МОДЕЛЬ ПАРНОЙ ЛИНЕЙНОЙ РЕГРЕССИИ

$$y = ax + b + \xi$$

y – зависимая (объясняемая) переменная
x – независимая (объясняющая) переменная
a b – неизвестные параметры модели

ξ - случайная составляющая

МОДЕЛЬ ПАРНОЙ ЛИНЕЙНОЙ РЕГРЕССИИ

$$y = ax + b + \xi$$

Предположим, что необходимо получить функцию спроса на некоторый товар в зависимости от дохода.

Проводится опрос домохозяйств.

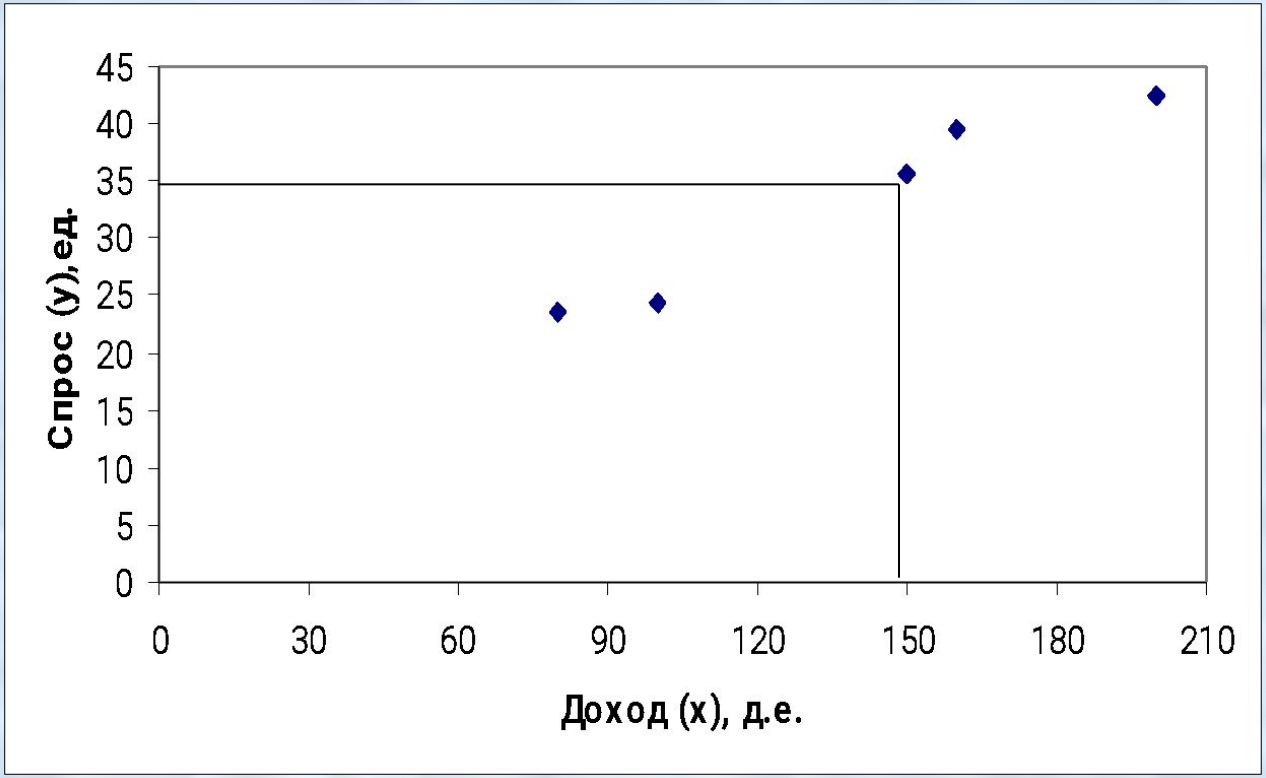
1. Среднедушевой доход домохозяйства?
2. Сколько единиц товара приобрело домохозяйство за месяц?

МОДЕЛЬ ПАРНОЙ ЛИНЕЙНОЙ РЕГРЕССИИ

№ домохозяйства	Среднедушевой доход домохозяйства, д.е.	Объем спроса, ед.
1	100	24
2	200	42
3	150	35
4	80	24
5	160	39

МОДЕЛЬ ПАРНОЙ ЛИНЕЙНОЙ РЕГРЕССИИ

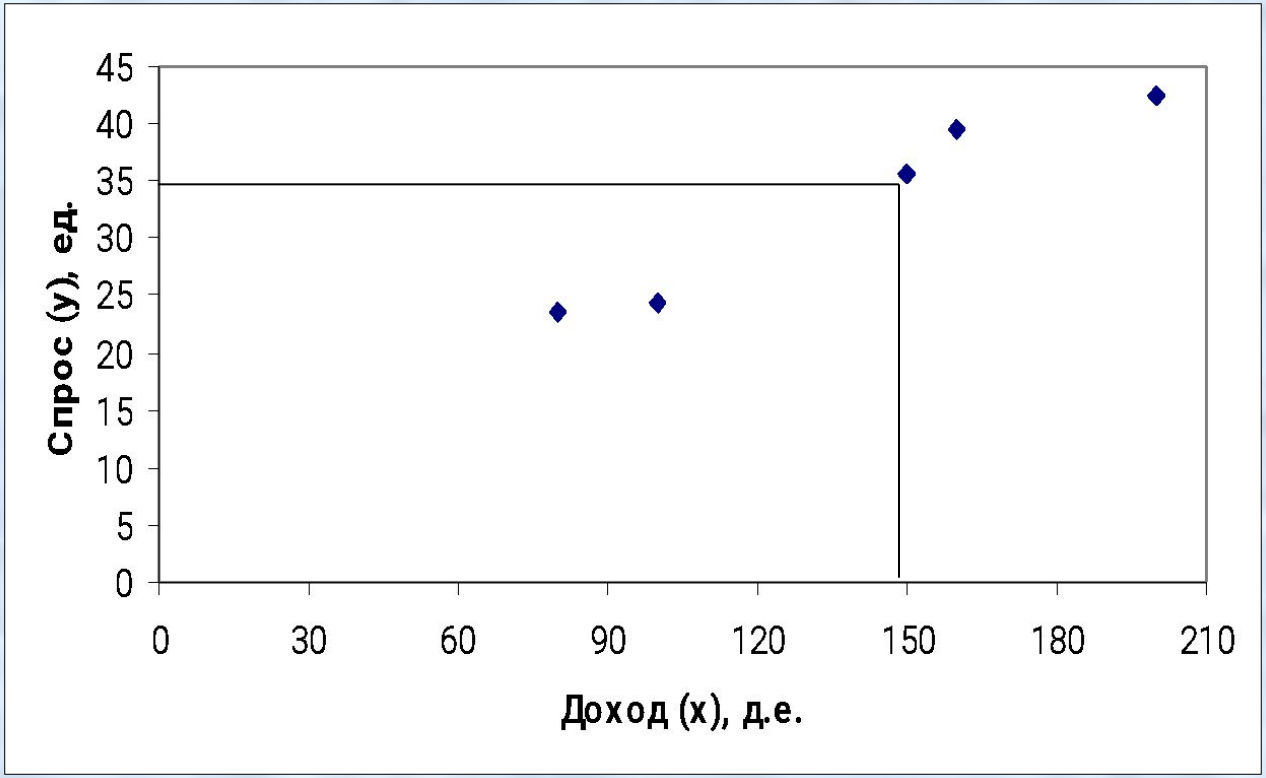
Нанесем точки на график



x	y
100	24
200	42
150	35
80	24
160	39

Метод наименьших квадратов

Нанесем точки на график

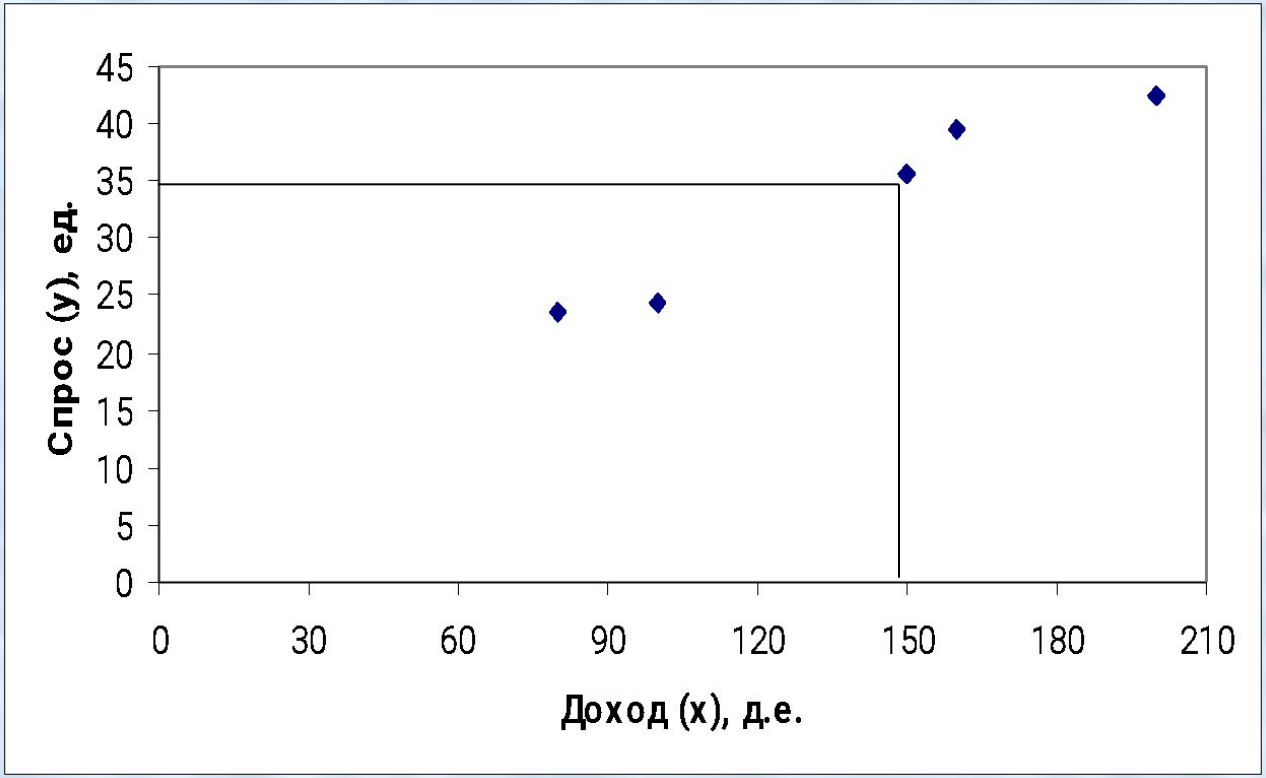


x	y
100	24
200	42
150	35
80	24
160	39

Точки разбросаны вокруг некоторой прямой!
Как ее найти?

Метод наименьших квадратов

Нанесем точки на график

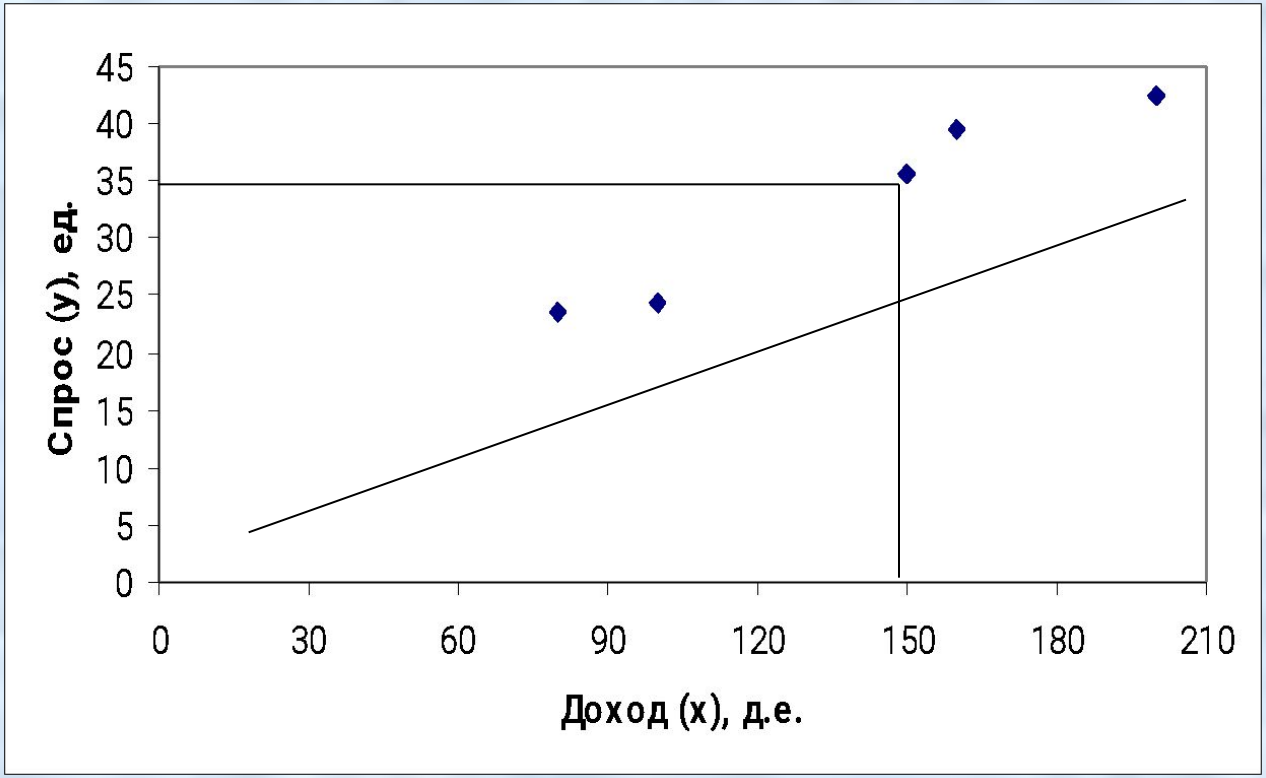


x	y
100	24
200	42
150	35
80	24
160	39

Расстояние от каждой точки до прямой должно быть как можно меньше!

Метод наименьших квадратов

Нанесем точки на график

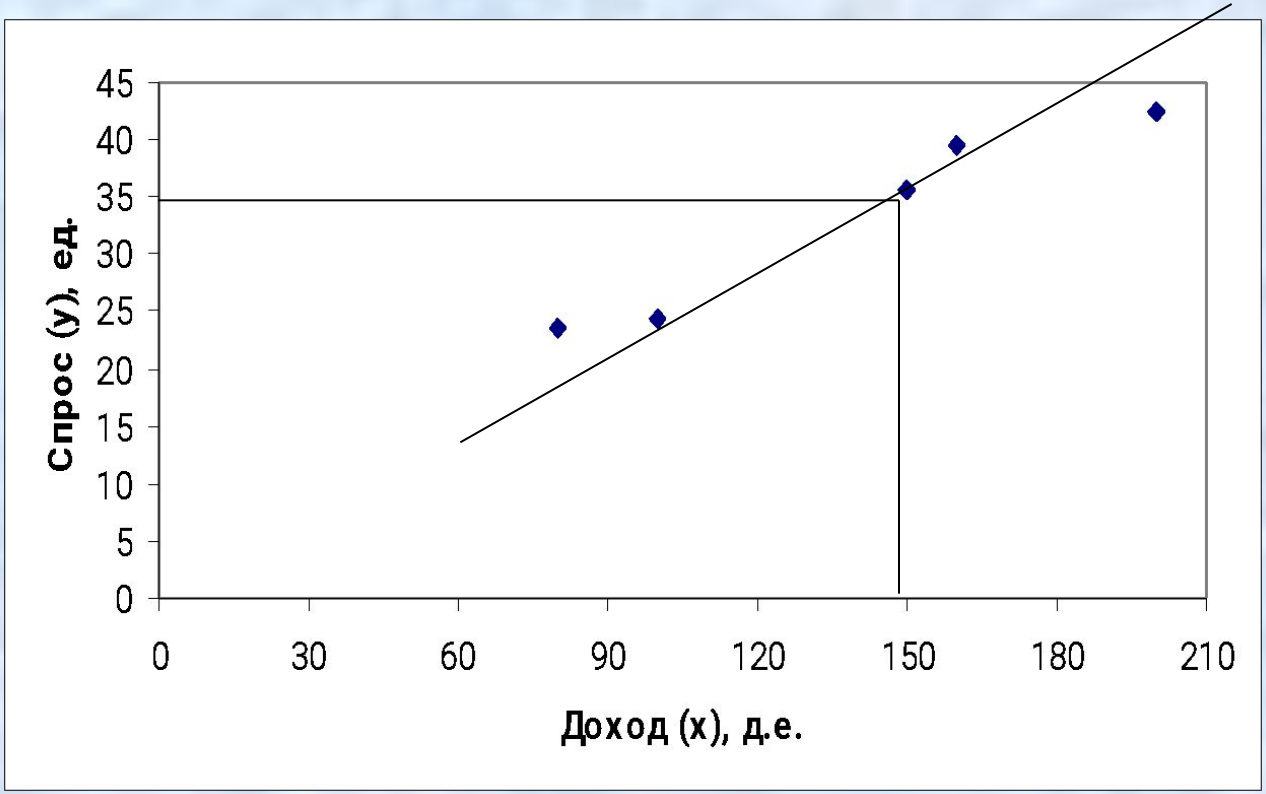


x	y
100	24
200	42
150	35
80	24
160	39

Плохая прямая!

Метод наименьших квадратов

Нанесем точки на график

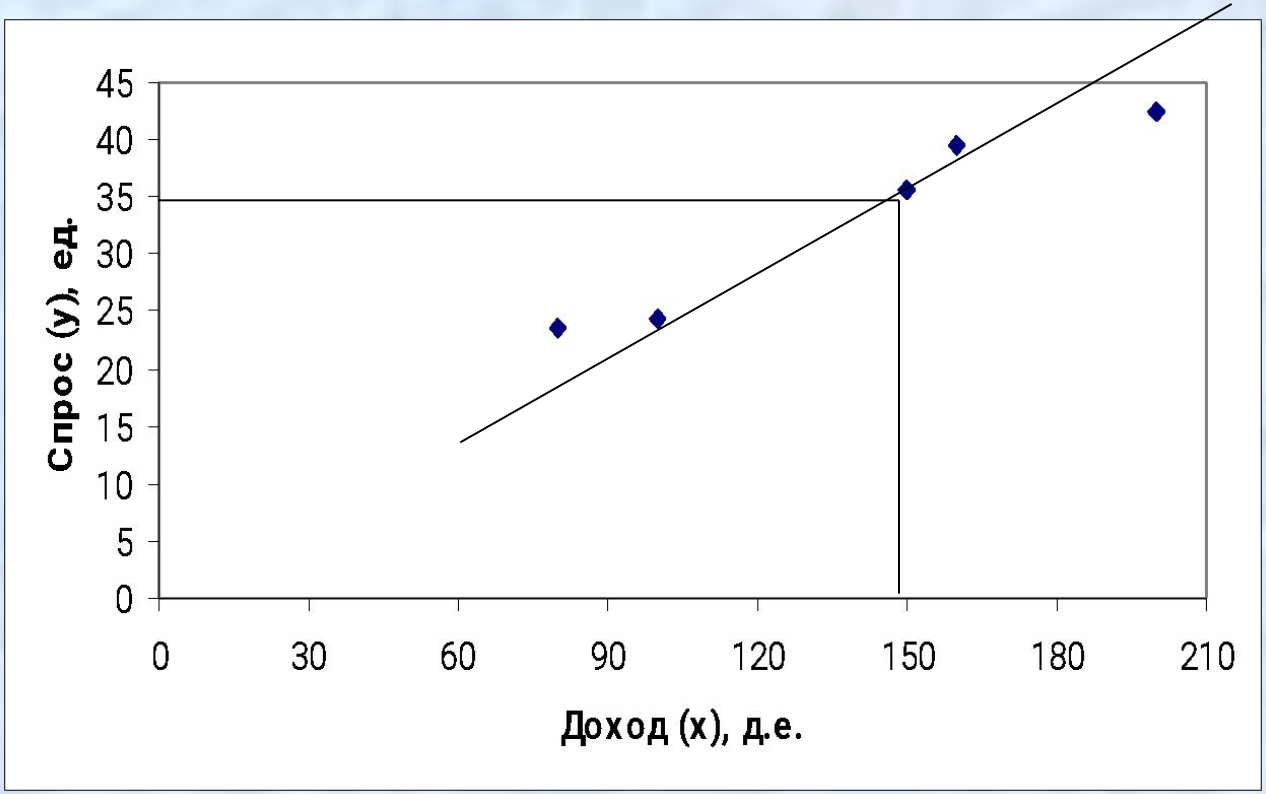


x	y
100	24
200	42
150	35
80	24
160	39

Хорошая прямая! Но может быть есть еще лучше?

Метод наименьших квадратов

Нанесем точки на график



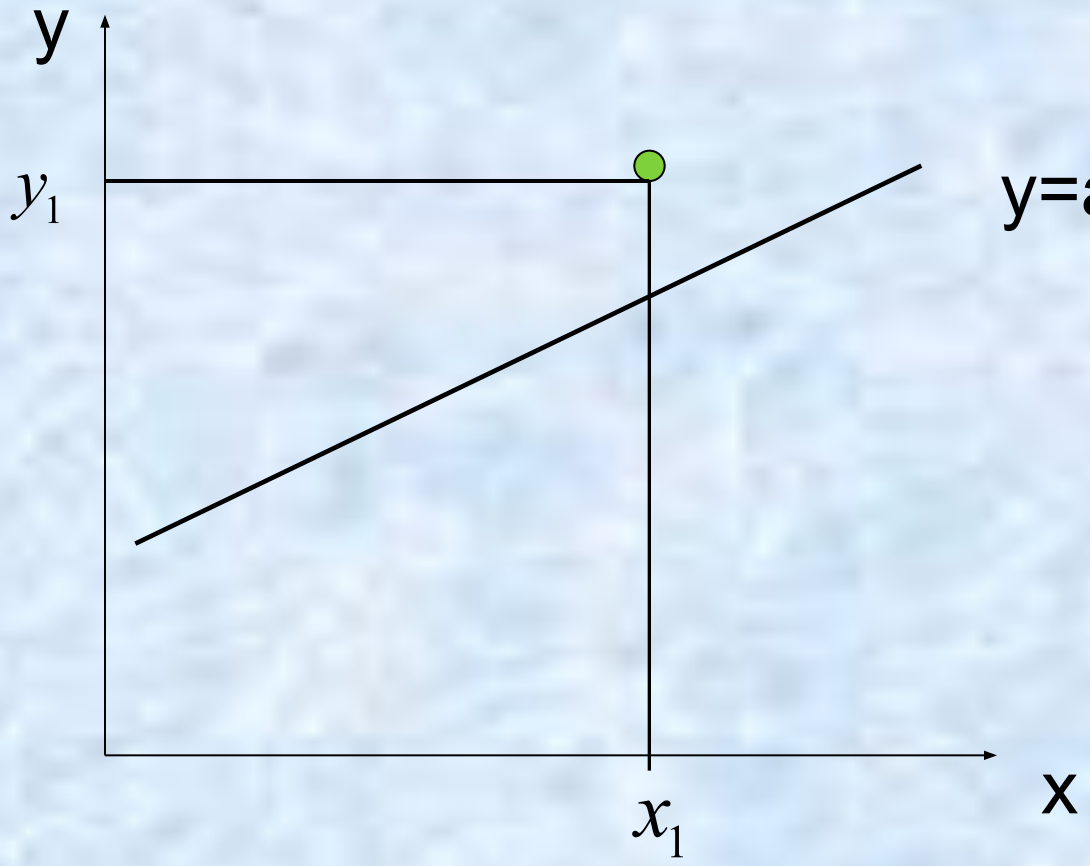
x	y
100	24
200	42
150	35
80	24
160	39

Уравнение прямой в общем виде y=ax+b. Надо найти наиболее подходящие a и b.

Обозначим

x_1 доход 1-го домохозяйства

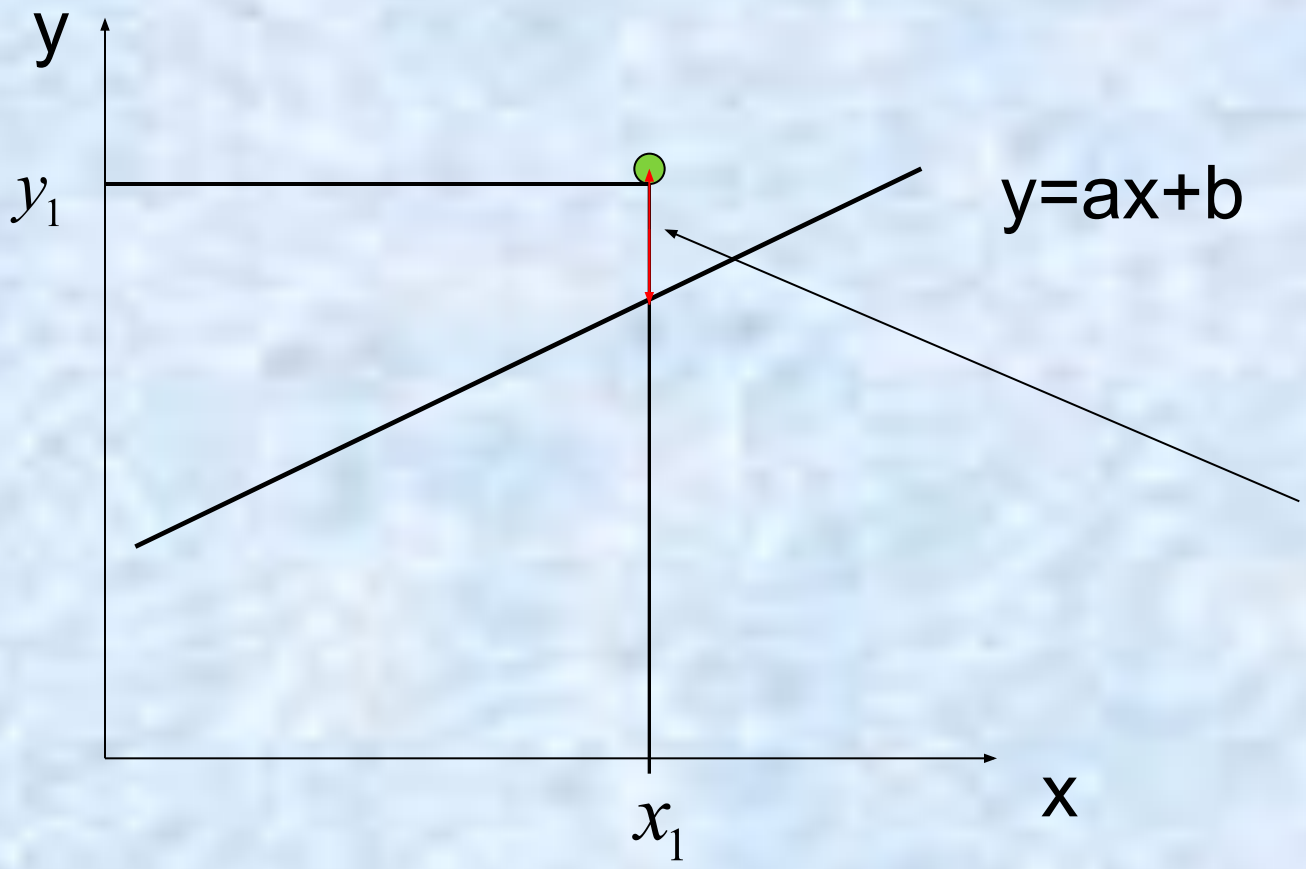
y_1 спрос 1-го домохозяйства на продукт



Обозначим

x_1 доход 1-го домохозяйства

y_1 спрос 1-го домохозяйства на продукт

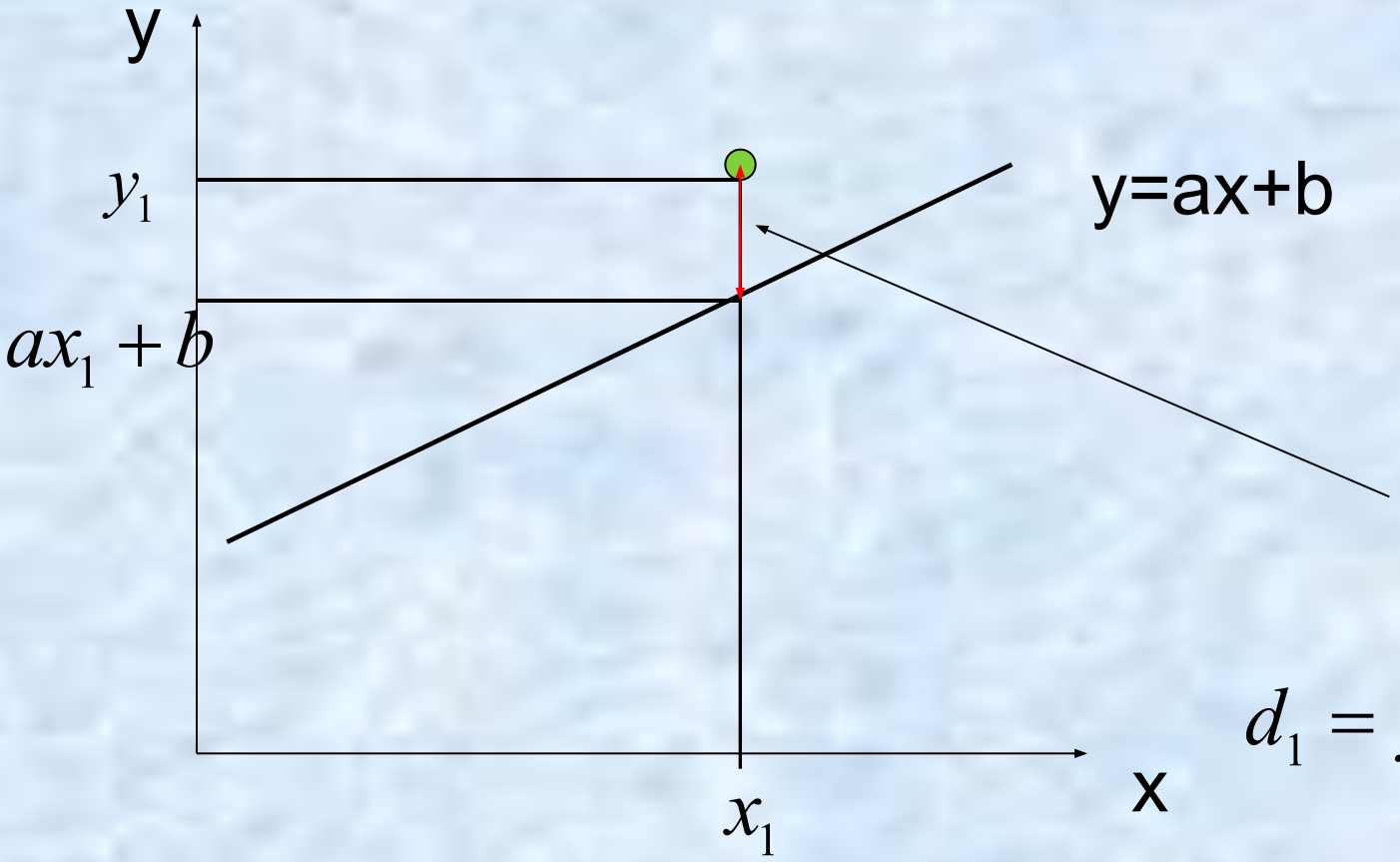


Отклонение точки
от прямой. Должно быть
как можно меньше!

Обозначим

x_1 доход 1-го домохозяйства

y_1 спрос 1-го домохозяйства на продукт

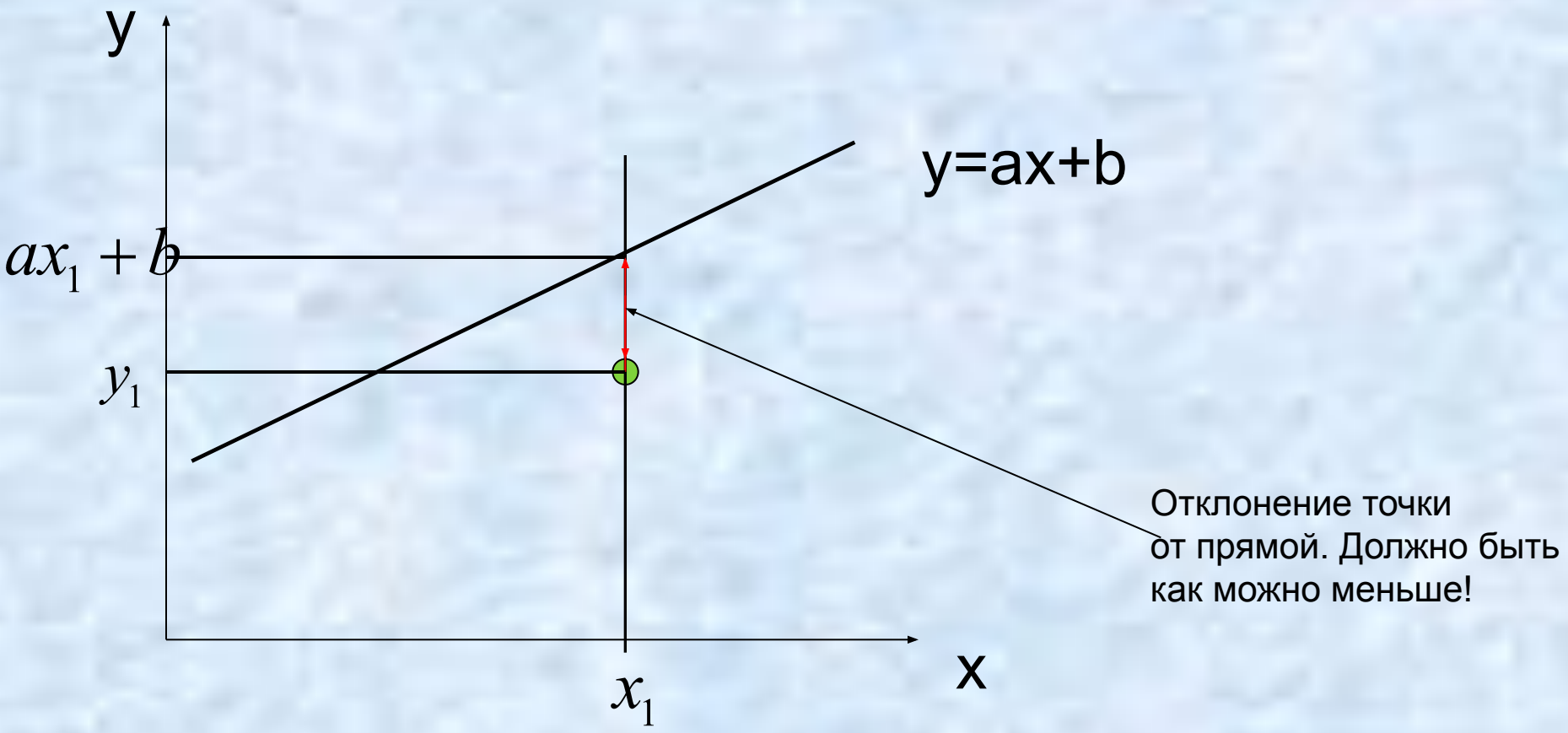


Отклонение точки
от прямой. Должно быть
как можно меньше!

$$d_1 = y_1 - (ax_1 + b)$$

А если точка лежит ниже прямой?

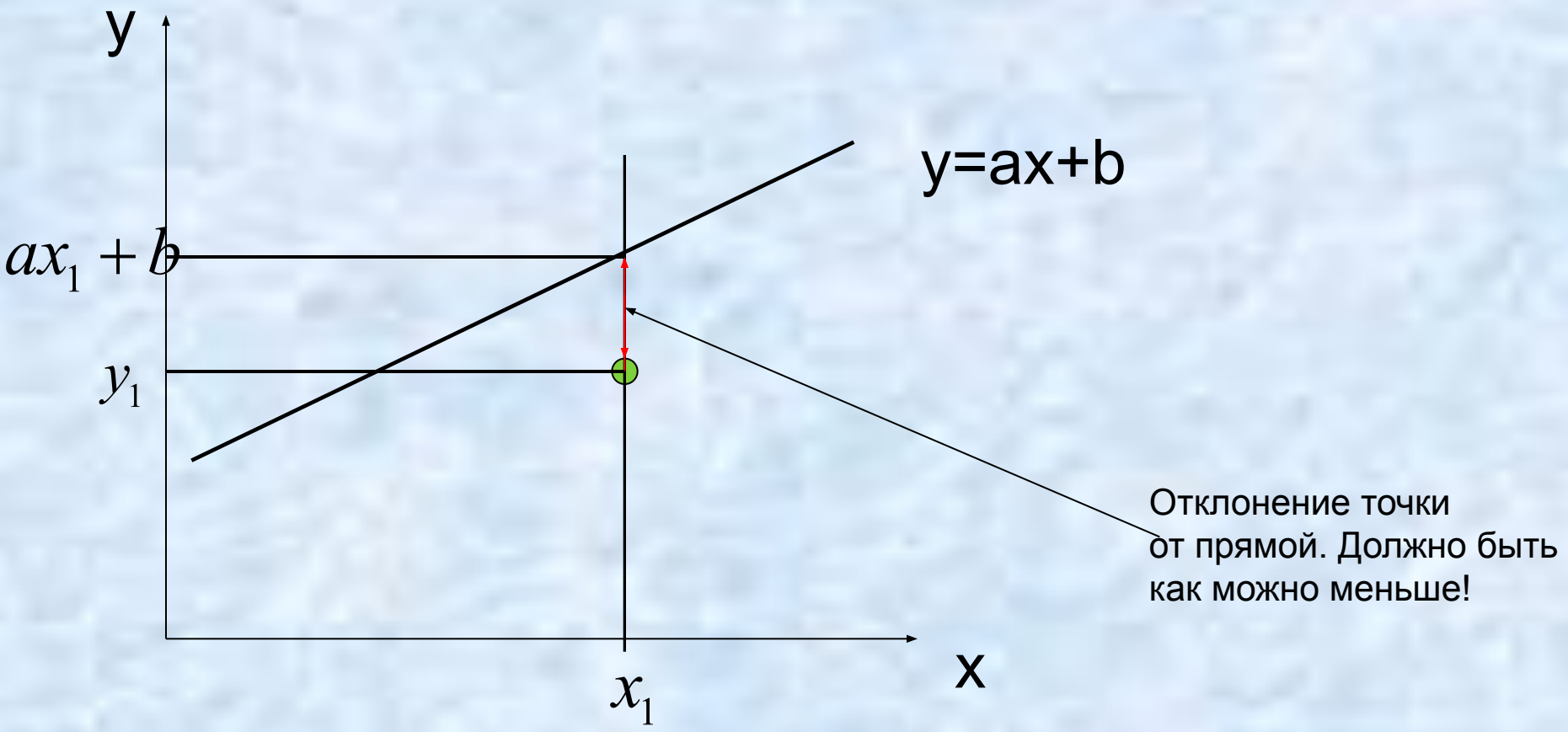
Тогда отклонение $d_1 = (ax_1 + b) - y_1$



Как учесть сразу оба случая?

Квадрат отклонения $d_1^2 = (y_1 - (ax_1 + b))^2$

должен быть как можно меньше.



Квадрат отклонения до второй точки тоже должен быть как можно меньше.

$$d_2^2 = (y_2 - (ax_2 + b))^2 \to \min$$

Квадрат отклонения до второй точки тоже должен быть как можно меньше.

$$d_2^2 = (y_2 - (ax_2 + b))^2 \to \min$$

И для третьей точки

$$d_3^2 = (y_3 - (ax_3 + b))^2 \to \min$$

Предположим, что у нас n точек.
Тогда и для последней точки

$$d_n^2 = (y_n - (ax_n + b))^2 \to \min$$

Как учесть все точки сразу?

$$d_1^2 + d_2^2 + d_3^2 + \ldots + d_n^2 \to \min$$

Сумма квадратов расстояний от точек до прямой должна быть как можно меньше.

Как учесть все точки сразу?

$$d_1^2 + d_2^2 + d_3^2 + \ldots + d_n^2 \to \min$$

Сумма квадратов расстояний от точек до прямой должна быть как можно меньше.

$$d_1^2 + d_2^2 + d_3^2 + \ldots + d_n^2 = \sum_{i=1}^{n} d_i^2$$

обозначение

Как учесть все точки сразу?

$$\sum_{i=1}^{n} d_i^2 \to \min$$

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

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

Получили функцию двух переменных, для которой надо найти минимум, т.е. надо исследовать на экстремум.

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

x_i и y_i это просто числа, нам известные

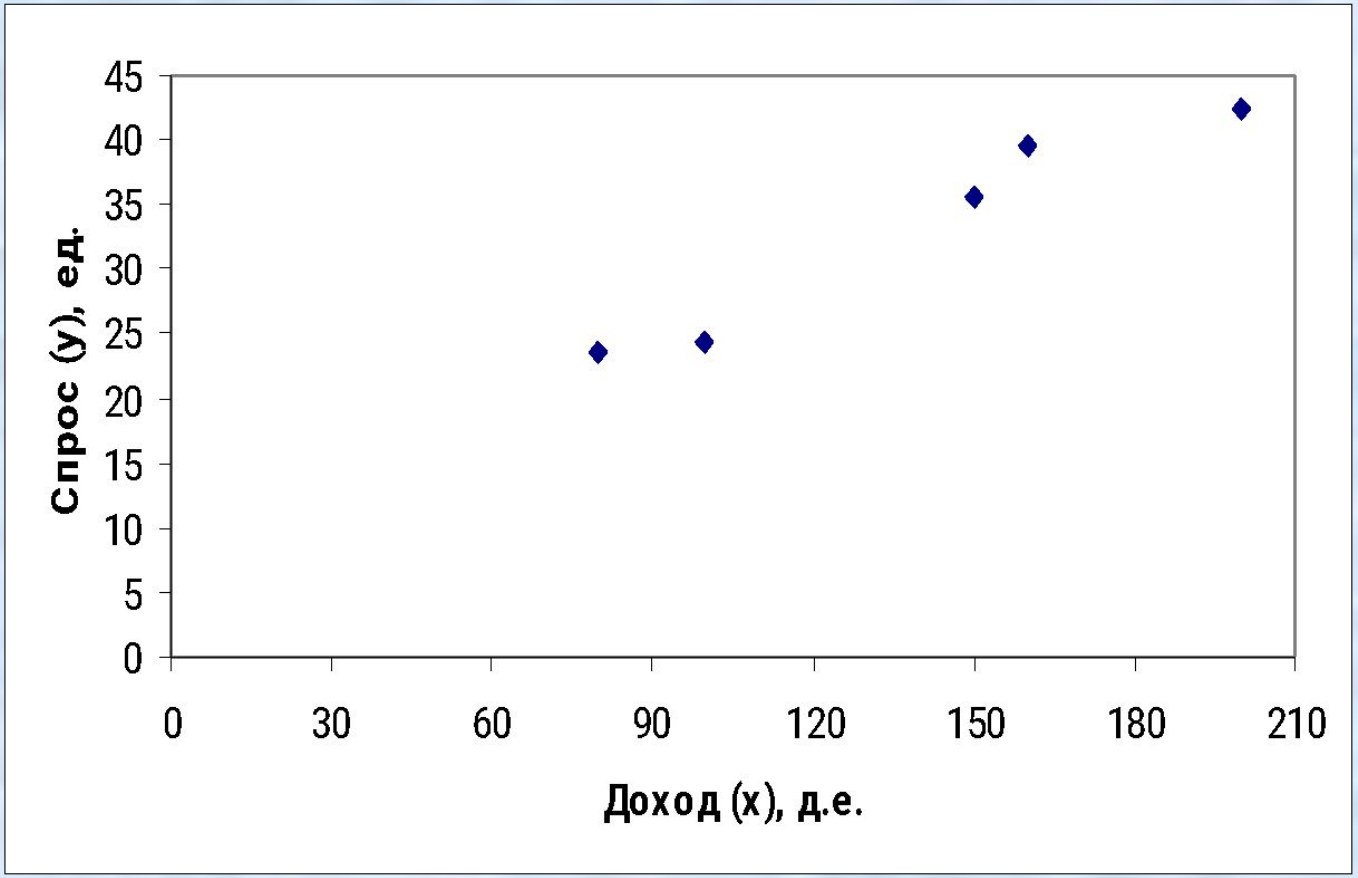
$$S(a,b)=\sum_{i=1}^{n}(y_i-ax_i-b)^2$$

x_i и y_i это просто числа, нам известные

$$a=\frac{\sum xy-n\cdot\bar{x}\cdot\bar{y}}{\sum x^2-n(\bar{x})^2}$$

$$b=\bar{y}-a\bar{x}$$

Вернемся к примеру

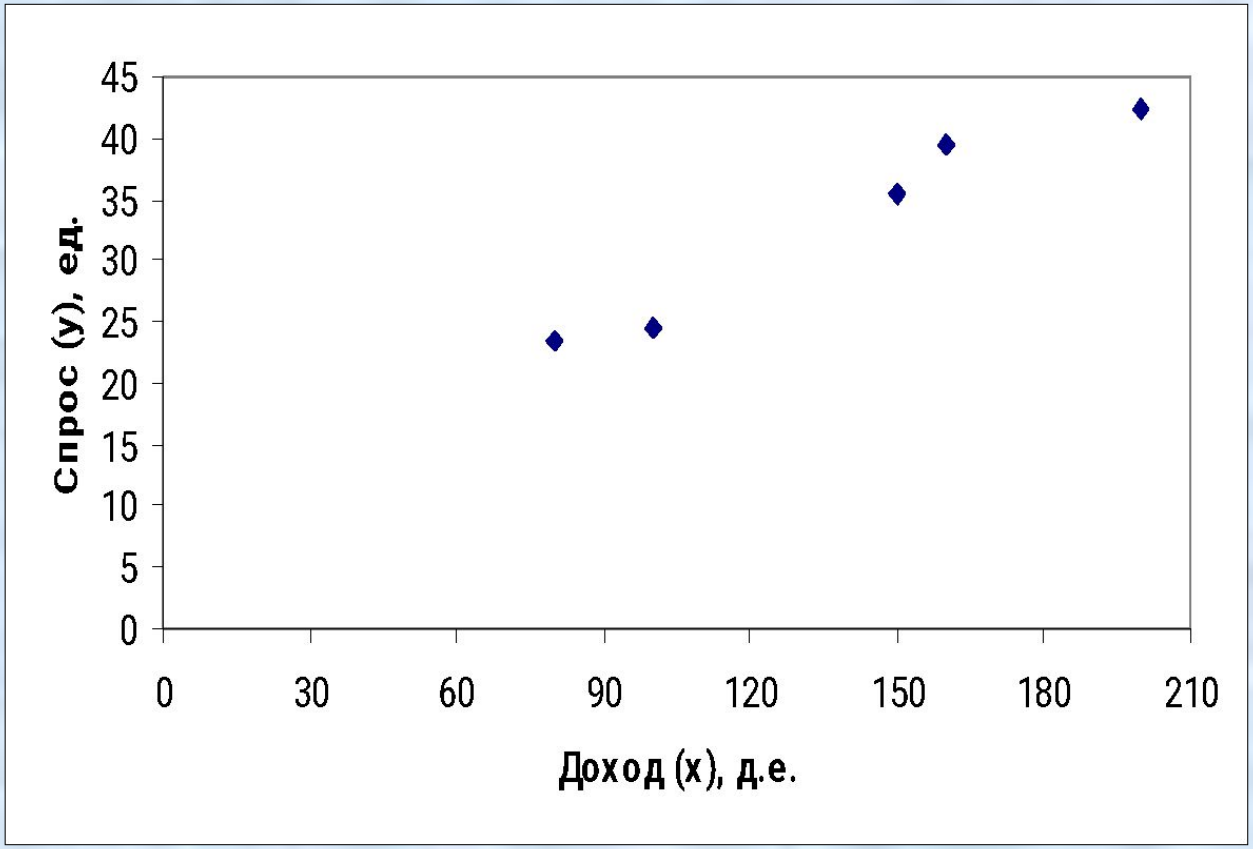


x	y
100	24
200	42
150	35
80	24
160	39

Надо найти

$$\bar{x}, \bar{y}, \sum xy, \sum x^2$$

Вернемся к примеру



x	y
100	24
200	42
150	35
80	24
160	39

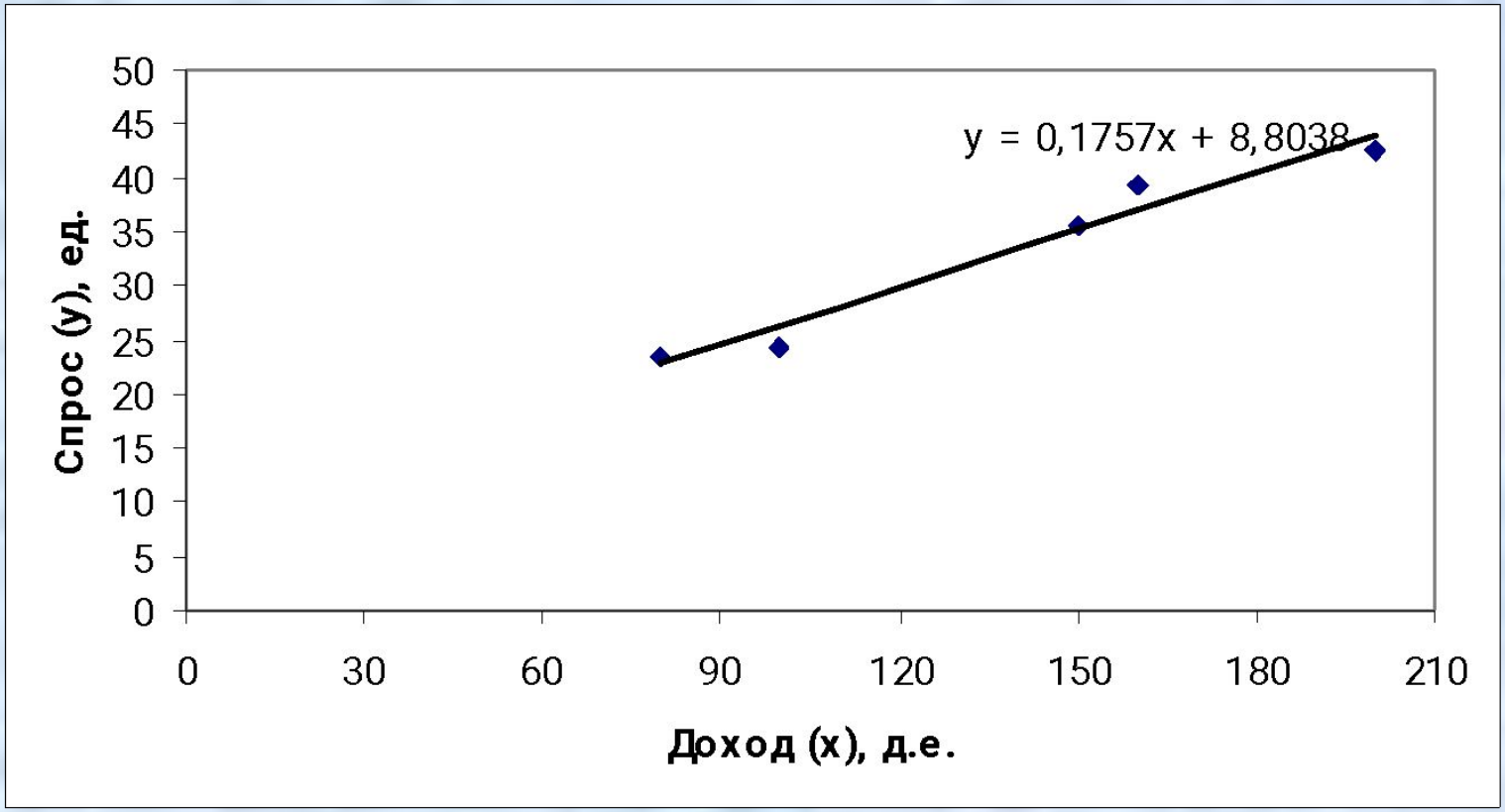
$$\bar{x} = 138,\ \bar{y} \approx 33,\ \sum xy = 24400,\ \sum x^2 = 104500$$

$\bar{x} = 138,\ \bar{y} \approx 33,\ \sum xy = 24400,\ \sum x^2 = 104500,\ n = 5$

$$a = \frac{\sum xy - n \cdot \bar{x} \cdot \bar{y}}{\sum x^2 - n(\bar{x})^2} \qquad b = \bar{y} - a\bar{x}$$

a=0,18, b=8,8

y=0,18x+8,8 - уравнение прямой, которая проходит ближе всего к точкам.



y=0,18x+8,8 - функция спроса в зависимости от дохода.

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*
Y-пересечение	9,334052	3,296116	2,831833	0,06609
Переменная X 1	0,170043	0,0228	7,458124	0,004991

y=0,18x+8,8 - функция спроса в зависимости от дохода.

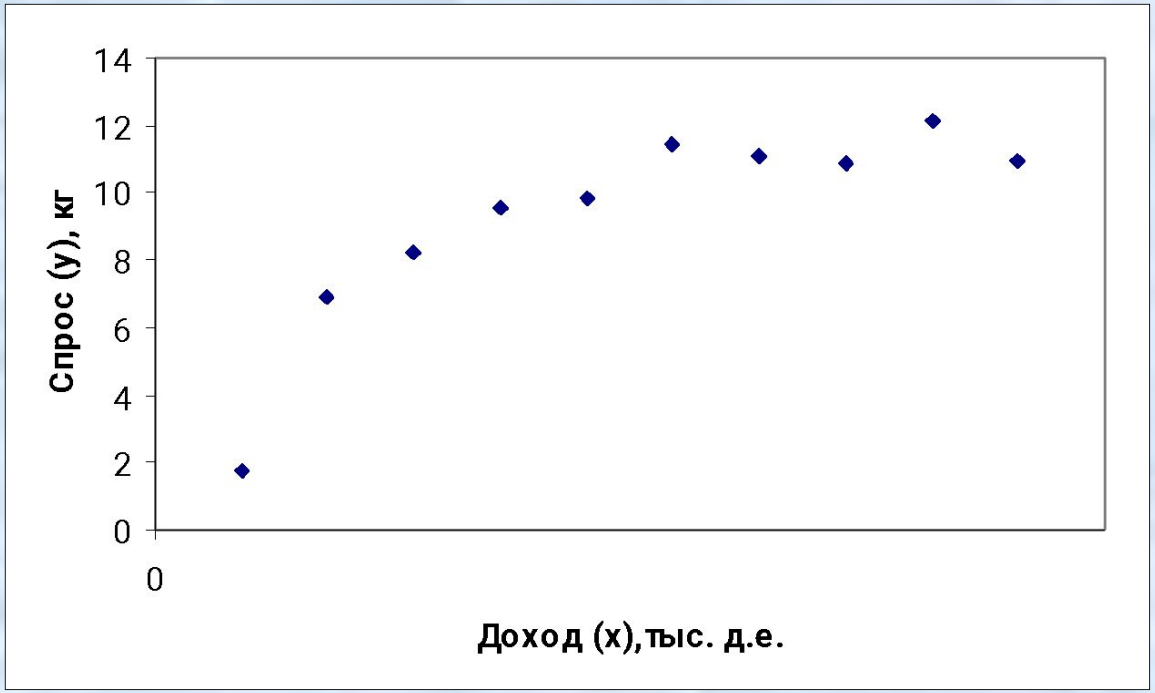
y=0,18x+8,8 - функция спроса в зависимости от дохода.

1) *Выполнить прогноз потребления продукта домохозяйством с доходом 200 д.е.*

2)Найти предельную склонность к потреблению продукта.

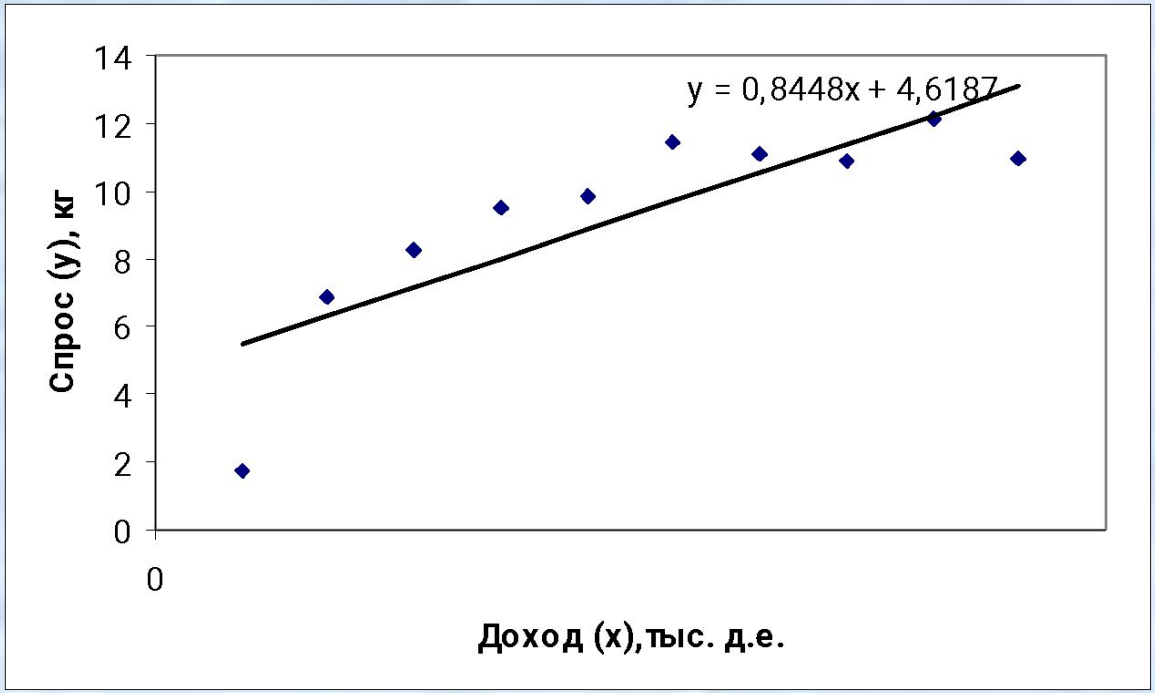
3) Найти эластичность спроса по доходу при доходе 100 д.е. и 50 д.е.

№ домохозяйства	Среднедушевой доход домохозяйства, тыс. д.е.	Объем спроса, кг в месяц
1	1	1,71
2	2	6,88
3	3	8,25
4	4	9,52
5	5	9,81
6	6	11,43
7	7	11,09
8	8	10,87
9	9	12,15
10	10	10,94



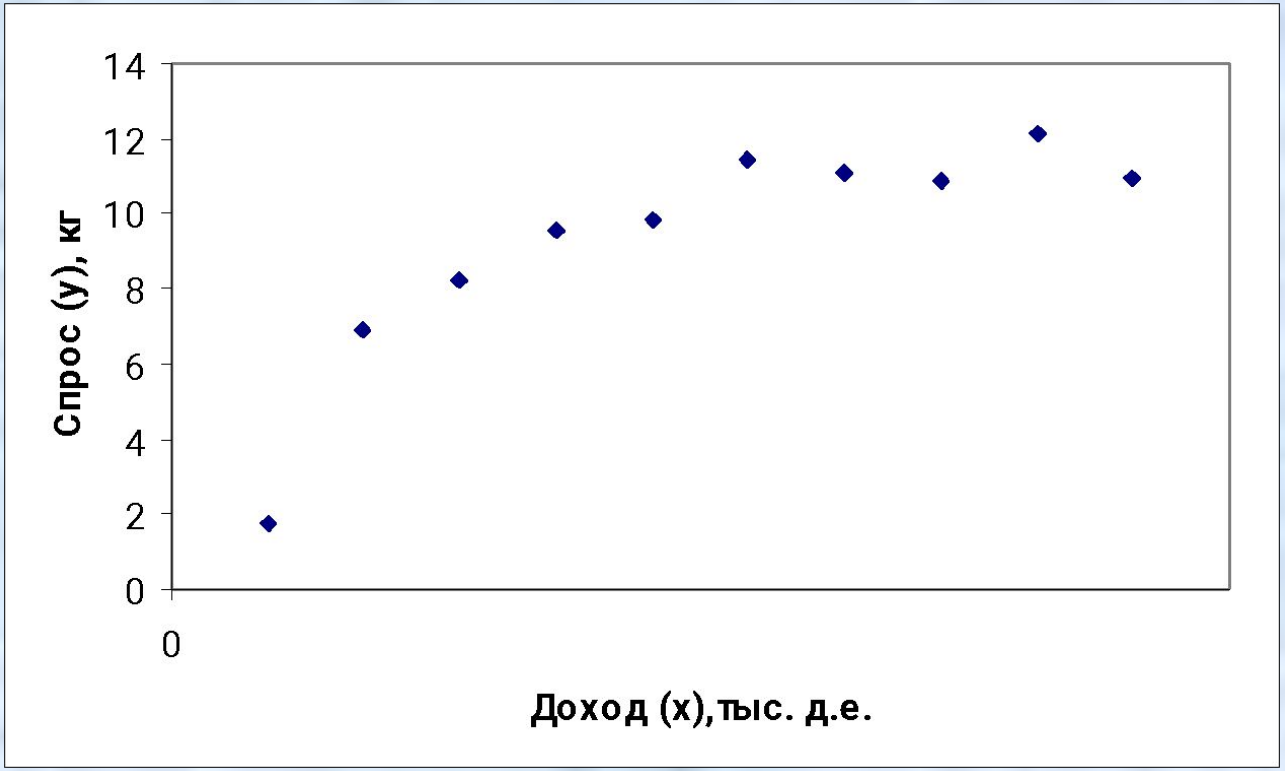
x	y
1	1,71
2	6,88
3	8,25
4	9,52
5	9,81
6	11,43
7	11,09
8	10,87
9	12,15
10	10,94

Зависимость нелинейная!



x	y
1	1,71
2	6,88
3	8,25
4	9,52
5	9,81
6	11,43
7	11,09
8	10,87
9	12,15
10	10,94

Попытка провести прямую



x	y
1	1,71
2	6,88
3	8,25
4	9,52
5	9,81
6	11,43
7	11,09
8	10,87
9	12,15
10	10,94

Попробуем провести гиперболу $y = a\frac{1}{x} + b$ наилучшим образом.

$$\sum_{i=1}^{n} d_i^2 \to \min$$

$$\sum_{i=1}^{n} \left(y_i - \left(a \cdot \frac{1}{x_i} + b \right) \right)^2 \to \min$$

$$S(a,b) = \sum_{i=1}^{n} \left(y_i - \left(a \cdot \frac{1}{x_i} + b \right) \right)^2$$

Получили функцию двух переменных, для которой надо найти минимум, т.е. надо исследовать на экстремум.

$$S(a,b)=\sum_{i=1}^{n}\left(y_i-\left(a\cdot\frac{1}{x_i}+b\right)\right)^2$$

Можно исследовать на экстремум, но лучше заменить

$$z_i=\frac{1}{x_i}$$

тогда $$S(a,b)=\sum_{i=1}^{n}\left(y_i-\left(a\cdot z_i+b\right)\right)^2$$

А это такая же функция, что и для линейной регрессии!
Поэтому можно воспользоваться готовым результатом!

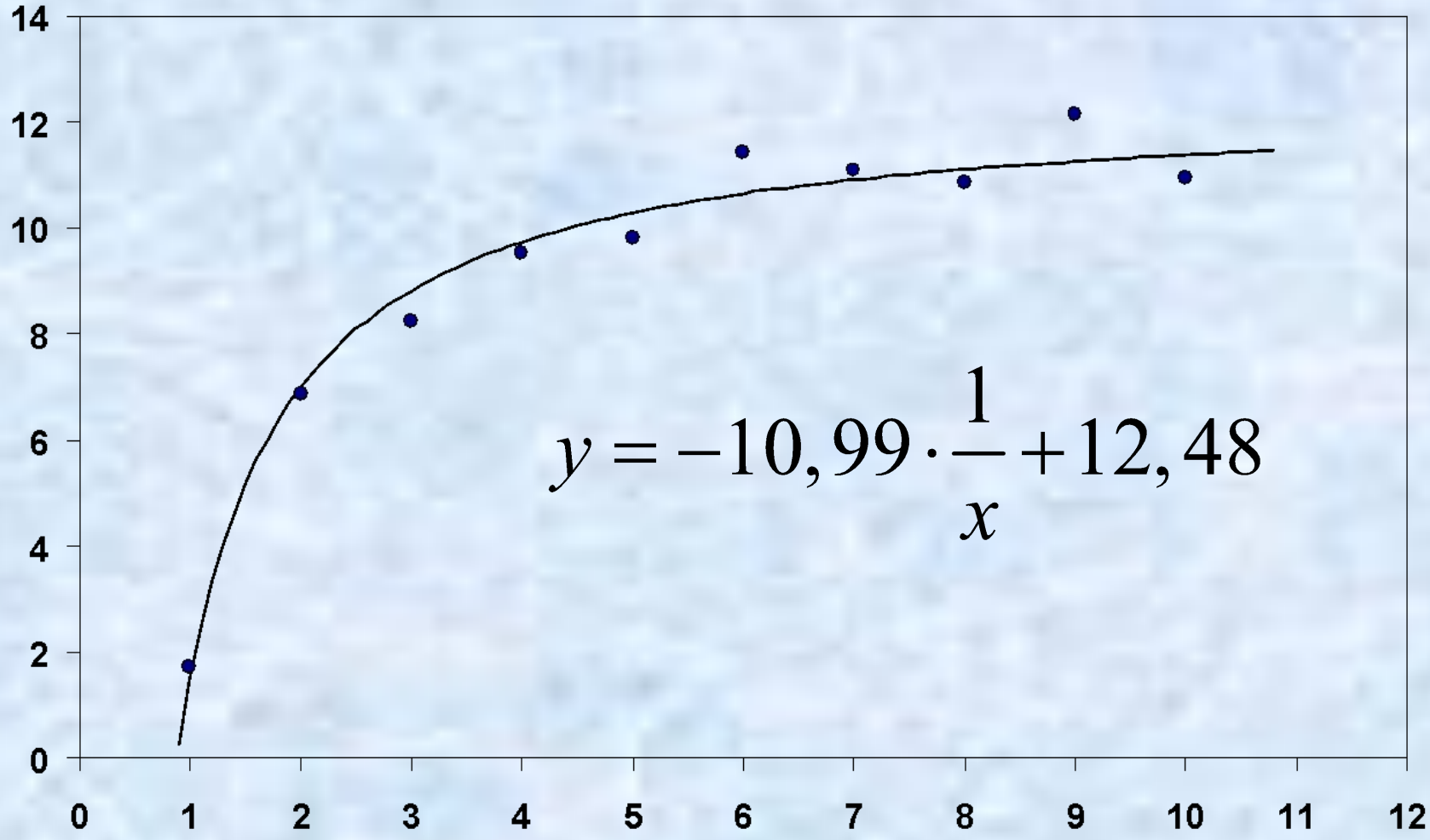
x	y	z
1	1,71	1,00
2	6,88	0,50
3	8,25	0,33
4	9,52	0,25
5	9,81	0,20
6	11,43	0,17
7	11,09	0,14
8	10,87	0,13
9	12,15	0,11
10	10,94	0,10

Сначала рассчитаем столбик z=1/x

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*
Y-пересечение	12,48354	0,255751	48,81128	3,43E-11
z=1/x	-10,9887	0,649657	-16,9145	1,51E-07

$$y = -10,99 \cdot z + 12,48$$

$$y = -10,99 \cdot \frac{1}{x} + 12,48$$


14
12
10
8
6
4
2
0
0
1
2
3
4
5
6
7
8
9
10
11
12
$y = -10,99 \cdot \frac{1}{x} + 12,48$

$y = -10,99 \cdot \frac{1}{x} + 12,48$ - *функция спроса в зависимости от дохода.*

1) *Выполнить прогноз потребления продукта домохозяйством с доходом 4 тыс.д.е.*
2) *Имеется ли уровень насыщения для данного продукта? Если да, найти его.*
2)*Найти предельную склонность к потреблению продукта.*
3) *Найти эластичность спроса по доходу при доходе 1000 д.е. и 10000 д.е.*

МОДЕЛЬ МНОЖЕСТВЕННОЙ ЛИНЕЙНОЙ РЕГРЕССИИ

$$y = a_1 x_1 + a_2 x_2 + \ldots + a_{r-1} x_{r-1} + a_r + \xi$$

y – зависимая или объясняемая переменная

$x_1, x_2 \ldots , x_{r-1}$ - независимые или объясняющие переменные

ξ - случайная составляющая.

Задача множественного регрессионного анализа – оценить $a_1, a_2 \ldots a_r$

Пример: Множественная регрессия

Мы хотим определить связь между заработной платой, числом лет обучения и опытом работы.

- y – почасовая заработная плата ($).
- x_1 – число лет обучения
- x_2 – опыт работы (лет)

$$y = a_1 x_1 + a_2 x_2 + a_3 + \xi$$

МОДЕЛЬ МНОЖЕСТВЕННОЙ ЛИНЕЙНОЙ РЕГРЕССИИ

$$y = a_1 x_1 + a_2 x_2 + a_3 + \xi$$

Для оценки необходима ***выборка (большое количество респондентов)***

№ семьи	y заработная плата	x_1 число лет обучения	x_2 опыт работы
1	10	5	10
2	12	6	13
3	15	6	20
4	6	2	4
5	20	4	18

n – объем выборки

y_i заработная плата i-го респондента

x_{i1} число лет обучения i-го респондента

x_{i2} опыт работы i-го респондента

$$i = 1 \ldots \ n$$

Уравнение для i-й семьи

$$y_i = a_1 x_{i1} + a_2 x_{i2} + a_3 + \xi_i$$

Чтобы подобрать наилучшие a_1, a_2, a_3

$$S(a_1, a_2, a_3) = \sum_{i=1}^{n} \left(y_i - a_1 x_{i1} - a_2 x_{i2} - a_3 \right)^2$$

$$\min_{a_1, a_2 \ldots a_r} S(a_1, a_2 \ldots a_r)$$

Пример оценки параметров в модели зависимости заработной платы от числа лет обучения и опыта работы

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*
Y-пересечение	-26,93164811	4,523407834	-5,95384	4,73E-09
N	2,674036105	0,231999296	11,52605	1,28E-27
Nrab	0,59409725	0,137923673	4,307435	1,96E-05

Zpl = 2.67*N + 0.59*NRab -26.93

ИНТЕРПРЕТАЦИЯ ПАРАМЕТРОВ ЛИНЕЙНОЙ МНОЖЕСТВЕННОЙ РЕГРЕССИИ

$$\hat{y} = \hat{a}_1 x_1 + \hat{a}_2 x_2 + \ldots + \hat{a}_{r-1} x_{r-1} + \hat{a}_r$$

Интерпретация: коэффициент регрессии при переменной x_i показывает на сколько единиц изменится переменная y при изменении переменной x_i на 1 единицу, при условии постоянства других переменных:

Пример оценки параметров в модели зависимости заработной платы от числа лет обучения и опыта работы

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*
Y-пересечение	-26,93164811	4,523407834	-5,95384	4,73E-09
N	2,674036105	0,231999296	11,52605	1,28E-27
Nrab	0,59409725	0,137923673	4,307435	1,96E-05

Zpl = 2.67*N + 0.59*NRab -26.93

Пример оценки параметров в модели зависимости заработной платы от числа лет обучения и опыта работы

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*
Y-пересечение	-26,93164811	4,523407834	-5,95384	4,73E-09
N	2,674036105	0,231999296	11,52605	1,28E-27
Nrab	0,59409725	0,137923673	4,307435	1,96E-05

Zpl = 2.67*N + 0.59*NRab -26.93

Каждый дополнительный год обучения при данном опыте работы увеличивает часовой заработок на 2,67$

Каждый дополнительный год опыта работы при данной продолжительности обучения увеличивает часовой заработок на 0,59$

-26,93 не имеет содержательной интерпретации.

ИНТЕРПРЕТАЦИЯ ПАРАМЕТРОВ ЛИНЕЙНОЙ МНОЖЕСТВЕННОЙ РЕГРЕССИИ

Пример y – затраты на питание (млрд. $)

x_1 – личный располагаемый доход (млрд. $)

x_2 – индекс цен на продукты питания (%)

$$\hat{y} = 0,112x_1 - 0,739x_2 + 116,7$$

ИНТЕРПРЕТАЦИЯ ПАРАМЕТРОВ ЛИНЕЙНОЙ МНОЖЕСТВЕННОЙ РЕГРЕССИИ

Пример y – затраты на питание (млрд. $)

x_1 – личный располагаемый доход (млрд. $)

x_2 – индекс цен на продукты питания (%)

$$\hat{y} = 0,112x_1 - 0,739x_2 + 116,7$$

При увеличении личного располагаемого дохода на 1 млрд. $ (при сохранении неизменной цены) расходы на питание увеличатся на 112 млн.$

При увеличении индекса цен на 1 процентный пункт (при сохранении постоянных доходов) расходы на питание сократятся на 739 млн.$

116,7 не интерпретируется, т.к. x_1 и x_2 не могут быть равными 0.

Коэффициент детерминации -это доля дисперсии признака y, объясненная регрессией в общей дисперсии признака y. Чем ближе к 1, тем лучше!

Регрессионная статистика	
Множественный R	0,446161
R-квадрат	0,19906
Нормированный R-квадрат	0,196077
Стандартная ошибка	13,09197
Наблюдения	540

Коэффициент детерминации

Дисперсионный анализ					Значимость F
	df	*SS*	*MS*	*F*	1,31E-26
Регрессия	2	22875,36	11437,68	66,73107	
Остаток	537	92041,6	171,3996		
Итого	539	114917			

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*
Y-пересечение	-26,9316	4,523408	-5,95384	4,73E-09
N	2,674036	0,231999	11,52605	1,28E-27

Множественный коэффициент корреляции -это корень квадратный из коэффициента детерминации. Чем ближе к 1, тем лучше!

Регрессионная статистика

Множественный R	0,446161	Множественный коэффициент корреляции
R-квадрат	0,19906	
Нормированный R-квадрат	0,196077	
Стандартная ошибка	13,09197	
Наблюдения	540	

Дисперсионный анализ

	df	*SS*	*MS*	*F*	Значимость F
Регрессия	2	22875,36	11437,68	66,73107	1,31E-26
Остаток	537	92041,6	171,3996		
Итого	539	114917			

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*
Y-пересечение	-26,9316	4,523408	-5,95384	4,73E-09
N	2,674036	0,231999	11,52605	1,28E-27

Значимость F - это вероятность того, что полученная зависимость случайна. При значимости больше 0,05 обычно считают, что построенная зависимость незначима. Моделью нельзя пользоваться для прогнозирования.

Значимость F
1,31E-26

Дисперсионный анализ

	df	*SS*	*MS*	*F*
Регрессия	2	22875,36	11437,68	66,73107
Остаток	537	92041,6	171,3996	
Итого	539	114917		

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*
Y-пересечение	-26,9316	4,523408	-5,95384	4,73E-09
N	2,674036	0,231999	11,52605	1,28E-27
Nrab	0,594097	0,137924	4,307435	1,96E-05

P-значение - это вероятность того, что соответствующая переменная не влияет на зависимую переменную y. При P-значении больше 0,05 обычно считают, что соответствующая переменная незначима и ее можно исключить из уравнения регрессии.

Замечание. Константу из уравнения регрессии удалять нельзя, даже если она незначима.

Дисперсионный анализ

	df	*SS*	*MS*	*F*
Регрессия	2	22875,36	11437,68	66,73107
Остаток	537	92041,6	171,3996	
Итого	539	114917		

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*
Y-пересечение	-26,9316	4,523408	-5,95384	4,73E-09
N	2,674036	0,231999	11,52605	1,28E-27
Nrab	0,594097	0,137924	4,307435	1,96E-05