Министерство образования Республики Беларусь

УО «Витебский государственный технологический университет»

Тема4. «РЕГРЕССИОННЫЙ АНАЛИЗ.»

Кафедра теоретической и прикладной математики.

разработана доц. Е.Б.Дуниной

4.1 Измерения и их погрешности.

Под измерениями понимают сравнение некоторой величины с другой (однородной) величиной, принятой за единицу.

Различают прямые и косвенные измерения.

При прямых измерениях исследуемая величина сравнивается с единицей измерения непосредственно или с помощью измерительного прибора (например, измерение длин линейкой)

При косвенных измерениях величины ее значения определяются по результатам прямых измерений других величин, связанных с рассматриваемой величиной заданной функциональной зависимостью (например, измерение плотности тела по измерениям его массы и объема).

В результате измерений получают приближенные значения величины, а не ее точное значение.

При измерениях погрешности неизбежны.

Погрешностью, или ошибкой, измерения называется разность $x - a$

между результатом измерения *x* и точным значением a **измеряемой величины.**

Погрешности измерений обычно неизвестны, т.к. неизвестно и точное значение измеряемой величины.

Различают следующие виды погрешностей измерений:

грубые ошибки – ошибки, сделанные вследствие неверной записи показаний прибора, неправильно прочитанного отсчета и т.д.;

систематические погрешности – погрешности, связанные с ограниченной точностью прибора, неправильной его установкой и т.д.

Систематические ошибки можно устранить путем введения соответствующих поправок.

Случайные погрешности вызываются большим числом случайных причин, действие которых на каждое измерение различно и заранее не может быть учтено.

Случайные ошибки являются неустранимыми.

В качестве закона распределения случайных погрешностей измерения чаще всего принимается нормальный закон распределения.

4.2 Метод наименьших квадратов.

Основная цель *регрессионного анализа* состоит в определении связи между некоторой характеристикой *Y* наблюдаемого явления или объекта и величинами x_1, x_2, …, x_n, которые обусловливают, объясняют изменения *Y*. Переменная *Y* называется *зависимой переменной* (*откликом*), влияющие переменные x_1, x_2, …, x_n называются *факторами* (регрессорами).

При измерении двух величин *x* и *y* получены следующие данные

x	x_1	x_2	$\cdots$	x_n
y	y_1	y_2	$\cdots$	y_n

Известен также вид функциональной зависимости $y = f(x).$

Данная функция зависит от некоторых параметров значения которых требуется определить.

Значения *y*, полученные из формулы при заданных значениях $x_i,$ **как правило, не совпадают с экспериментальными значениями, приведенными в таблице.**

Разности

$$f(x_i) - y_i = \varepsilon_i, \quad i = 1,\ldots,n$$

называются уклонениями или погрешностями.

Требуется подобрать параметры функции *f(x)* так, чтобы уклонения ε_i оказались наименьшими.

Критерий, лежащий в основе метода наименьших квадратов:

параметры функции *f(x)* выбирают так, чтобы оказалась минимальной сумма квадратов уклонений:

$$\sum_{i=1}^{n} \varepsilon_i^2 = \varepsilon_1^2 + \varepsilon_2^2 + \ldots + \varepsilon_n^2 \geq 0.$$

!

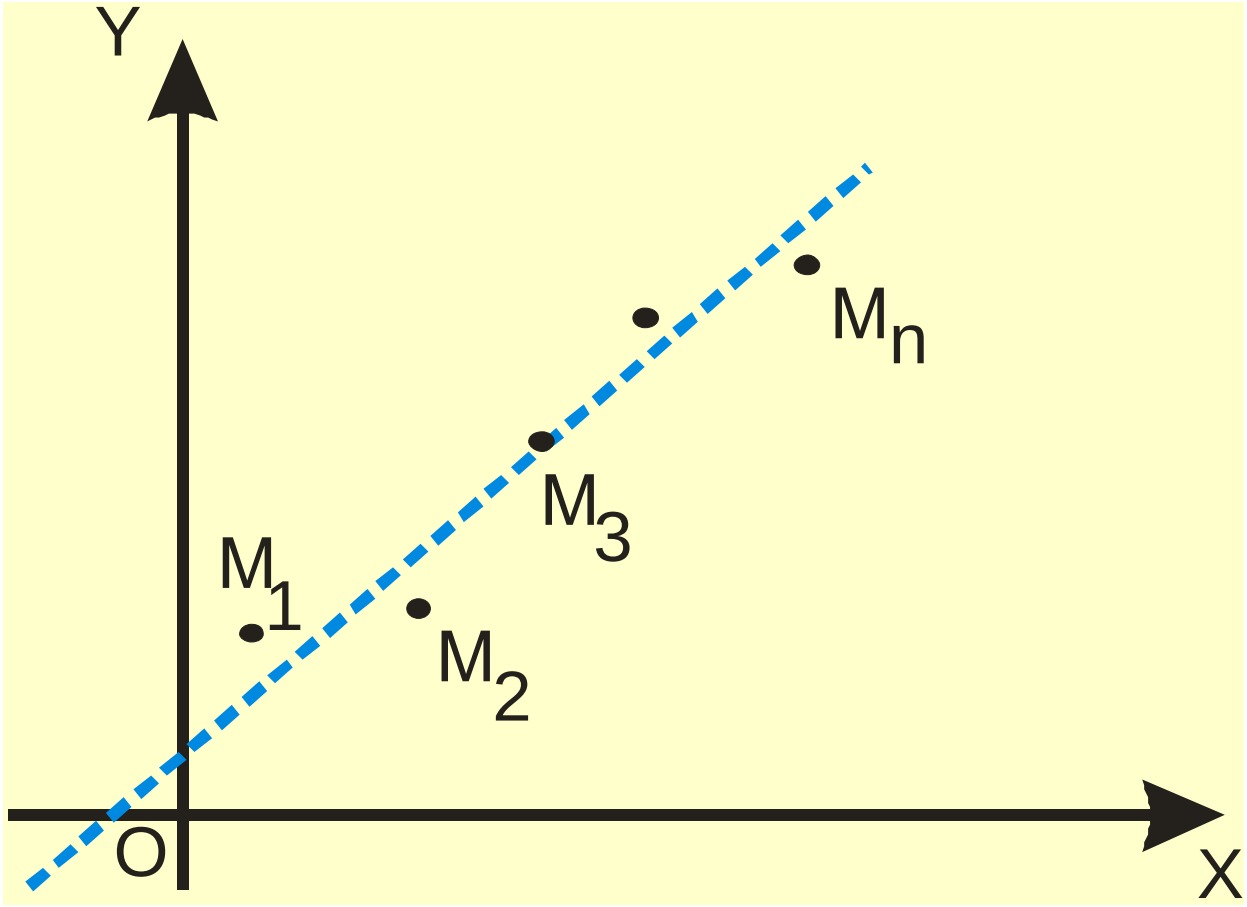
4.3 Определение параметров эмпирических формул в случае линейной зависимости.

Пусть необходимо установить зависимость между двумя величинами *x* и *y*, результаты измерений которых занесены в таблицу

x	x_1	x_2	$\cdots$	x_n
y	y_1	y_2	$\cdots$	y_n

Рассмотрим в декартовой системе координат *xOy* точки

$$M_1(x_1, y_1); M_2(x_2, y_2), \ldots, M_n(x_n, y_n)$$



Если эти точки почти лежат на некоторой прямой, естественно предположить, что между *x* и *y* существует линейная зависимость, выражающаяся формулой

$$y = ax + b, \quad (7.3)$$

где a, b **- некоторые постоянные коэффициенты (параметры), подлежащие определению.**

Запишем выражение (7.3) в виде

$$ax + b - y = 0. \quad (7.4)$$

Подставляя координаты точек

x_i и y_i $i = 1,\ldots,n$

в формулу (7.4) получим

$$\begin{cases} ax_1 + b - y_1 = \varepsilon_1 \\ ax_2 + b - y_2 = \varepsilon_2 \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ ax_n + b - y_n = \varepsilon_n \end{cases} \quad (7.5)$$

где $\varepsilon_1, \varepsilon_2, \ldots \varepsilon_n$ **- уклонения.**

По методу наименьших квадратов подберем неизвестные параметры a, b

таким образом, чтобы полученная величина

$$U = \sum_{i=1}^{n} \varepsilon_i^2 \geq 0$$

была наименьшей.

С учетом (7.5) имеем

$$U = \sum_{i=1}^{n} \varepsilon_i^2 = (ax_1 + b - y_1)^2 +$$

$$+ (ax_2 + b - y_2)^2 + \ldots + (ax_n + b - y_n)^2. \quad (7.6)$$

Переменная величина U

является функцией двух переменных $a \quad \text{и} \quad b.$

Чтобы функция U

получила возможно меньшее значение при выбранных $a \quad \text{и} \quad b,$

необходимо выполнение условия

$$\frac{\partial U}{\partial a} = 0; \frac{\partial U}{\partial b} = 0. \qquad (7.7)$$

Найдем частные производные

$$\frac{\partial U}{\partial a} = \left[(ax_1 + b - y_1)^2 + (ax_2 + b - y_2)^2 + \ldots + (ax_n + b - y_n)^2\right]'_a =$$

$$= 2(ax_1 + b - y_1)x_1 + 2(ax_2 + b - y_2)x_2 + \ldots + 2(ax_n + b - y_n)x_n =$$

$$= 2a\sum_{i=1}^{n} x_i^2 + 2b\sum_{i=1}^{n} x_i - 2\sum_{i=1}^{n} x_i y_i .$$

$$\frac{\partial U}{\partial b} = \left[(ax_1 + b - y_1)^2 + (ax_2 + b - y_2)^2 + \ldots + (ax_n + b - y_n)^2\right]'_b =$$

$$= 2(ax_1 + b - y_1) + 2(ax_2 + b - y_2) + \ldots + 2(ax_n + b - y_n) =$$

$$= 2a\sum_{i=1}^{n} x_i + 2b \cdot n - 2\sum_{i=1}^{n} y_i .$$

С учетом (7.7) получим

$$\begin{cases} a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i, \\ a\sum_{i=1}^{n} x_i + bn = \sum_{i=1}^{n} y_i. \end{cases} \qquad (7.8)$$

Решив систему определяют оптимальные значения неизвестных параметров a и b.

Замечание.

Эмпирическая формула $y = ax + b$

неплохо отражает вид зависимости между *x* и *y* в случае, когда первые разделенные разности

$$f(x_2, x_1) = \frac{y_2 - y_1}{x_2 - x_1}$$

$$f(x_3, x_2) = \frac{y_3 - y_2}{x_3 - x_2}$$

,…, $$f(x_n, x_{n-1}) = \frac{y_n - y_{n-1}}{x_n - x_{n-1}}.$$

мало отличаются друг от друга.

Если таблица результатов имеет постоянный шаг, то достаточно сравнивать неразделенные разности

$$\Delta y_1 = y_2 - y_1,$$

$$\Delta y_2 = y_3 - y_2,$$

…

$$\Delta y_{n-1} = y_n - y_{n-1}.$$

Пример.

Результаты измерений величин *x* и *y* представлены в виде таблицы

X	-2	-1	0	1	2	3
Y	5,6	5	4,3	4	3,6	3

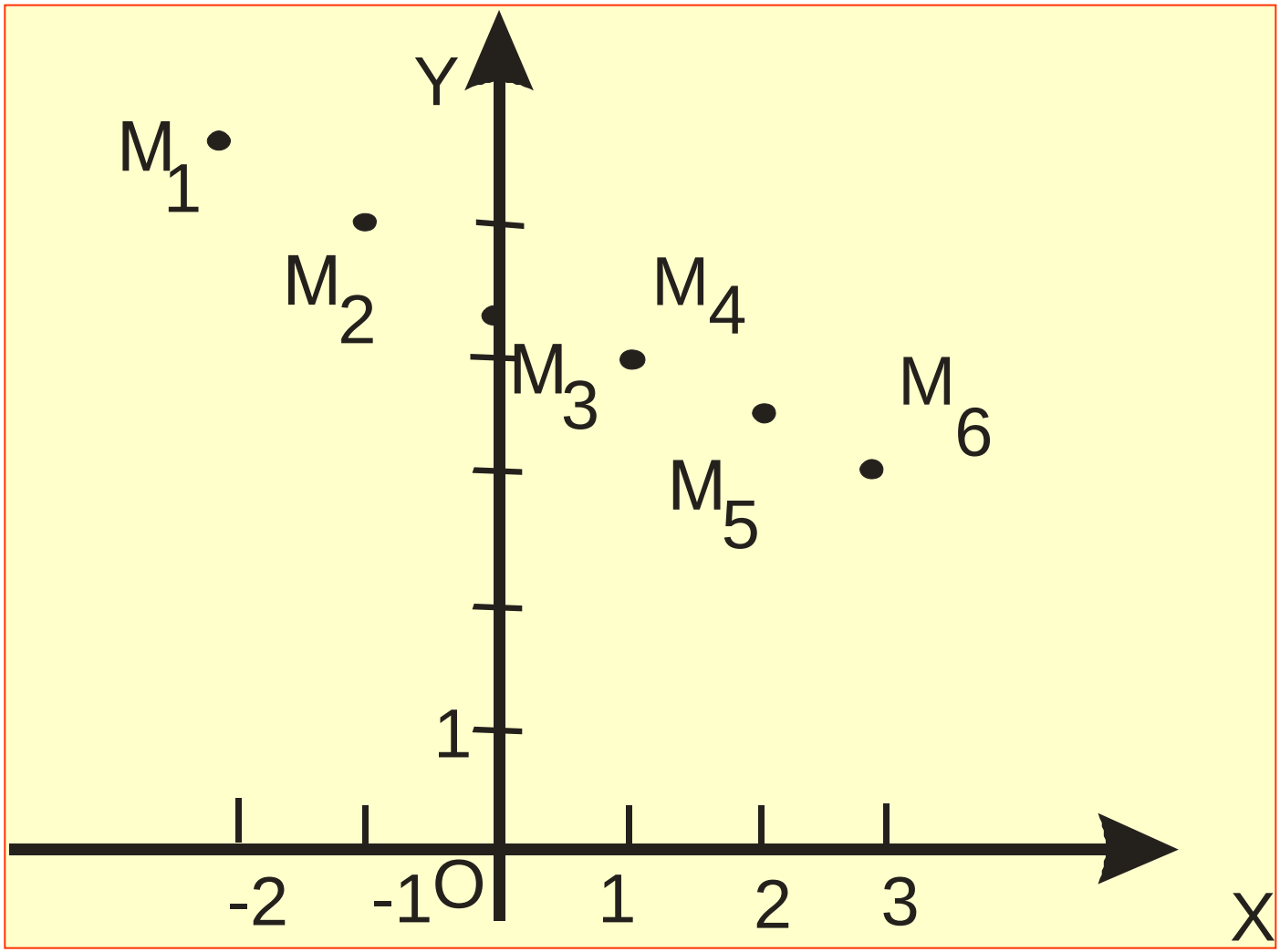
Установить вид зависимости между этими величинами и найти параметры эмпирических формул.

Решение.

Построим в декартовой системе координат точки

$M_1(-2;5,6), M_2(-1;5), M_3(0;4,3), M_4(1;4),$

$M_5(2;3,6), M_6(3,3).$



Эти точки располагаются приблизительно на одной прямой.

Можно предположить существование между *x* и *y* линейной зависимости.

Уравнение прямой запишем в виде

$$y = ax + b.$$

Определим параметры a и b

используя систему (7.8).

Для упрощения расчетов составим таблицу

i	x_i	y_i	$x_i y_i$	x_i^2
1	-2	5,6	-11,2	4
2	-1	5	-5	1
3	0	4,3	0	0
4	1	4	4	1
5	2	3,6	7,2	4
6	3	3	9	9
Σ	3	25,5	4	19

С учетом таблицы система (7.8) примет вид

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

$$\begin{cases} 19a + 3b = 4 \\ 3a + 6b = 25{,}5 \end{cases}$$

Решаем систему любым методом, например, Крамера

$$\Delta = \begin{vmatrix} 19 & 3 \\ 3 & 6 \end{vmatrix} = 19 \cdot 6 - 3 \cdot 3 = 105$$

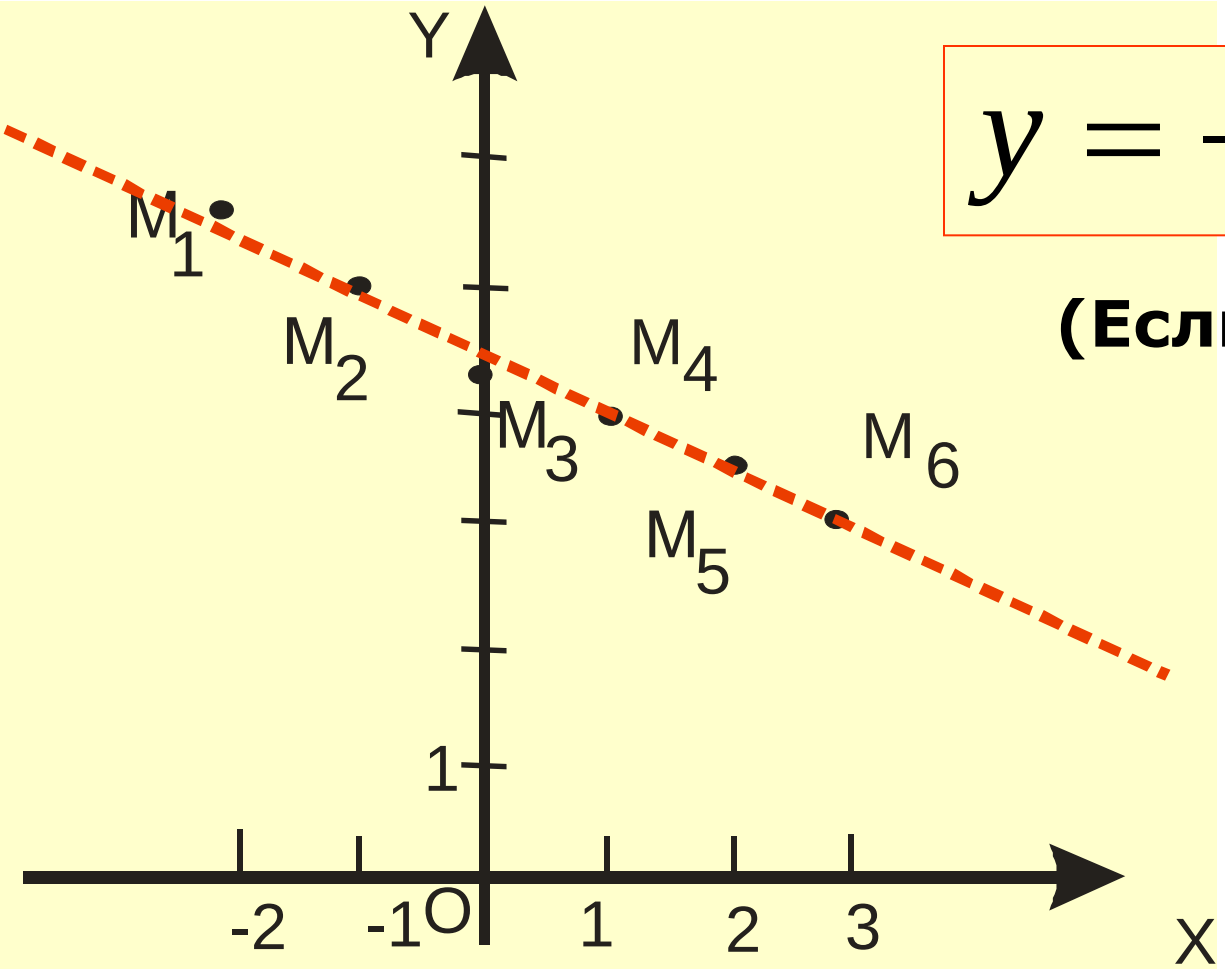
$$\Delta_a = \begin{vmatrix} 4 & 3 \\ 25{,}5 & 6 \end{vmatrix} = 4 \cdot 6 - 3 \cdot 25{,}5 = 24 - 76{,}5 = -52{,}5$$

$$\Delta_b = \begin{vmatrix} 19 & 4 \\ 3 & 25{,}5 \end{vmatrix} = 19 \cdot 25{,}5 - 3 \cdot 4 = 484{,}5 - 12 = 472{,}5$$

$$a = \frac{\Delta_a}{\Delta} = \frac{-52{,}5}{105} = -0{,}5$$

$$b = \frac{\Delta_b}{\Delta} = \frac{472{,}5}{105} = 4{,}5$$

Следовательно, зависимость между величинами *x* и *y* выражается формулой



$$y = -0{,}5x + 4{,}5.$$

(Если $x = 0, y = 4{,}5;$

$x = 3, y = 3)$

4.4 Определение параметров эмпирических формул в случае квадратичной зависимости.

Пусть в результате измерений двух зависимых величин *x* и *y* получена следующая таблица

x	x_1	x_2	$\cdots$	x_n
y	y_1	y_2	$\cdots$	y_n

Предположим, что точки

$$M_1(x_1, y_1); M_2(x_2, y_2), \ldots, M_n(x_n, y_n)$$

почти лежат на некоторой параболе.

В этом случае можно предположить, что между *x* и *y* существует квадратичная зависимость

$$y = ax^2 + bx + c \quad (7.9)$$

Запишем выражение (7.9) в виде

$$ax^2 + bx + c - y = 0. \quad (7.10)$$

Подставляя координаты точек

x_i и y_i $\quad i = 1,\ldots,n$

в формулу (7.10) получим

$$\begin{cases} ax_1^2 + bx_1 + c - y_1 = \varepsilon_1 \\ ax_2^2 + bx_2 + c - y_2 = \varepsilon_2 \\ \quad \ldots\ldots\ldots\ldots\ldots\ldots \\ ax_n^2 + bx_n + c - y_n = \varepsilon_n \end{cases} \quad (7.11)$$

По методу наименьших квадратов подберем неизвестные параметры таким образом, чтобы полученная величина

$$U = \sum_{i=1}^{n} \varepsilon_i^2 \geq 0$$

была наименьшей.

С учетом (7.11) имеем

$$U = \sum_{i=1}^{n} \varepsilon_i^2 =$$

$$= (ax_1^2 + bx_1 + c - y_1)^2 + (ax_2^2 + bx_2 + c - y_2)^2 +$$

$$+ \ldots + (ax_n^2 + bx_n + c - y_n)^2. \qquad (7.12)$$

Переменная величина U

является функцией трех переменных a , b и c.

Чтобы функция U **получила возможно меньшее значение необходимо выполнение условия**

$$\frac{\partial U}{\partial a}=0;\frac{\partial U}{\partial b}=0;\frac{\partial U}{\partial c}=0. \quad (7.13)$$

Найдем частные производные

$$\frac{\partial U}{\partial a}=[(ax_1^2+bx_1+c-y_1)^2+(ax_2^2+bx_2+c-y_2)^2+$$

$$+\ldots+(ax_n^2+bx_n+c-y_n)^2]'_a=$$

$$= 2(ax_1^2 + bx_1 + c - y_1)x_1^2 + 2(ax_2^2 + bx_2 + c - y_2)x_2^2 +$$

$$+ \ldots + 2(ax_n^2 + bx_n + c - y_n)x_n^2 =$$

$$= 2a\sum_{i=1}^{n} x_i^4 + 2b\sum_{i=1}^{n} x_i^3 + 2c\sum_{i=1}^{n} x_i^2 - 2\sum_{i=1}^{n} x_i^2 y_i .$$

$$\frac{\partial U}{\partial b} = [(ax_1^2 + bx_1 + c - y_1)^2 + (ax_2^2 + bx_2 + c - y_2)^2 +$$

$$+ \ldots + (ax_n^2 + bx_n + c - y_n)^2]'_b =$$

$$= 2(ax_1^2 + bx_1 + c - y_1)x_1 + 2(ax_2^2 + bx_2 + c - y_2)x_2 +$$

$$+ \ldots + 2(ax_n^2 + bx_n + c - y_n)x_n =$$

$$= 2a\sum_{i=1}^{n} x_i^3 + 2b\sum_{i=1}^{n} x_i^2 + 2c\sum_{i=1}^{n} x_i - 2\sum_{i=1}^{n} x_i y_i .$$

$$\frac{\partial U}{\partial c} = [(ax_1^2 + bx_1 + c - y_1)^2 + (ax_2^2 + bx_2 + c - y_2)^2 +$$

$$+ \ldots + (ax_n^2 + bx_n + c - y_n)^2]'_c =$$

$$= 2(ax_1^2 + bx_1 + c - y_1) + 2(ax_2^2 + bx_2 + c - y_2) +$$

$$+ \ldots + 2(ax_n^2 + bx_n + c - y_n) =$$

$$= 2a\sum_{i=1}^{n} x_i^2 + 2b\sum_{i=1}^{n} x_i + 2c \cdot n - 2\sum_{i=1}^{n} y_i .$$

С учетом (7.13) получим

$$\begin{cases} a\sum_{i=1}^{n} x_i^4 + b\sum_{i=1}^{n} x_i^3 + c\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i^2 y_i. \\ a\sum_{i=1}^{n} x_i^3 + b\sum_{i=1}^{n} x_i^2 + c\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i. \\ a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i + c\cdot n = \sum_{i=1}^{n} y_i. \end{cases} \quad (7.14)$$

Решив систему определяют оптимальные значения неизвестных параметров a, b и c.

Пример.

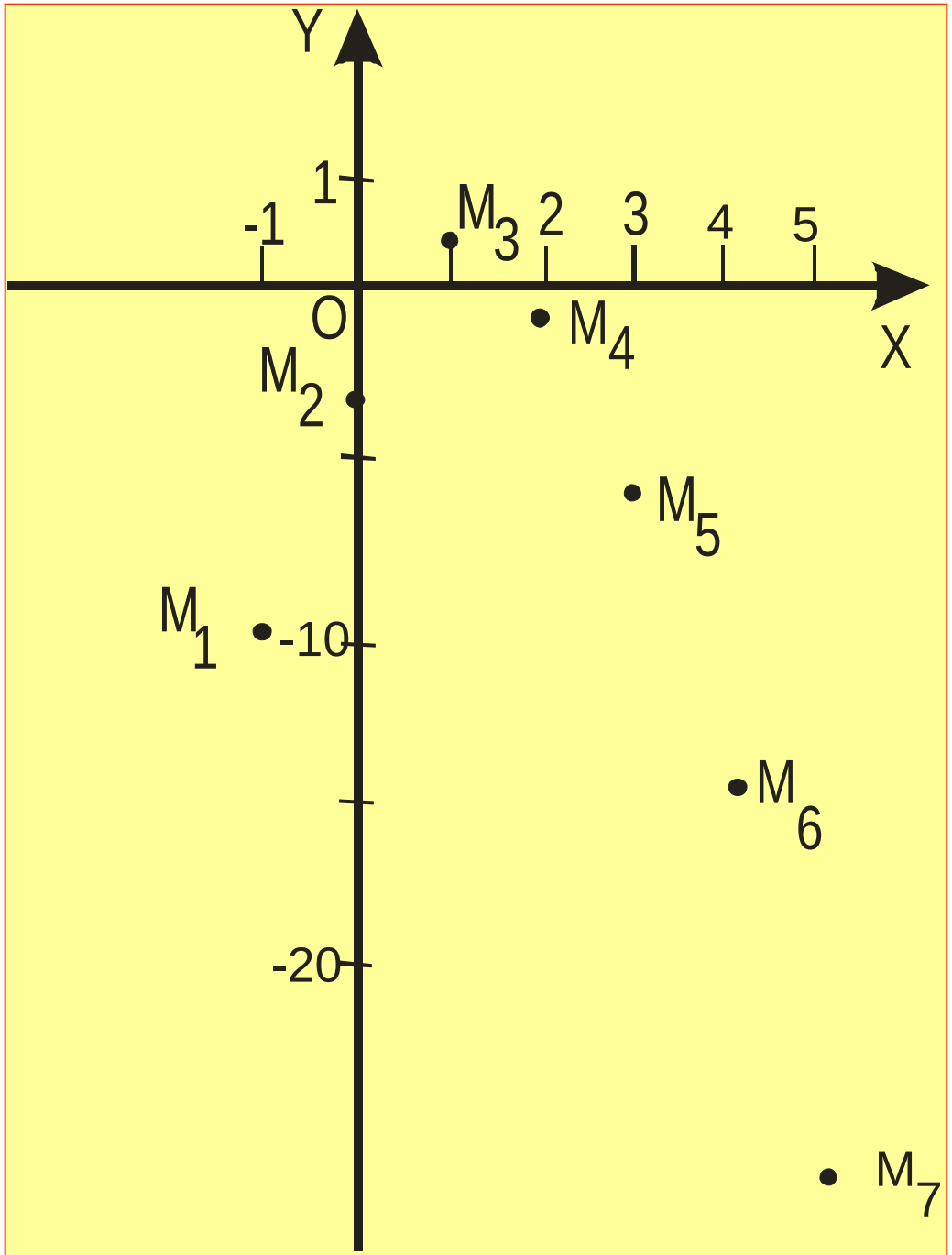
Результаты измерений величин *x* и *y* представлены в виде таблицы

X	-1	0	1	2	3	4	5
Y	-9,8	-3.1	0,3	-1,2	-6,1	-14,7	-28,2

Установить вид зависимости между этими величинами и найти параметры эмпирических формул.

Решение

Построим в декартовой системе координат точки



$M_1(-1;-9{,}8), M_2(0;-3{,}1),$

$M_3(1;0{,}3), M_4(2;-1{,}2),$

$M_5(3;-6{,}1), M_6(4,-14{,}7),$

$M_7(5;-28{,}2).$

Составим таблицу

i	x_i	x_i^2	x_i^3	x_i^4	y_i	$x_i y_i$	$x_i^2 y_i$
1	**-1**	**1**	**-1**	**1**	**-9,8**	**9,8**	**-9,8**
2	**0**	**0**	**0**	**0**	**-3,1**	**0**	**0**
3	**1**	**1**	**1**	**1**	**0,3**	**0,3**	**0,3**
4	**2**	**4**	**8**	**16**	**-1,2**	**-2,4**	**-4,8**
5	**3**	**9**	**27**	**81**	**-6,1**	**-18,3**	**-54,9**
6	**4**	**16**	**64**	**256**	**-14,8**	**-58,8**	**-233,2**
7	**5**	**25**	**125**	**625**	**-28,2**	**-141,0**	**-705,0**
$\sum$	**14**	**56**	**224**	**980**	**-62,8**	**-210,4**	**-1007,4**

Тогда система (7.14) принимает вид

$$\begin{cases} a\sum_{i=1}^{n} x_i^4 + b\sum_{i=1}^{n} x_i^3 + c\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i^2 y_i. \\ a\sum_{i=1}^{n} x_i^3 + b\sum_{i=1}^{n} x_i^2 + c\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i. \\ a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i + c \cdot n = \sum_{i=1}^{n} y_i. \end{cases}$$

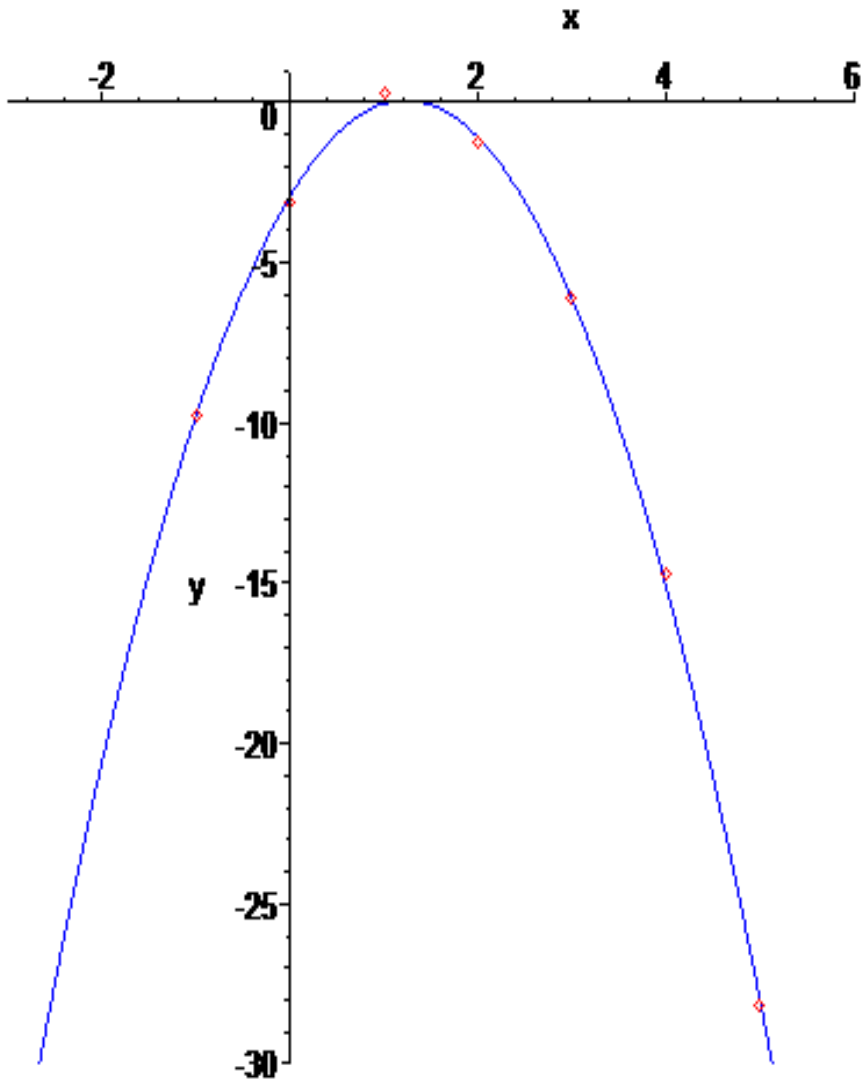
$$\begin{cases} 980a + 224b + 56c = -1007{,}4 \\ 224a + 56b + 14c = -210{,}4 \\ 56a + 14b + 7c = -62{,}8 \end{cases}$$

Решая систему получи

$$a = -1{,}974,$$

$$b = 4{,}867,$$

$$c = -2{,}914$$



Искомая зависимость между *x* и *y* имеет вид

$$y = -1{,}974x^2 + 4{,}867x - 2{,}914.$$

4.5 Определение параметров эмпирических формул для гиперболической и показательной зависимостей.

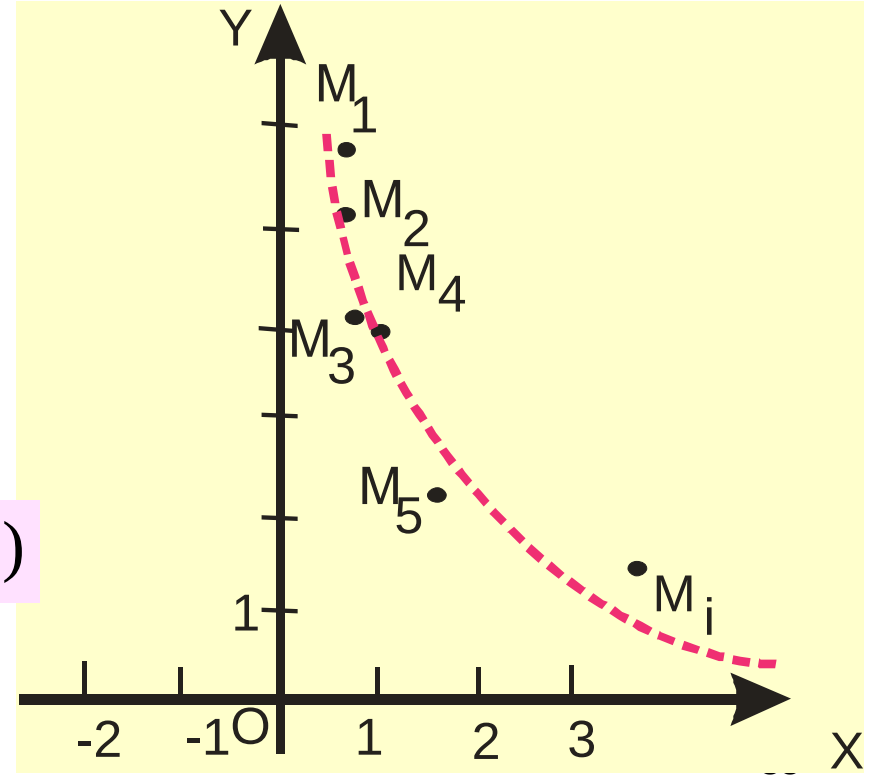
Пусть в результате измерений двух зависимых величин *x* и *y* получена следующая таблица

x	x_1	x_2	$\cdots$	x_n
y	y_1	y_2	$\cdots$	y_n

Предположим, что точки

$$M_1(x_1, y_1); M_2(x_2, y_2), \ldots, M_n(x_n, y_n)$$

почти лежат на некоторой гиперболе



$$y = \frac{a}{x} + b, \quad (7.15)$$

где a и b **- параметры, подлежащие определению.**

Введем новую переменную $z = \frac{1}{x},$

тогда уравнение (7.15) будет иметь вид

$$y = az + b,$$

т.е. y и z **связаны линейной зависимостью с параметрами** a и $b.$

В этом случае по методу наименьших квадратов параметры определяются из системы (7.8), где вместо x_i

произведена замена

$$z_i = \frac{1}{x_i}, i = 1,\ldots, n.$$

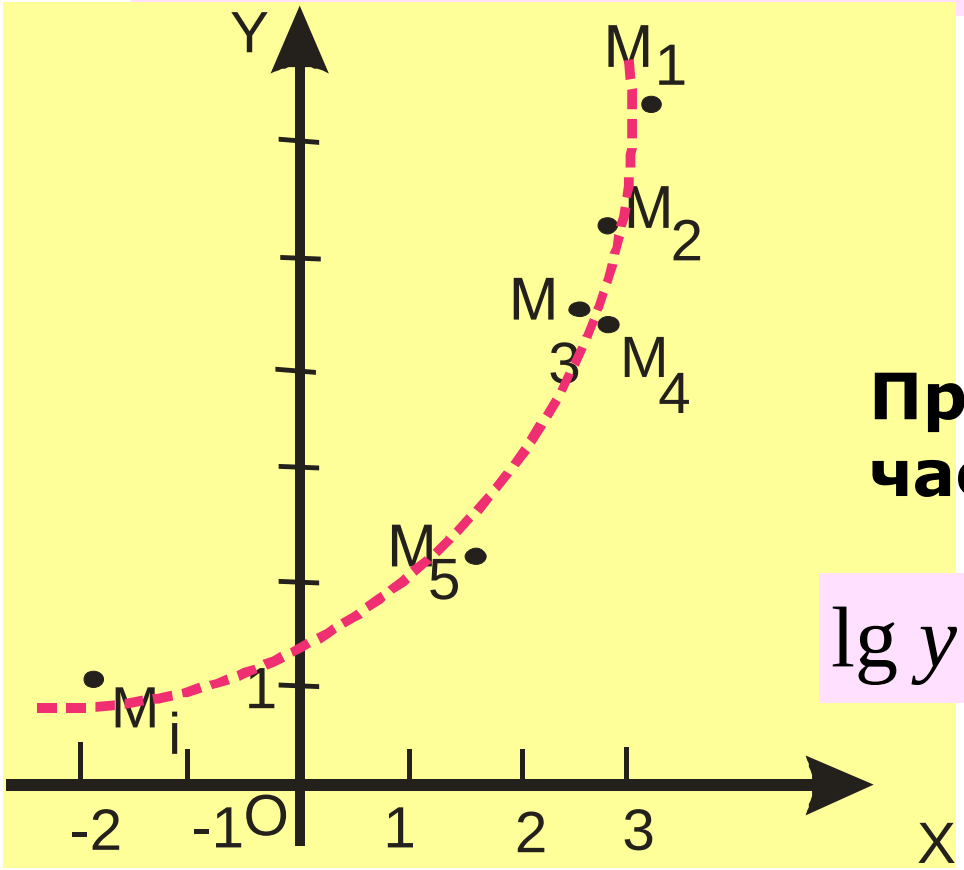
Таким образом получим систему

$$\begin{cases} a\sum_{i=1}^{n} z_i^2 + b\sum_{i=1}^{n} z_i = \sum_{i=1}^{n} z_i y_i. \\ a\sum_{i=1}^{n} z_i + bn = \sum_{i=1}^{n} y_i. \end{cases}$$

$$\begin{cases} a\sum_{i=1}^{n} \frac{1}{x_i^2} + b\sum_{i=1}^{n} \frac{1}{x_i} = \sum_{i=1}^{n} \frac{1}{x_i} y_i. \\ a\sum_{i=1}^{n} \frac{1}{x_i} + bn = \sum_{i=1}^{n} y_i. \end{cases} \quad (7.16)$$

Пусть анализ связи между *x* и *y* привел к выбору в качестве формы зависимости *y* от *x* показательной формы

$$y = b \cdot a^x, a > 0; a \neq 1. \quad (7.17)$$



Пролагорифмируем обе части уравнения (7.17)

$$\lg y = \lg(b \cdot a^x) = \lg b + \lg a^x,$$

$$\lg y = x \lg a + \lg b. \quad (7.18)$$

Переменные $\lg y$ и x

связаны линейной зависимостью с параметрами

$$\lg a \text{ и } \lg b.$$

Воспользуемся системой (7.8), в которой заменяем

$$a \text{ на } \lg a,\ \text{b на } \lg b,\ y_i \text{ на } \lg y_i.$$

Получаем систему

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

$$\begin{cases} \lg a\sum_{i=1}^{n} x_i^2 + \lg b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i \lg y_i. \\ \lg a\sum_{i=1}^{n} x_i + \lg b \cdot n = \sum_{i=1}^{n} \lg y_i. \end{cases} \quad (7.19)$$

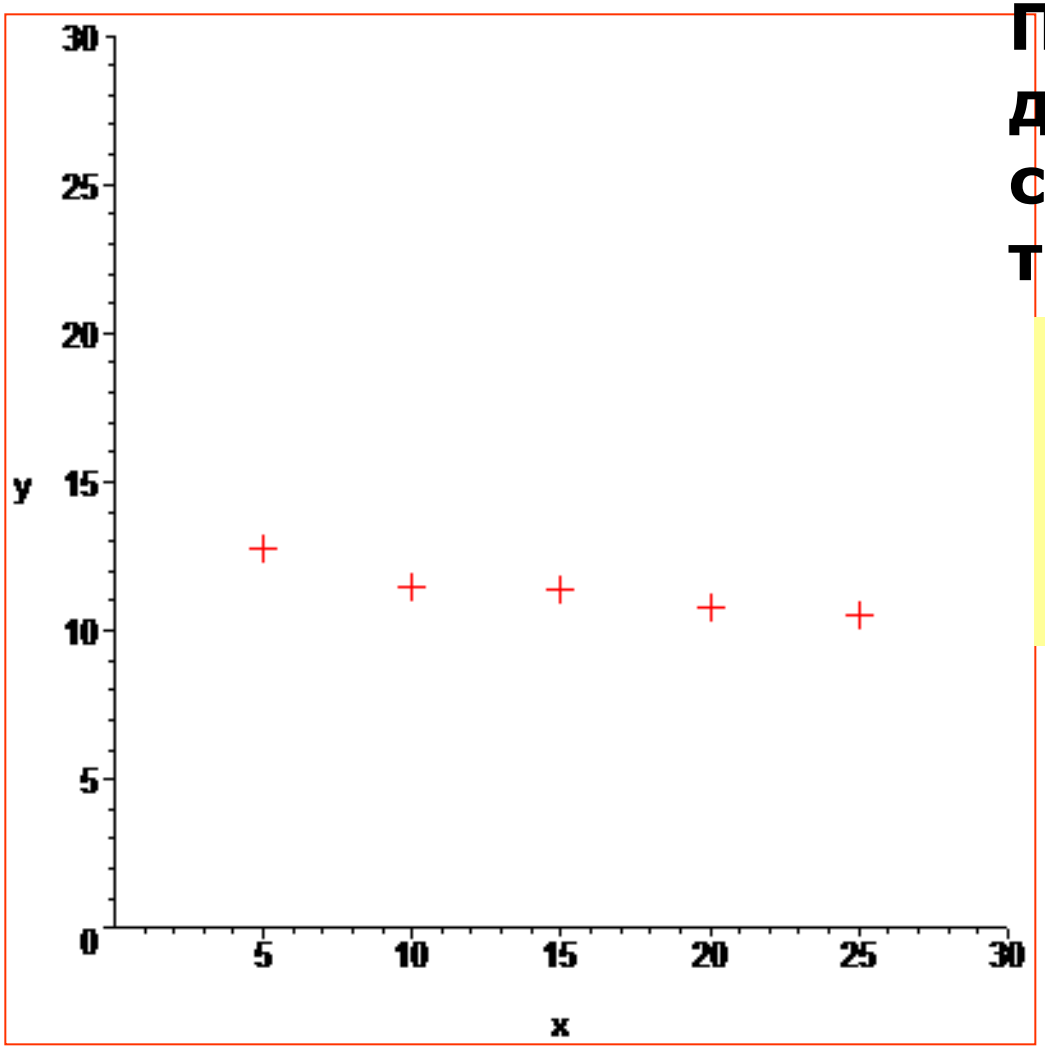
Решив систему (7.19), находим, $\lg a$ и $\lg b$,

а потом определяем параметры a и b.

Пример

Установить форму связи между зависимыми величинами *x* и *y* и найти уравнение этой зависимости

X	5	10	15	20	25
Y	12,8	11.5	11,4	10,8	10,5



Решение

Построим в декартовой системе координат точки

$M_1(5;12,8), M_2(10;11,5),$

$M_3(15;11,4),$

$M_4(20;10,8), M_5(25;10,5).$

Из рисунка видно, что между переменными существует гиперболическая зависимость

$$y = \frac{a}{x} + b.$$

i	x_i	$\frac{1}{x_i}$	$\frac{1}{x_i^2}$	y_i	$\frac{y_i}{x_i}$
1	5	0,2	0,04	12,8	2,56
2	10	0,1	0,01	11,5	1,15
3	15	0,067	0,0045	11,4	0,76
4	20	0,05	0,0025	10,8	0,54
5	25	0,04	0,0016	10,5	0,42
Σ	75	0,457	0,0586	57,0	5,43

Результаты расчетов приведем в таблице

Система уравнений (7.16)

$$\begin{cases} a\sum_{i=1}^{n}\frac{1}{x_i^2}+b\sum_{i=1}^{n}\frac{1}{x_i}=\sum_{i=1}^{n}\frac{1}{x_i}y_i. \\ a\sum_{i=1}^{n}\frac{1}{x_i}+bn=\sum_{i=1}^{n}y_i. \end{cases}$$

примет вид

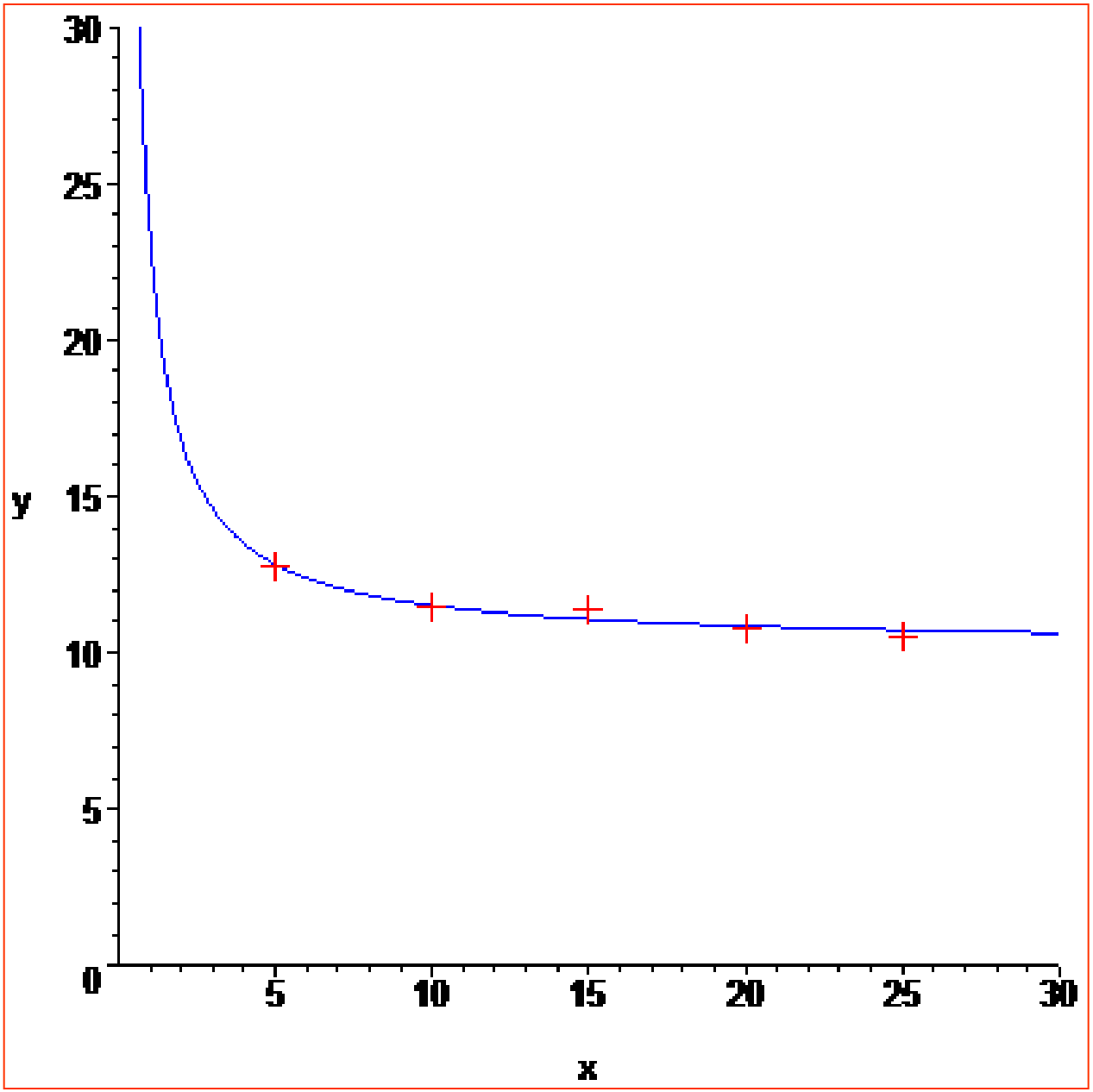
$$\begin{cases} 0{,}0586a+0{,}457b=5{,}43 \\ 0{,}457a+5b=57. \end{cases}$$

Решая ее, найдем

$$a=13{,}113;\quad b=10{,}2$$

и тогда уравнение примет вид

$$y=\frac{13{,}113}{x}+10{,}2.$$


30
25
20
y
15
10
5
0
5
10
15
20
25
30
x