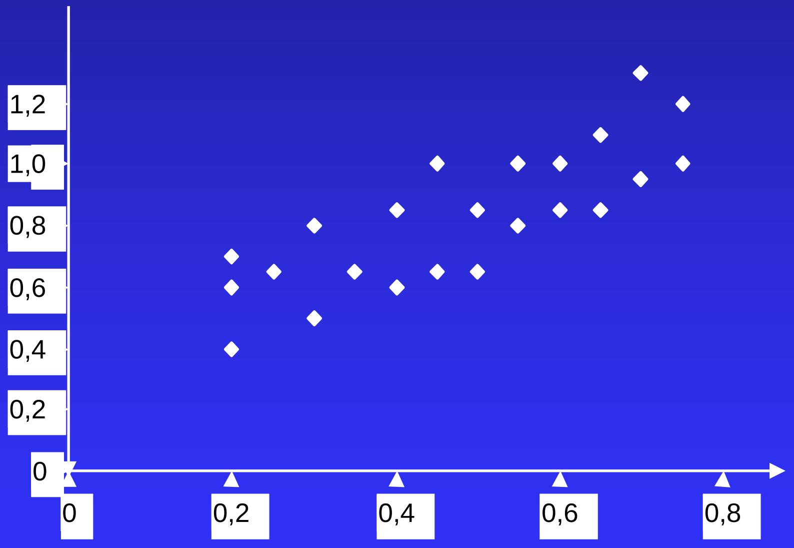
Корреляционно-регрессионный анализ выполняется на основе анализа эмпирических данных.

Методы такого анализа являются составной частью эконометрики, которая устанавливает и исследует количественные закономерности в экономике на основе методов теории вероятностей и математической статистики.



Детерминированная (функциональная) модель

В отличие от детерминированных (функциональных) моделей, которые характеризуются однозначным соответствием между факторными и результативными показателями (рис. 3.1), в стохастических (корреляционных) моделях каждому значению факторного признака может соответствовать несколько значений результативного показателя



Стохастическая (корреляционная) связь между факторным и результативным показателями

Методы корреляционно-регрессионного анализа базируются на следующих положениях.

Результативный показатель рассматривается как сумма двух случайных составляющих по формуле

$$y = f\left(x_i\right) + \varepsilon$$

где $f\left(x_i\right)$ – составляющая результативного показателя, определенная в виде функции от влияющих факторов;

ε — составляющая, которая отражает влияние неучтенных случайных факторов на анализируемый результативный показатель.

Процесс корреляционно-регрессионного анализа включает следующие этапы

Постановочно-информационный этап.

Устанавливается набор показателей, регистрируемых на каждом из наблюдаемых объектов, определяются цели исследования. Проводится сбор необходимой информации.

Корреляционный анализ

Решаются следующие задачи:

– выбор измерителя корреляционной связи (коэффициента корреляции, корреляционного отношения, коэффициента детерминации и т.д.);

– расчет тесноты связи по имеющимся выборочным данным;

– установление структуры связей между исследуемыми факторными и результативным показателями.

Регрессионный анализ.

Выбирается общий вид функции регрессии, определяются коэффициенты уравнения регрессии.

В зависимости от числа исследуемых признаков различают парную и множественную корреляцию. Если исследуется связь между двумя признаками (результативным и факторным), то корреляция называется *парной*. Если факторных показателей несколько, то корреляцию называют *множественной*. В процессе анализа парных зависимостей решаются следующие задачи:

установление вида зависимости между факториальными и результативным показателями;

определение коэффициентов уравнения регрессии;

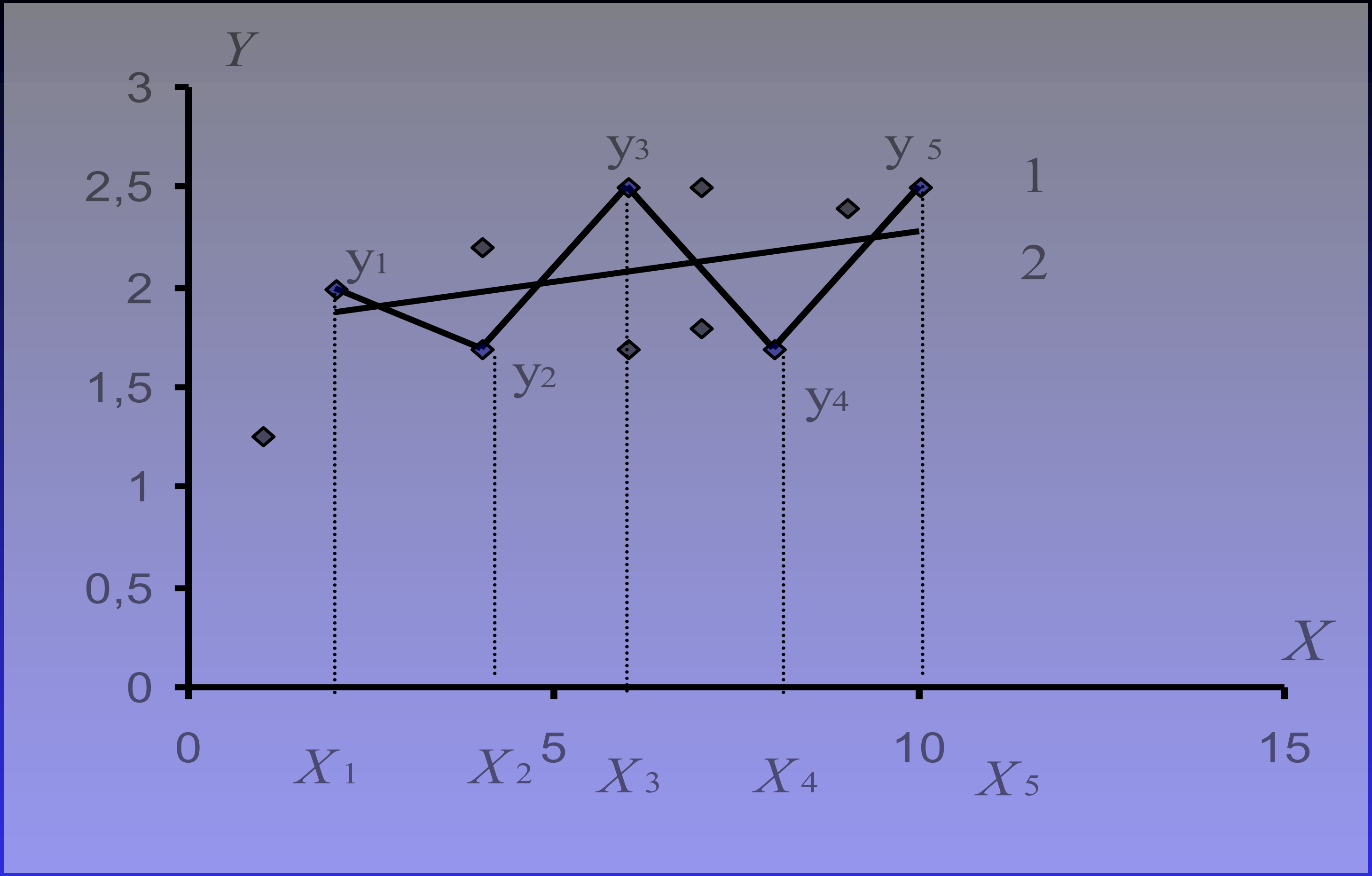
определение тесноты (силы) связи.

Установление вида зависимости

Исходные данные для корреляционного анализа, представленные в виде таблицы

y	x
y_1	x_1
y_2	x_2
…	…
y_i	x_i
…	…
y_m	x_m

изображаются на графике



— Поле корреляции

1 — эмпирическая линия регрессии;

2 — теоретическая линия регрессии

Устанавливается диапазон изменения факторного признака — x_{max}, x_{min}. Массив данных делится на группы. Длина интервала по каждой группе определяется по формуле,

$$\Delta x = \frac{x_{max} - x_{min}}{1 + 3{,}2 lgm}$$

где *m* — число наблюдений.

Для каждой группы данных определяются средние значения:

$\bar{x}$ и $\bar{y}$

$$\bar{x}_1 = x_{min} + \frac{\Delta x}{2};$$

$$\bar{x}_2 = \bar{x}_1 + \Delta x;$$

$$\ldots \quad \ldots \quad \ldots \quad \ldots$$

$$\bar{x}_k = \bar{x}_{k\text{-}1} + \Delta x;$$

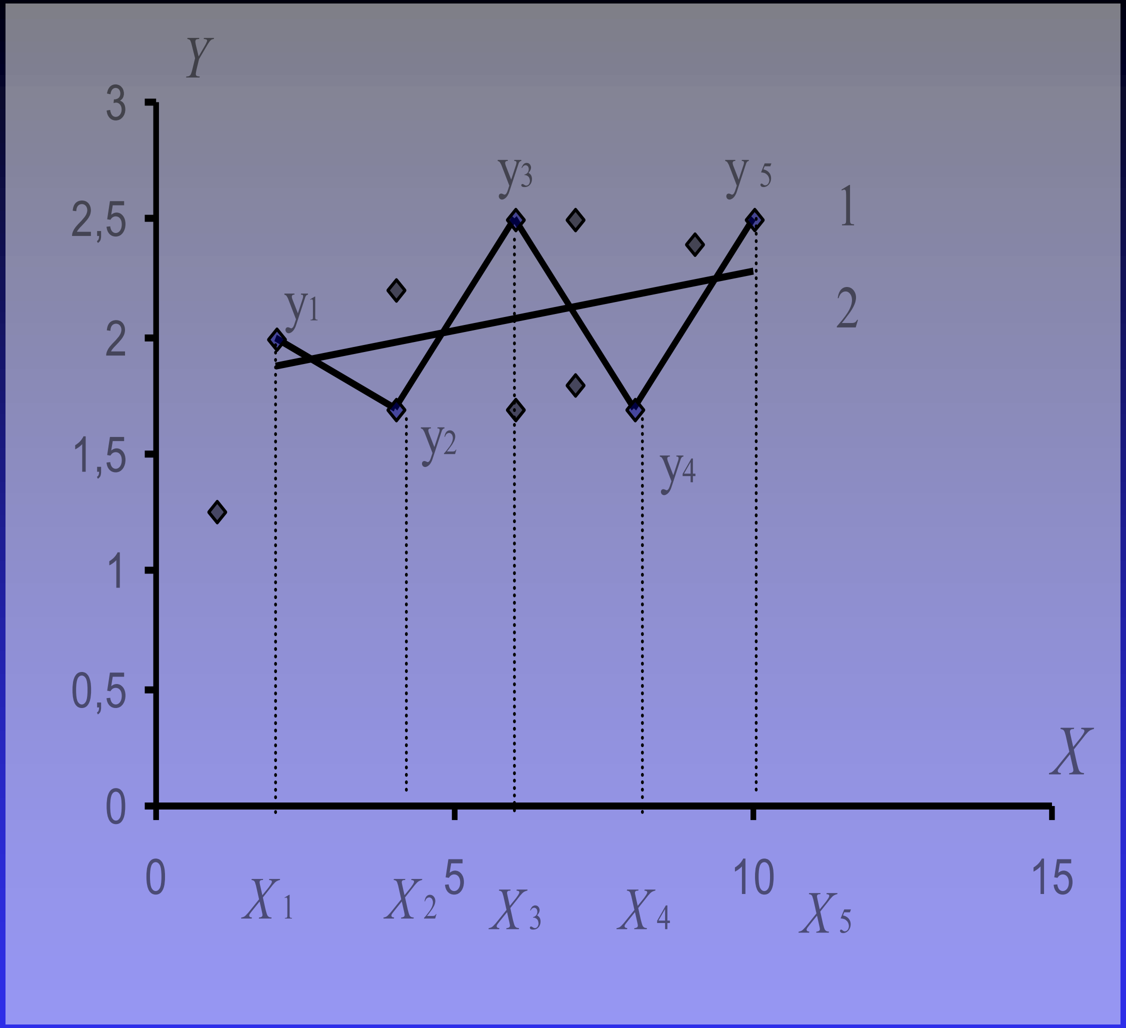
$$\bar{y}_k = \frac{1}{n} \sum_{j=1}^{n} y_j,$$

где *j*=1, …, *n* — номера точек, попавших в интервал *k*.

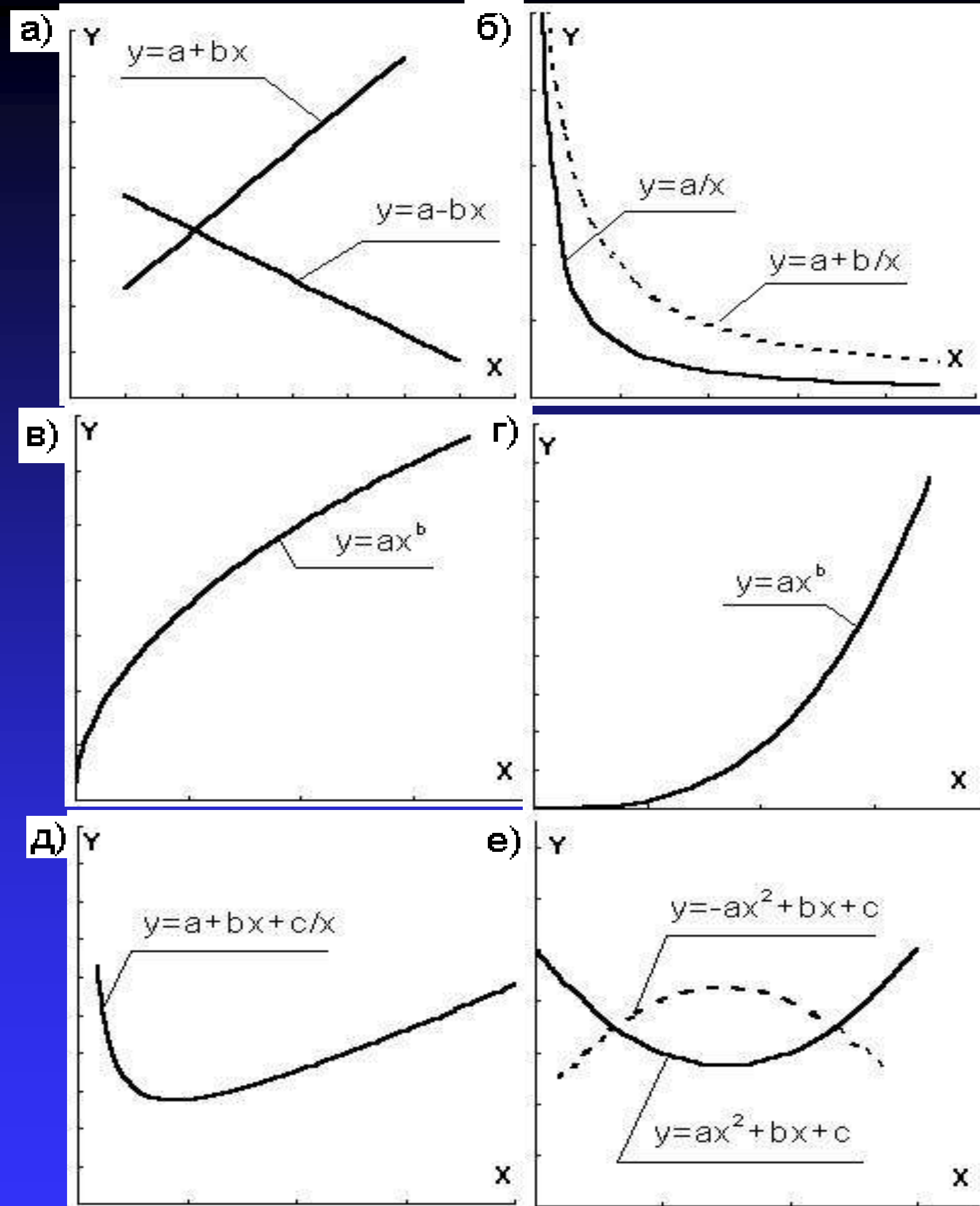
Ломанная линия, полученная путем соединения соседних точек, носит название эмпирической линии регрессии. По внешнему виду эмпирической линии регрессии можно сделать вывод о виде факторной модели.

Прежде всего, необходимо гипотетически предположить является ли искомая зависимость линейной или нелинейной. При этом учитываются причинно-следственные связи между фактором и результативным показателем, экономический смысл исследуемой зависимости, опыт предыдущих исследований данной пары признаков.

Конечным результатом анализа является установление вида теоретической линии регрессии.



Поле корреляции
1 — эмпирическая линия регрессии;
2 — теоретическая линия регрессии



Факторные модели, используемые в экономическом анализе

Определение коэффициентов уравнения регрессии

Коэффициенты уравнения регрессии (теоретической линии регрессии) определяются методом наименьших квадратов. Этот метод широко используется при обработке результатов наблюдений в эконометрии.

Суть метода состоит в том, чтобы подобрать параметры (коэффициенты) теоретической линии регрессии с условием минимизации суммы квадратов отклонений теоретических значений результативного показателя от его эмпирических значений.

Определение тесноты связи

Теснота связи для парных зависимостей отражает долю влияния рассматриваемого фактора *x* среди всех факторов (в том числе неучтенных и случайных), оказывающих влияние на результативный показатель *y*.

Теснота связи оценивается по величине рассеяния значений *y*. Небольшое рассеяние эмпирических точек относительно теоретической линии регрессии свидетельствует о сильной связи между *x* и *y*, большое рассеяние указывает на слабую связь.

Мера рассеяния признака относительно его среднего значения оценивается величиной дисперсии или среднего квадратического отклонения.

Среднее значение признака х

$$\bar{x} = \frac{\sum_{i=1}^{m} x_i}{m}$$

Дисперсия определяется как средний квадрат отклонения признака *x* от его среднего значения

$$D_x = \frac{\sum_{i=1}^{m}\left(x_i - \bar{x}\right)^2}{m}$$

Среднеквадратическое отклонение

$$\sigma_x = \sqrt{D_x} = \sqrt{\frac{\sum_{i=1}^{m}\left(x_i - \bar{x}\right)^2}{m}}$$

Преобразуем формулу дисперсии:

$$D_x = \frac{\sum\limits_{i=1}^{m} \left(x_i^2 - 2x_i\bar{x} + \bar{x} \right)}{m} =$$

$$= \frac{\sum\limits_{i=1}^{m} x_i^2}{m} - \frac{2\bar{x}\sum x_i}{m} + \frac{m\left(\bar{x}\right)^2}{m} =$$

$$= \overline{x^2} - 2\bar{x}\bar{x} + \left(\bar{x}\right)^2 = \overline{x^2} - 2\left(\bar{x}\right)^2 + \left(\bar{x}\right)^2;$$

Дисперсия равна разности среднего значения квадратов признака *x* и квадрата среднего значения *x*.

$$D_x = \overline{x^2} - \left(\bar{x}\right)^2$$

При оценке тесноты связи долю влияния фактора *x* можно определить как отношение дисперсии результативного показателя *y* за счет фактора *x* к общей дисперсии *y*.

Этот показатель называется *коэффициентом детерминации*

$$R^2 = \frac{D_{y_x}}{D_y}$$

Если уравнение теоретической линии имеет вид

$$y_x = a + bx$$

то их среднее значение

$$\bar{y}_x = a + b\bar{x}$$

Тогда дисперсия за счет фактора *x* определяется из выражения:

$$D_{y_x} = \frac{\sum_{i=1}^{m} \left[(a + bx) - (a + b\bar{x}) \right]^2}{m}$$

Поскольку

$$\frac{\sum (x - \bar{x})^2}{m} = D_x$$

то

$$D_{y_x} = b^2 \cdot D_x$$

Общая дисперсия показателя *y*

$$D_y = \frac{\sum_{i=1}^{m} (y_i - \bar{y})^2}{m}$$

Коэффициент детерминации можно определить следующим образом

$$R^2 = \frac{b^2 D_x}{D_y}$$

Для оценки тесноты связи линейных зависимостей также используется показатель, который называется *коэффициентом корреляции (r)*:

$$r = \sqrt{R^2}$$

$$r = b\frac{\sigma_x}{\sigma_y}$$

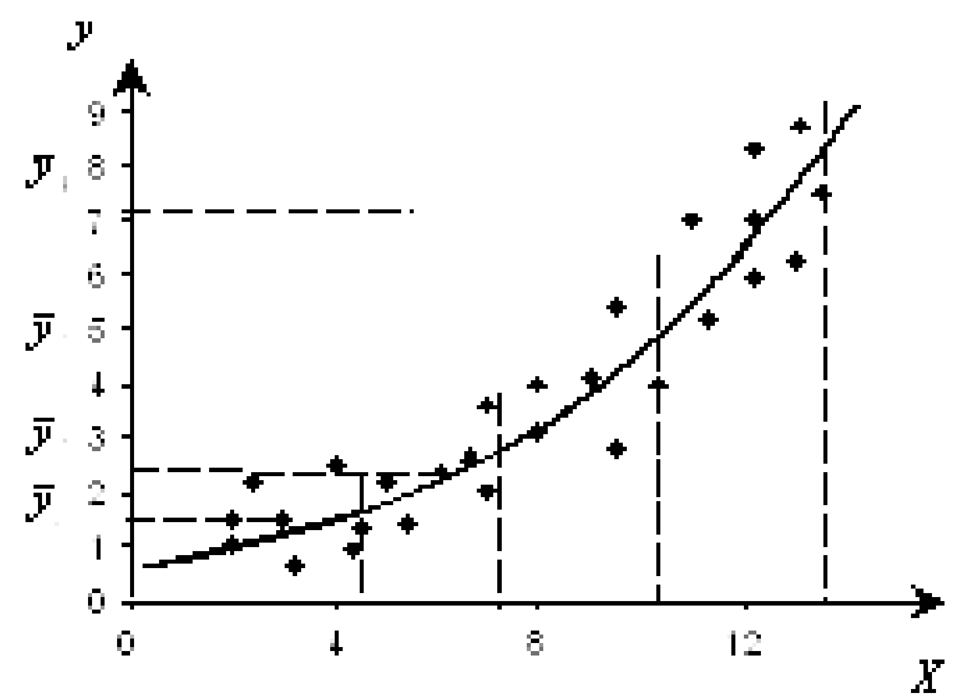
Сильная	0,7÷1,0	0,49–1,00
Средняя	0,3÷0,7	0,09–0,49
Слабая	<0,3	<0,09

Коэффициент корреляции может быть рассчитан до решения задачи определения коэффициентов уравнения регрессии.

В этом случае его нужно выразить через статистические характеристики исходных данных.

Теснота связи нелинейных зависимостей

Рассмотрим поле корреляции для нелинейной зависимости



Поле корреляции нелинейной зависимости

Всю совокупность исходных данных разделим на несколько групп. Номера групп обозначим буквой j (j=1, …, n).

В каждой группе определим средние значения показателя y:

где y_j — значения показателя y, опавшие в j–ю группу;

m_j — количество точек, попавших в j–ю группу.

Полученные средние значения y называются *групповыми средними*.

$$\bar{y}_1 = \frac{\sum y_1}{m_1} \qquad \bar{y}_2 = \frac{\sum y_2}{m_2} \qquad \bar{y}_n = \frac{\sum y_n}{m_n}$$

Определим межгрупповую дисперсию, которая представляет собой дисперсию групповых средних относительно общей средней

где $\bar{y}$ общая средняя показателя *y*

$$\bar{y} = \frac{\sum_{j=1}^{n} m_j y_j}{\sum_{j=1}^{n} m_j}$$

$$D_{межгр} = \frac{\sum_{j=1}^{n} m_j (\bar{y}_j - \bar{y})^2}{\sum_{j=1}^{n} m_j}$$

Теснота связи нелинейных зависимостей оценивается корреляционным отношением, которое представляет собой отношение межгруппового среднего квадратического отклонения к общему среднему квадратическому отклонению показателя *y*

$$\eta = \frac{\sigma_{\text{межгр}}}{\sigma_{\text{общ}}}$$

$$\sigma_{\text{межгр}} = \sqrt{D_{\text{межгр}}}$$

$$\sigma_{\text{общ}} = \sqrt{D_{\text{общ}}} = \sqrt{\frac{\sum m_j \left(y_j - \bar{y}\right)^2}{\sum m_j}}$$

Корреляционное отношение служит мерой тесноты связи любой, в том числе и линейной связи. Если корреляционное отношение равно коэффициенту корреляции, то это свидетельствует о том, что исследуемая зависимость может быть выражена в виде линейной корреляционной связи

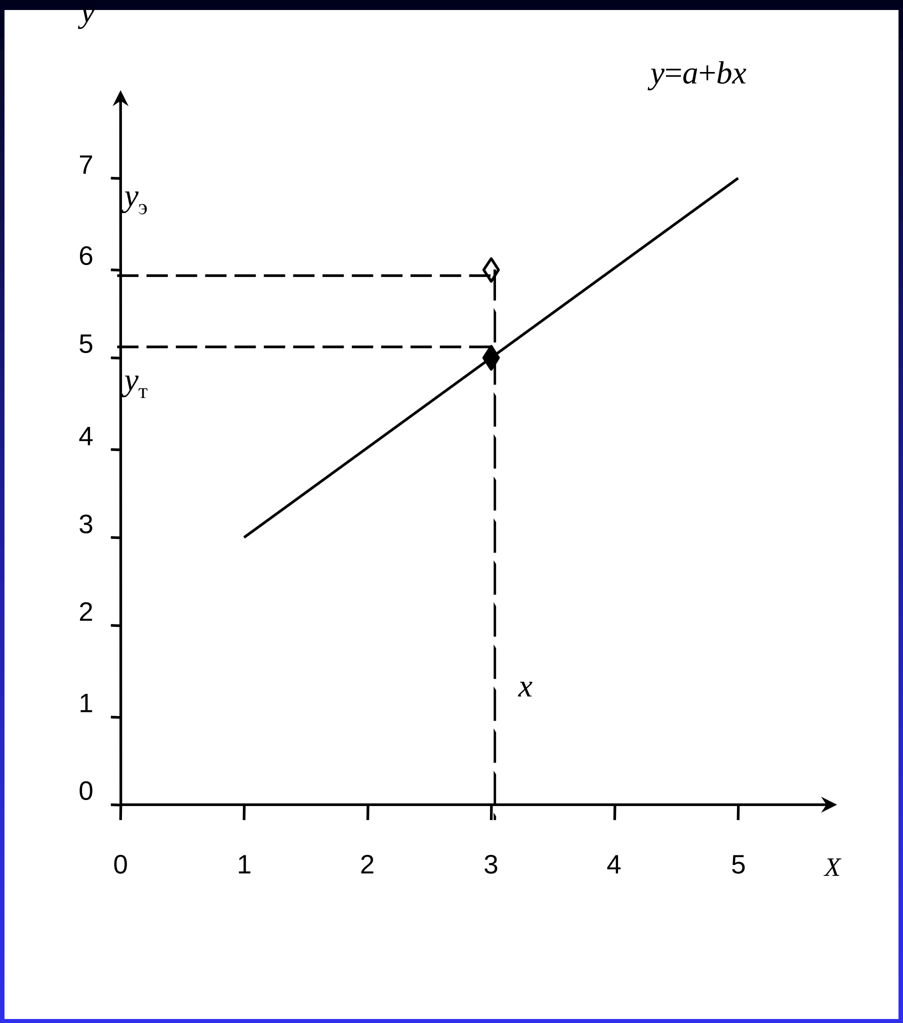
В процессе многомерного корреляционно-регрессионного анализа устанавливается зависимость результативного показателя от нескольких факторов.

Уравнение множественной регрессии имеет следующий вид

$$y = (x_1, \ldots, x_i, \ldots, x_m)$$

Например, для линейного уравнения регрессии необходимо определить такие значения коэффициентов *a* и *b*, которые минимизировали бы сумму квадратов отклонений теоретических значений $y_т$ от эмпирических $y_э$ для всех точек, построенных по данным наблюдений

$$S = \sum_{i=1}^{m} \left(y_{m_i} - y_{э_i} \right)^2 \rightarrow min$$



Геометрический смысл метода наименьших квадратов

Линейная зависимость

В соответствии с методом наименьших квадратов минимизируемая сумма имеет следующих вид

$$S = \sum_{i=1}^{m} (a + bx_i - y_i)^2 \to min$$

Для отыскания минимума приравняем нулю соответствующие частные производные:

$$\frac{\partial S}{\partial a} = 2\sum_{i=1}^{m} (a + bx_i - y_i) = 0 \qquad \frac{\partial S}{\partial b} = 2\sum_{i=1}^{m} (a + bx_i - y_i) x_i = 0$$

Выполним простейшие преобразования

$$ma + b\sum x_i - \sum y_i = 0 \qquad a\sum x_i + b\sum x_i^2 - \sum x_i y_i = 0 .$$

Первое уравнение умножим на , $\sum x_i$ а второе — на $(-m)$

$$ma\sum x_i + b\left(\sum x_i\right)^2 - \sum x_i \sum y_i = 0$$

$$-ma\sum x_i - bm\sum x_i^2 + m\sum x_i y_i = 0$$

Решив эту систему уравнений, найдем искомые коэффициенты:

$$a = \frac{\sum y_i - b\sum x_i}{m}$$

$$b = \frac{\sum x_i \sum y_i - m\sum x_i y_i}{\left(\sum x_i\right)^2 - m\sum x_i^2}$$

Гиперболическая зависимость

Гиперболическую факторную модель вида можно линеаризовать путем замены переменной *x* на x`=1/x , после этого коэффициенты *a* и *b* можно определить по формулам, выведенным для линейной зависимости.

$$y = ax^b$$

$$\ln y = \ln a + b \ln x$$

Степенная зависимость

Степенная факторная модель линеар $y = ax^b$

методом логарифмирования

Для определения коэффициентов линии регрессии производится замена переменных:

На основе расчетов, выполненных по формулам линейной зависимости, определяются параметры линейной модели

$$y' = a' + bx'$$

Коэффициент *a* степенной модели определяется из выражения

$$a = e^{a'}$$

$$a' = \ln a$$

$$y' = \ln y$$

$$x' = \ln x$$

В конечном итоге записывается степенная факторная модель

Параболическая зависимость

$$\left(y = ax^2 + bx + c\right)$$

$$S = \sum \left(ax^2 + bx + c - y\right)^2 \to min$$

$$\frac{\partial S}{\partial a} = 2\sum \left(ax^2 + bx + c - y\right)x^2 = 0$$

$$\frac{\partial S}{\partial b} = 2\sum \left(ax^2 + bx + c - y\right)x = 0$$

$$\frac{\partial S}{\partial c} = 2\sum \left(ax^2 + bx + c - y\right) = 0$$

Для определения коэффициентов *a*, *b* и *c* используется следующая система уравнений:

$$a\sum x^4 + b\sum x^3 + c\sum x^2 - \sum yx^2 = 0$$

$$a\sum x^3 + b\sum x^2 + c\sum x - \sum xy = 0$$

$$a\sum x^2 + b\sum x + mc - \sum y = 0$$

Определение коэффициентов уравнения регрессии для случая двух влияющих факторов

В **случае линейной многомерной зависимости, минимизируемая сумма квадратов отклонений теоретических значений результативного показателя от эмпирических имеет следующий вид**

$$S = \Sigma(a + bx_1 + cx_2 - y)^2 \rightarrow min$$

где *a*, *b*,…, *k* — коэффициенты уравнения регрессии;

x_1, …, x_m — факторные показатели;

$y_э$ — эмпирические значения результативного показателя.

Определим коэффициенты уравнения регрессии для случая двух влияющих факторов

$$S = \Sigma(a + bx_1 + cx_2 - y)^2 \rightarrow min$$

$$\frac{\partial S}{\partial a} = 2\Sigma(a + bx_1 + cx_2 - y) = 0$$

$$\frac{\partial S}{\partial b} = 2\Sigma(a + bx_1 + cx_2 - y)x_1 = 0$$

$$\frac{\partial S}{\partial c} = 2\Sigma(a + bx_1 + cx_2 - y)x_2 = 0$$

Коэффициенты *a*, *b*, и *c*

определяются путем

решения системы

уравнений с использованием правила Крамера.

$$na + b\sum x_1 + c\sum x_2 = \sum y$$

$$a\sum x_1 + b\sum x_1^2 + c\sum x_1 \cdot x_2 = \sum x_1 y$$

$$a\sum x_2 + b\sum x_1 \cdot x_2 + c\sum x_2^2 = \sum x_2 y$$

Определитель первой матрицы получается из *D* путем замены первого столбца на столбец свободных членов

$$D = \det\begin{pmatrix} n & \Sigma x_1 & \Sigma x_2 \\ \Sigma x_1 & \Sigma x_1^2 & \Sigma x_1 x_2 \\ \Sigma x_2 & \Sigma x_1 x_2 & \Sigma x_2^2 \end{pmatrix}$$

$$D_1 = \det\begin{pmatrix} \Sigma y & \Sigma x_1 & \Sigma x_2 \\ \Sigma x_1 y & \Sigma x_1^2 & \Sigma x_1 x_2 \\ \Sigma x_2 y & \Sigma x_1 x_2 & \Sigma x_2^2 \end{pmatrix}$$

$$D_2 = \det\begin{pmatrix} n & \Sigma y & \Sigma x_2 \\ \Sigma x_1 & \Sigma x_1 y & \Sigma x_1 x_2 \\ \Sigma x_2 & \Sigma x_2 y & \Sigma x_2^2 \end{pmatrix}$$

$$D_3 = \det\begin{pmatrix} n & \Sigma x_1 & \Sigma y \\ \Sigma x_1 & \Sigma x_1^2 & \Sigma x_1 y \\ \Sigma x_2 & \Sigma x_1 x_2 & \Sigma x_2 y \end{pmatrix}$$

$$a = \frac{D_1}{D} \qquad b = \frac{D_2}{D} \qquad c = \frac{D_3}{D}$$

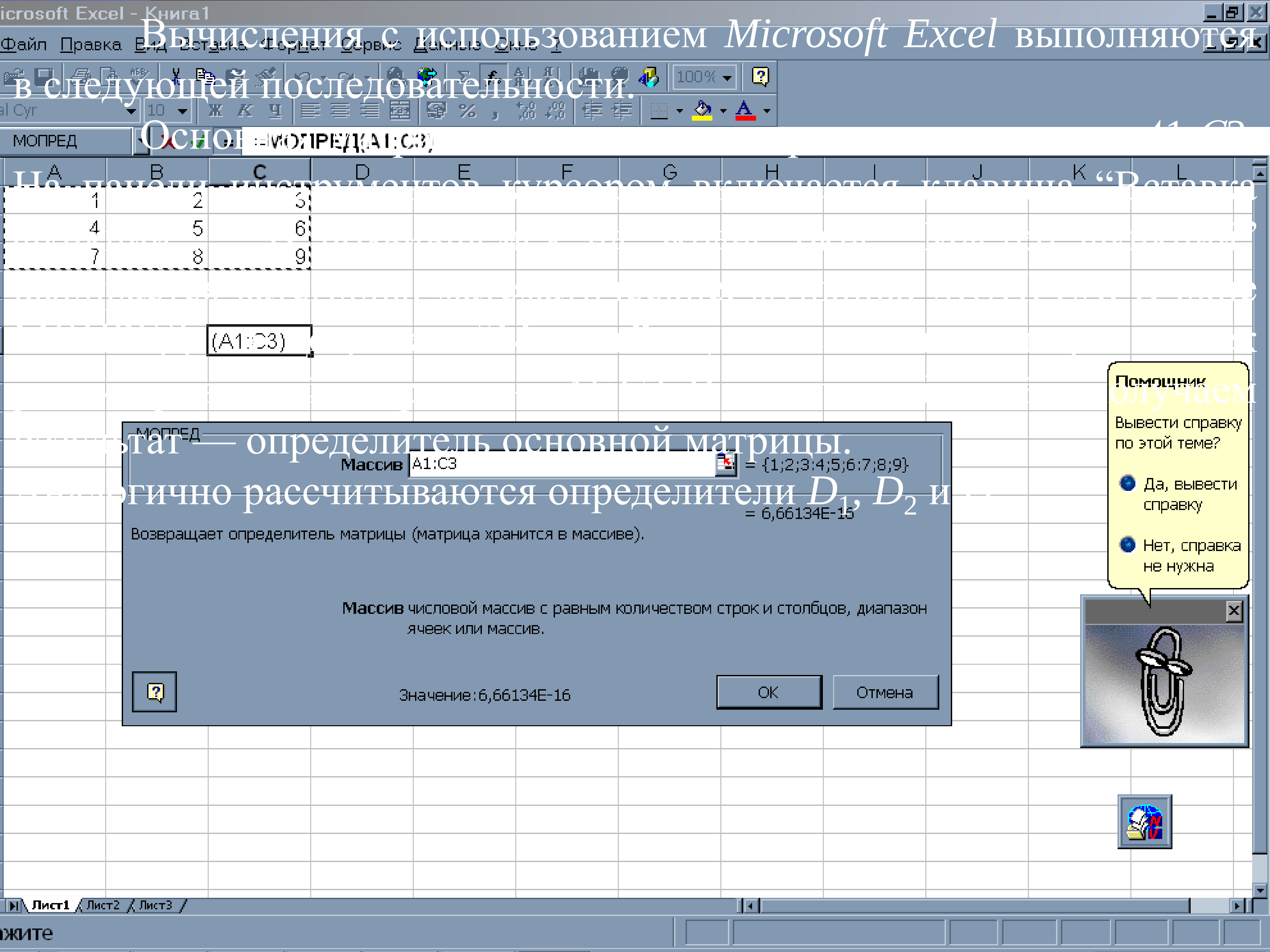
Вычисления с использованием *Microsoft Excel* выполняются в следующей последовательности.

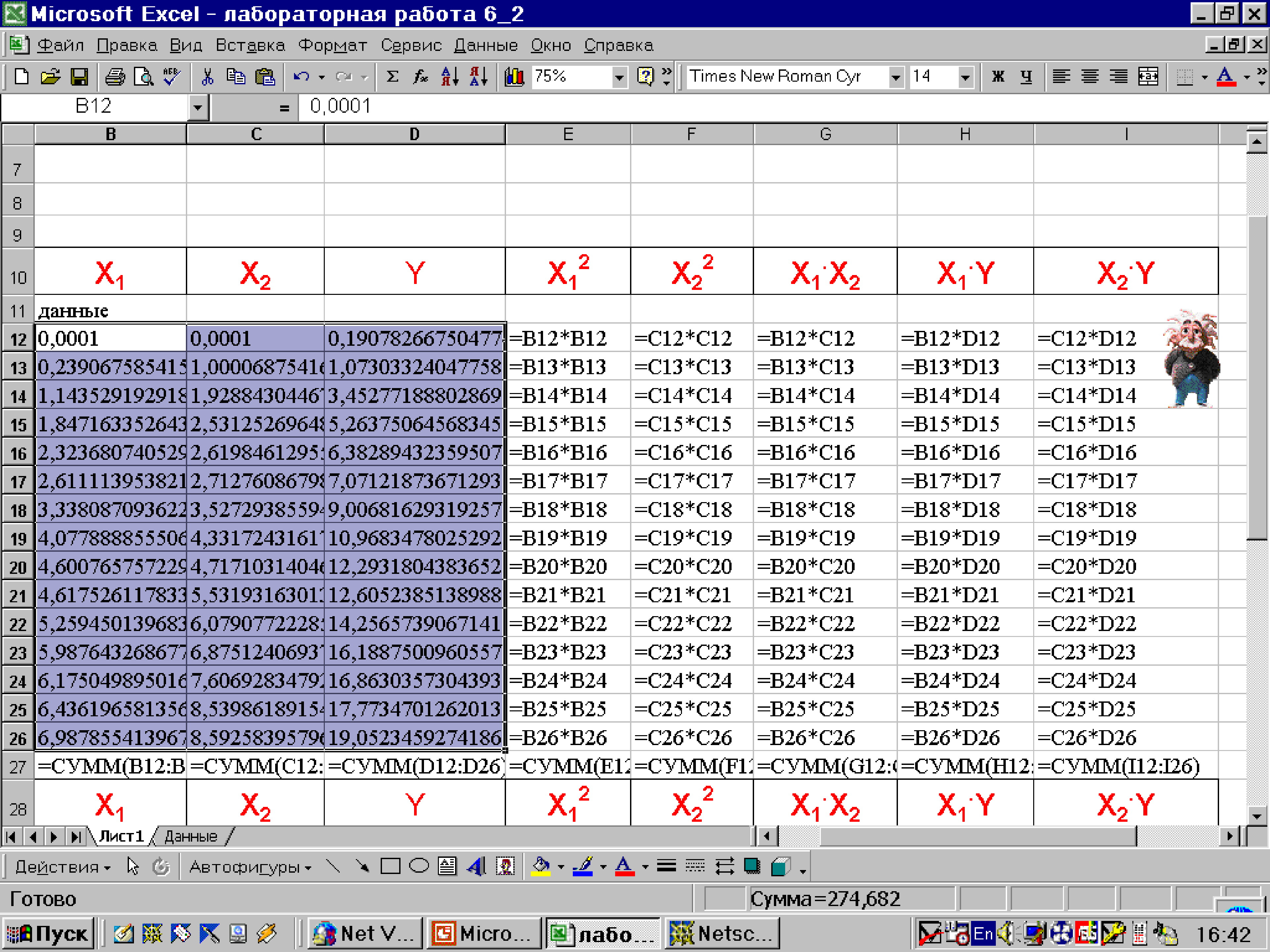
Основ… A1:C3

На панели инструментов курсором включается клавиша “Вставка

…олучаем …льтат — определитель основной матрицы.

…огично рассчитываются определители D_1, D_2 и





	B	C	D	E	F	G	H	I
7								
8								
9								
10	X_1	X_2	Y	X_1^2	X_2^2	$X_1 \cdot X_2$	$X_1 \cdot Y$	$X_2 \cdot Y$
11	данные							
12	0,0001	0,0001	0,19078266750477	=B12*B12	=C12*C12	=B12*C12	=B12*D12	=C12*D12
13	0,239067585415	1,0000687541	1,07303324047758	=B13*B13	=C13*C13	=B13*C13	=B13*D13	=C13*D13
14	1,143529192918	1,9288430446	3,45277188802869	=B14*B14	=C14*C14	=B14*C14	=B14*D14	=C14*D14
15	1,847163352643	2,5312526964	5,26375064568345	=B15*B15	=C15*C15	=B15*C15	=B15*D15	=C15*D15
16	2,323680740529	2,6198461295	6,38289432359507	=B16*B16	=C16*C16	=B16*C16	=B16*D16	=C16*D16
17	2,611139538821	2,7127608679	7,07121873671293	=B17*B17	=C17*C17	=B17*C17	=B17*D17	=C17*D17
18	3,338087093622	3,5272938559	9,00681629319257	=B18*B18	=C18*C18	=B18*C18	=B18*D18	=C18*D18
19	4,077888555506	4,3317243161	10,9683478025292	=B19*B19	=C19*C19	=B19*C19	=B19*D19	=C19*D19
20	4,600765757229	4,7171031404	12,2931804383652	=B20*B20	=C20*C20	=B20*C20	=B20*D20	=C20*D20
21	4,617526117833	5,5319316301	12,6052385138988	=B21*B21	=C21*C21	=B21*C21	=B21*D21	=C21*D21
22	5,259450139683	6,0790772228	14,2565739067141	=B22*B22	=C22*C22	=B22*C22	=B22*D22	=C22*D22
23	5,987643268677	6,8751240693	16,1887500960557	=B23*B23	=C23*C23	=B23*C23	=B23*D23	=C23*D23
24	6,175049895016	7,6069283479	16,8630357304393	=B24*B24	=C24*C24	=B24*C24	=B24*D24	=C24*D24
25	6,436196581356	8,5398618915	17,7734701262013	=B25*B25	=C25*C25	=B25*C25	=B25*D25	=C25*D25
26	6,987855413967	8,5925839579	19,0523459274186	=B26*B26	=C26*C26	=B26*C26	=B26*D26	=C26*D26
27	=СУММ(B12:B	=СУММ(C12:	=СУММ(D12:D26)	=СУММ(E12	=СУММ(F1	=СУММ(G12:	=СУММ(H12:	=СУММ(I12:I26)
28	X_1	X_2	Y	X_1^2	X_2^2	$X_1 \cdot X_2$	$X_1 \cdot Y$	$X_2 \cdot Y$

Microsoft Excel - лабораторная работа 6_2

Файл Правка Вид Вставка Формат Сервис Данные Окно Справка

65% Times New Roman Cyr 14

B12 = 0,0001

	A	B	C	D	E	F	G	H	I
27		55,645	66,594	152,44221	279,1545	398,26	332,09577	760,35	905,7
28	Σ	X_1	X_2	Y	X_1^2	X_2^2	$X_1 \cdot X_2$	$X_1 \cdot Y$	$X_2 \cdot Y$

$$D = \det \begin{pmatrix} n & \Sigma x_1 & \Sigma x_2 \\ \Sigma x_1 & \Sigma x_1^2 & \Sigma x_1 x_2 \\ \Sigma x_2 & \Sigma x_1 x_2 & \Sigma x_2^2 \end{pmatrix}$$

15	55,6451	66,5945	
55,6451	279,155	332,096	3427,27
66,5945	332,096	398,259	

$$a = \frac{D_1}{D}$$ 0,19052

$$D_1 = \det \begin{pmatrix} \Sigma y & \Sigma x_1 & \Sigma x_2 \\ \Sigma x_1 y & \Sigma x_1^2 & \Sigma x_1 x_2 \\ \Sigma x_2 y & \Sigma x_1 x_2 & \Sigma x_2^2 \end{pmatrix}$$

152,442	55,6451	66,5945	
760,353	279,155	332,096	652,965
905,697	332,096	398,259	

$$b = \frac{D_2}{D}$$ 2,28613

$$D_2 = \det \begin{pmatrix} n & \Sigma y & \Sigma x_2 \\ \Sigma x_1 & \Sigma x_1 y & \Sigma x_1 x_2 \\ \Sigma x_2 & \Sigma x_2 y & \Sigma x_2^2 \end{pmatrix}$$

15	152,442	66,5945	
55,6451	760,353	332,096	7835,18
66,5945	905,697	398,259	

$$c = \frac{D_3}{D}$$ 0,33595

$$D_3 = \det \begin{pmatrix} n & \Sigma x_1 & \Sigma y \\ \Sigma x_1 & \Sigma x_1^2 & \Sigma x_1 y \\ \Sigma x_2 & \Sigma x_1 x_2 & \Sigma x_2 y \end{pmatrix}$$

15	55,6451	152,442	
55,6451	279,155	760,353	1151,39
66,5945	332,096	905,697	

Лист1 / Данные

Действия Автофигуры

Готово Сумма=274,682

Пуск Net V... Micro... лабо... ERR... En 16:40