Тема 7.
Корреляция и регрессия

7.1. Корреляция

7.2. Значимость коэффициента корреляции

7.3. Регрессия

7.4. Надежность прогноза

Примеры

1. **Менеджер** интересуется, зависит ли объем продаж в этом месяце от объема рекламы в этом же периоде?
2. **Преподаватель** хочет выяснить, есть ли зависимость между количеством часов, потраченных студентом на занятия, и результатами экзамена?
3. **Врач** исследует, влияет ли кофеин на сердечные болезни и существует ли связь между возрастом человека и его кровяным давлением?
4. **Зоолог** стремится узнать, есть ли связь между весом определенного животного при рождении и его продолжительностью жизни.
5. **Социолог** исследует, какова связь между уровнем преступности и уровнем безработицы в регионе? Есть ли зависимость между расходами на жилье и совокупным доходом семьи? Связаны ли доход от профессиональной деятельности и продолжительность образования?

На эти вопросы можно ответить, используя методы корреляционного и регрессионного анализа, рассмотренные в материалах этой лекции.

Постановка проблемы

Четыре вопроса:

Вопрос 1. **Существует ли связь** между двумя или более переменными?

Вопрос 2. Какой **тип** имеет эта связь?

Вопрос 3. Насколько она **сильна**?

Вопрос 4. Какой можно сделать **прогноз**, основываясь на этой связи?

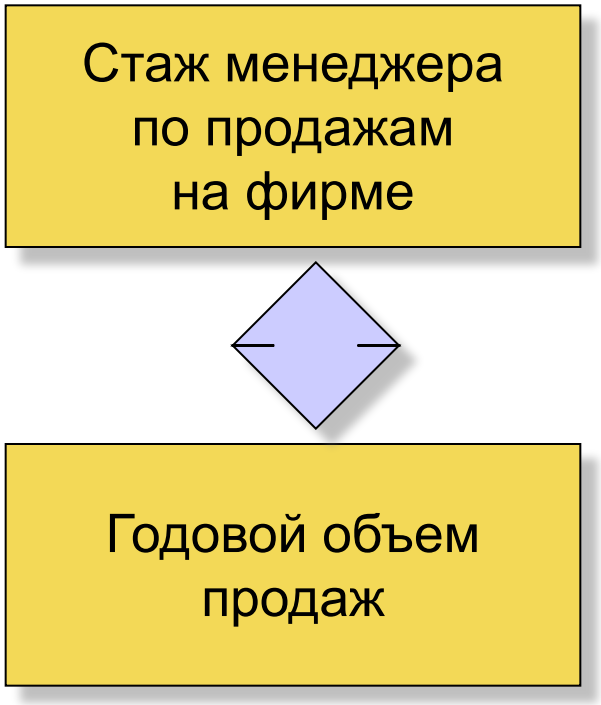
Методы

Корреляция – статистический метод, позволяющий определить, существует ли зависимость между переменными и на сколько она сильна.

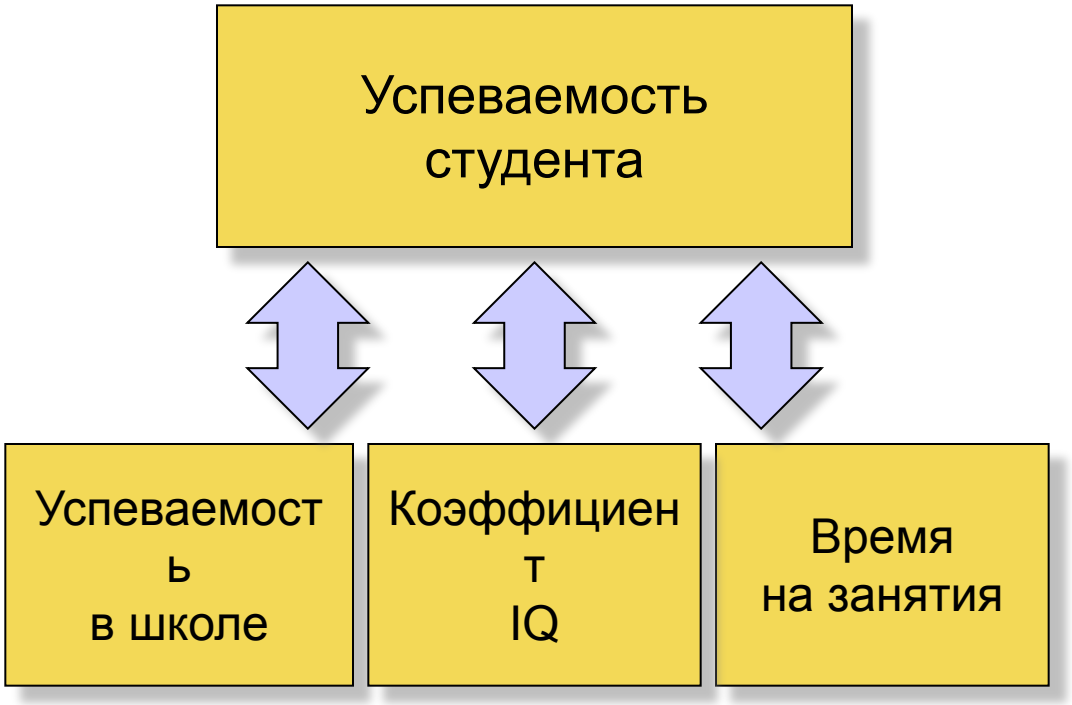
Регрессия – статистический метод, который используется для описания характера связи между переменными (положительная или отрицательная, линейная или нелинейная зависимость).

Простая и множественная связь

Простая связь означает изучение двух переменных.



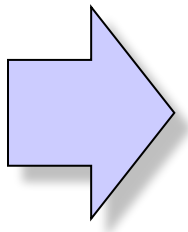
Множественная связь означает изучение несколько переменных.

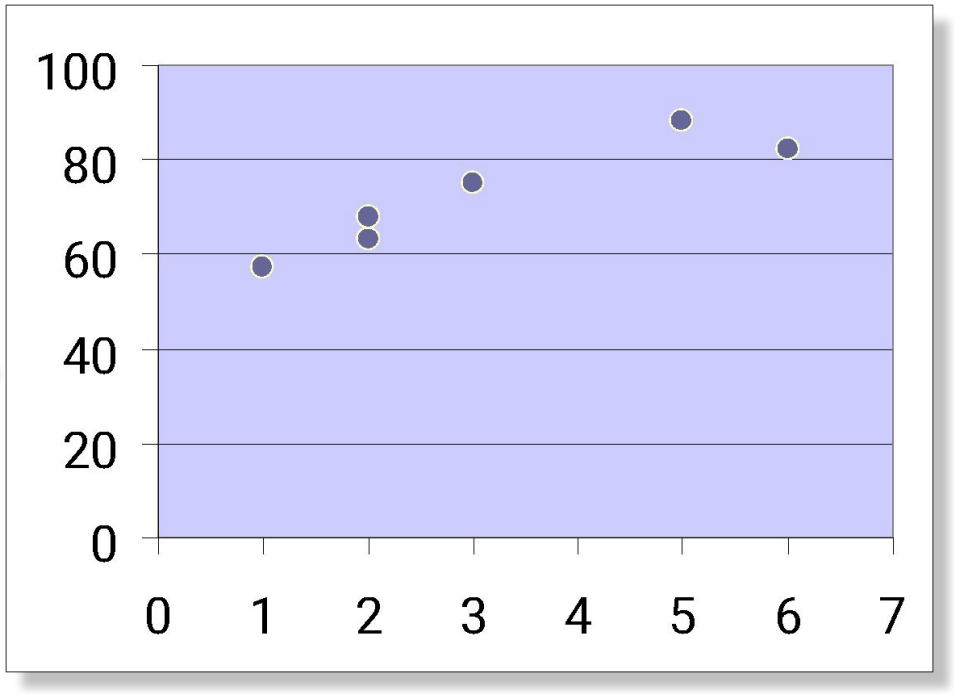


Визуальный анализ связи

Рассматриваем две переменные: «продолжительность занятий» студентов перед экзаменом и «итоговая оценка» (из 100 балов). Пытаемся визуально определить связь. Правда ли, что **чем больше времени занятий, тем выше оценка?**

Студент	Часы x	Оценка y
A	6	82
B	2	63
C	1	57
D	5	88
E	2	68
F	3	75





Независимая и зависимая переменные

Независимая переменная – это та переменная в регрессии, которую можно изменять. В данном случае, переменная «количество часов занятий» является независимой и обозначается как переменная *x*.

Зависимая переменная – это переменная в регрессии, которую нельзя изменять. «Экзаменационная оценка» является зависимой переменной. Она обозначается *y*.

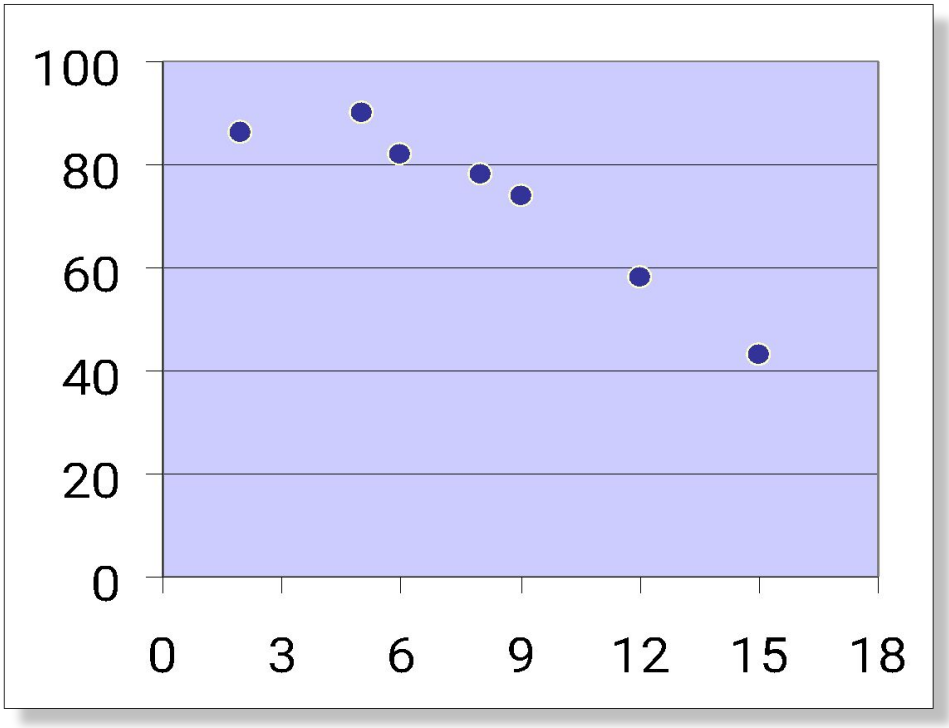
Причиной такого разделения переменных является то, что *предполагается*, что оценка, которую получает студент, зависит от количества часов, которые он посвятил занятиям. Предполагается также, что студенты могут регулировать количество часов, которое они тратят на занятия.

Не всегда можно ясно определить, какая переменная зависимая, а какая независимая, и выбор иногда делается произвольно.

Положительная и отрицательная зависимость

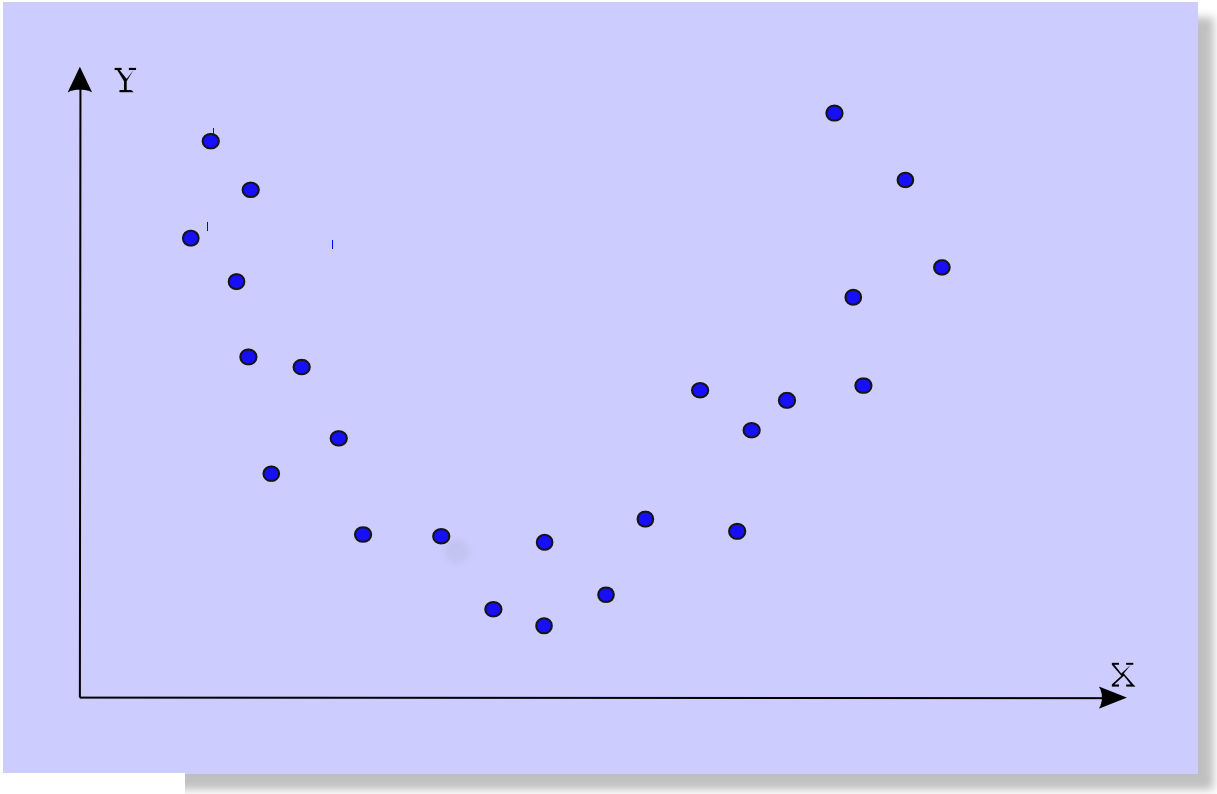
Визуально видно, что имеет место линейная зависимость, которая отрицательна. Это означает, что увеличение переменной *x* приводит к уменьшению второй переменной *y*.

Студент	Пропущено *x*	Оценка *y*
A	6	82
B	2	86
C	15	43
D	9	74
E	12	58
F	5	90
G	8	78



Нелинейная зависимость

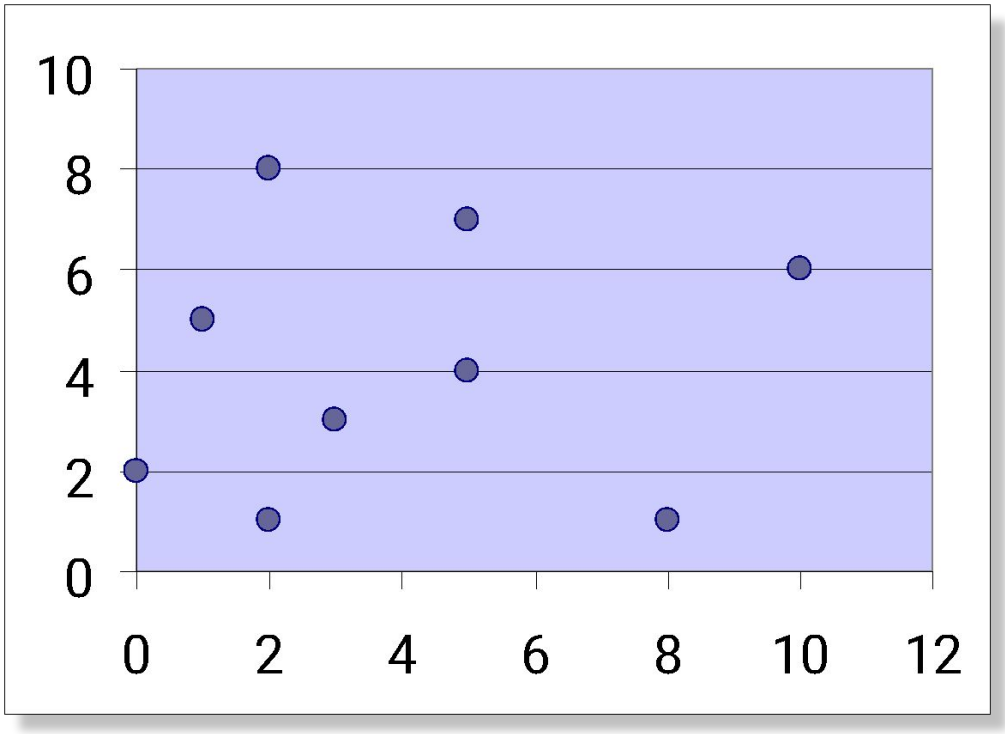
График показывает, что имеется зависимость, которая не является линейной. Возможно, эта зависимость квадратичная или какая-то иная.



Отсутствие зависимости

Студент	Часы занятий *x*	Бутылки пива *y*
A	3	3
B	0	2
C	2	1
D	5	7
E	8	1
F	5	4
G	10	6
H	2	8
I	1	5

График сообщает нам об отсутствии зависимости продолжительности занятий в неделю от количества выпиваемого пива (в бутылках).



7.1. Корреляция

Связь между двумя переменными

Коэффициент корреляции

Коэффициент корреляции измеряет силу и направление связи между двумя переменными.

Коэффициент корреляции

$$x_1, \dots, x_n \quad y_1, \dots, y_n$$

$\overline{x}$ выборочное среднее по x

$\overline{y}$ выборочное среднее по y

$x_1, \dots, x_n \quad y_1, \dots, y_n$

$\overline{x}$ выборочное среднее по x

$\overline{y}$ выборочное среднее по y

s_x^2 выборочная дисперсия по x

s_y^2 выборочная дисперсия по y

$$x_1, \dots, x_n \quad y_1, \dots, y_n$$

$$\operatorname{cov}(x, y) = \overline{x \cdot y} - \overline{x} \cdot \overline{y}$$

выборочная ковариация

$$\overline{x \cdot y} = \frac{1}{n} \sum_{i=1}^{n} x_i y_i$$

$$x_1, \dots, x_n \quad y_1, \dots, y_n$$

$$\operatorname{cov}(x, y) = \overline{x \cdot y} - \bar{x} \cdot \bar{y}$$

выборочная ковариация

$$\operatorname{cor}(x, y) = \frac{\operatorname{cov}(x, y)}{\sqrt{s_x^2 \cdot s_y^2}}$$

выборочный коэффициент корреляции

Свойства коэффициента корреляции

1) $-1 \le \operatorname{co} r(x, y) \le 1$

Свойства коэффициента корреляции

1) $-1 \le \mathrm{co}\, r(x, y) \le 1$

2) Если $y_i = ax_i + b$ для всех i=1,…n, то

$\mathrm{co}\, r(x, y) = 1$ при a>0

$\mathrm{co}\, r(x, y) = -1$ при a<0

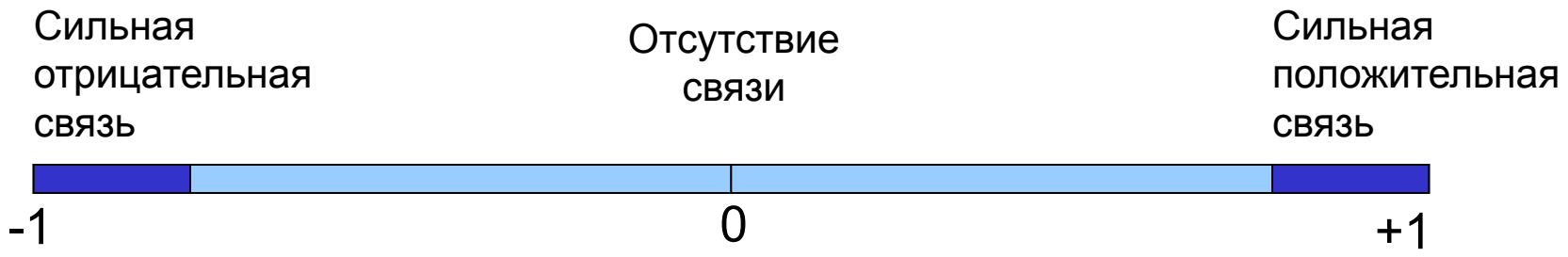
Коэффициент корреляции – мера линейной зависимости двух случайных величин

Значения коэффициента корреляции

Если между переменными существует сильная положительная связь, то значение *r* будет близко к +1.

Если между переменными существует сильная отрицательная связь, то значение *r* будет близко к –1.

Когда между переменными нет линейной связи или она очень слабая, значение r будет близко к 0.



$$cor(x, y) = -0,97$$

$$cor(x, y) = -0,06$$

Пример вычисления

Вычислим коэффициент корреляции для примера со студентами.

Студент	Часы x	Оценка y
A	6	82
B	2	63
C	1	57
D	5	88
E	2	68
F	3	75

Шаг 1. Достроим таблицу

Достраиваем таблицу тремя столбцами и итоговой строкой. Проводим необходимые вычисления.

Студент	Часы	Оценка			
	x	y	x^2	y^2	x*y
A	6	82	36	6724	492
B	2	63	4	3969	126
C	1	57	1	3249	57
D	5	88	25	7744	440
E	2	68	4	4624	136
F	3	75	9	5625	225
Среднее	3,17	72,17	13,17	5322,50	246,00

Шаги 2-3. Подставим в формулу, получим ответ

Подставим данные в формулу и найдем *r :*

$$\operatorname{cov}(x, y) = \overline{x \cdot y} - \overline{x} \cdot \overline{y}$$

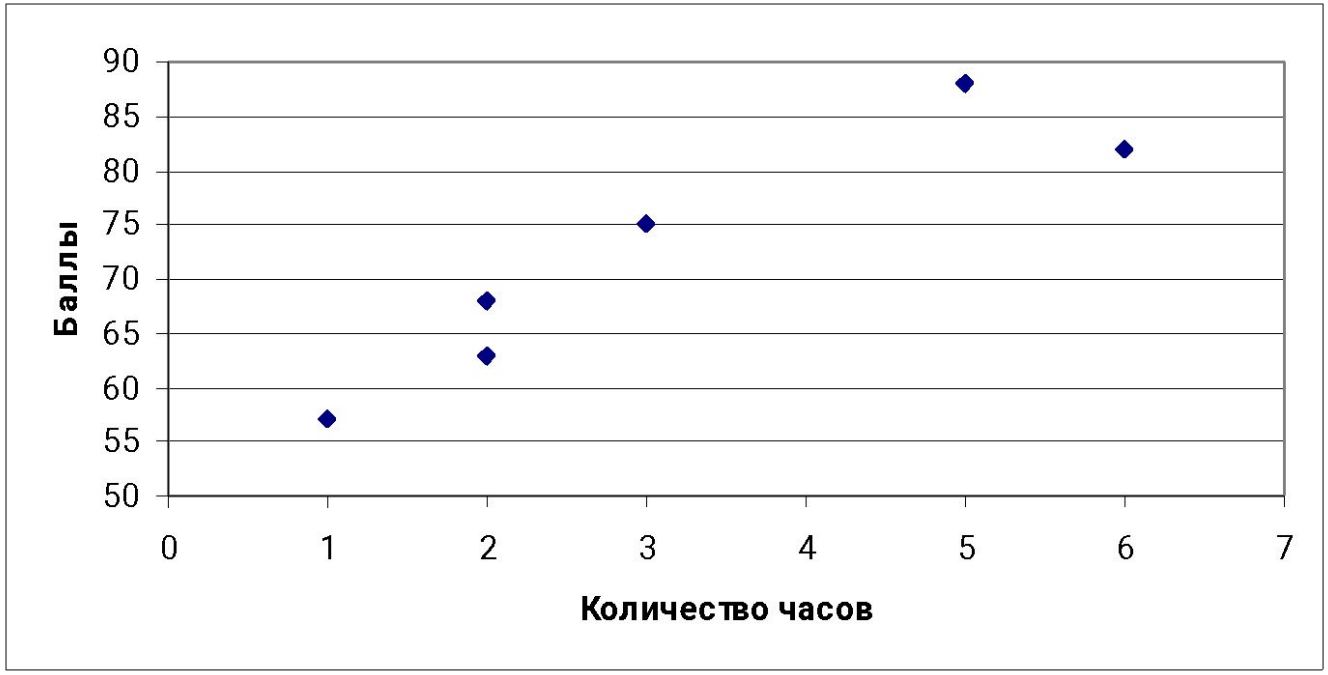
$$s_x^2 = \overline{x^2} - \left(\overline{x}\right)^2$$

$$s_y^2 = \overline{y^2} - \left(\overline{y}\right)^2$$

Ковариация	17,47
Выборочная дисперсия по x	3,14
Выборочная дисперсия по y	114,47
Коэффициент корреляции	0,92

Ответ. Значение коэффициента корреляции равно 0,92. Это означает, что существует сильная положительная связь.

Диаграмма рассеяния



Корреляция и причинная связь

Когда проверка гипотезы показывает, что существует значимая связь между переменными, необходимо получить уравнение, описывающее эту связь.

7.3. Регрессия

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

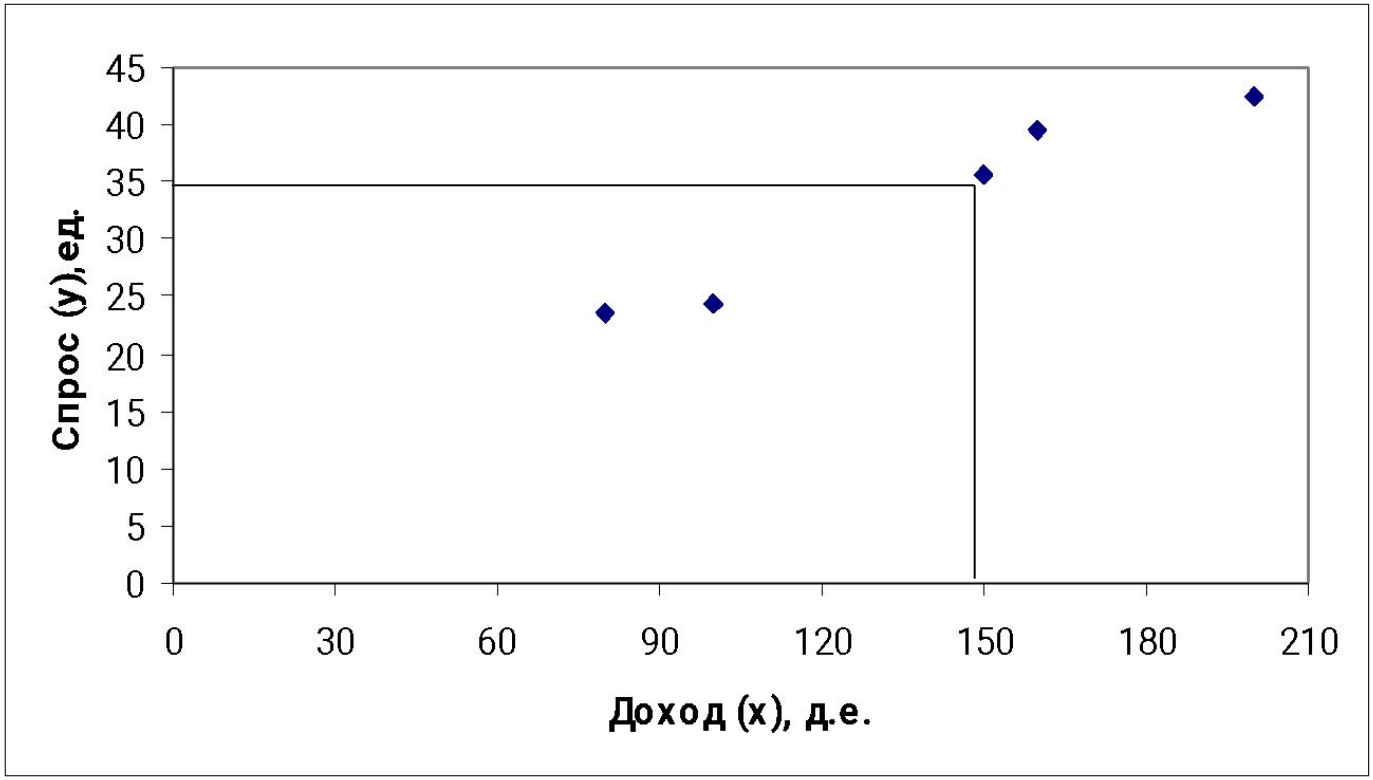
Проводится опрос домохозяйств.

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

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

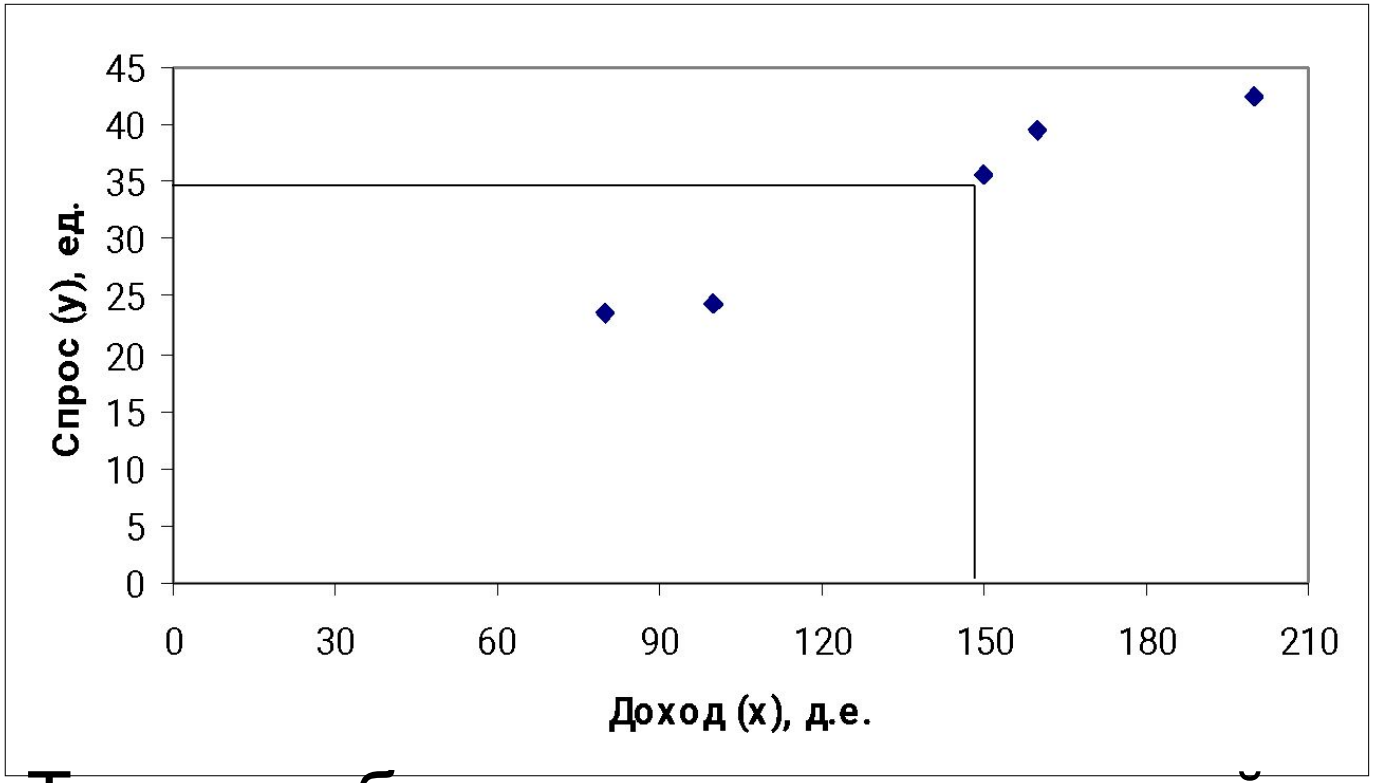
Нанесем точки на график

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



Нанесем точки на график

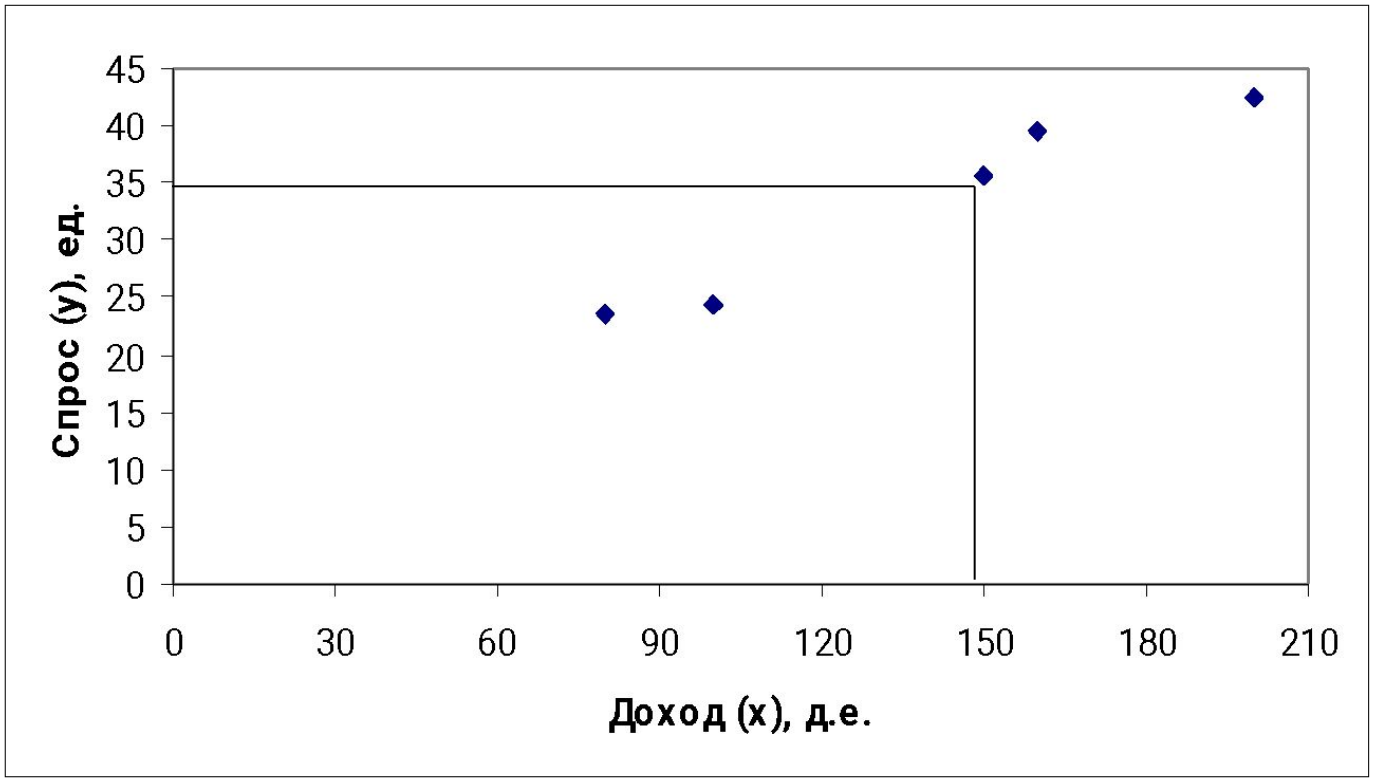
x	y
100	24
200	42
150	35
80	24
160	39



Точки разбросаны вокруг некоторой прямой!
Как ее найти?

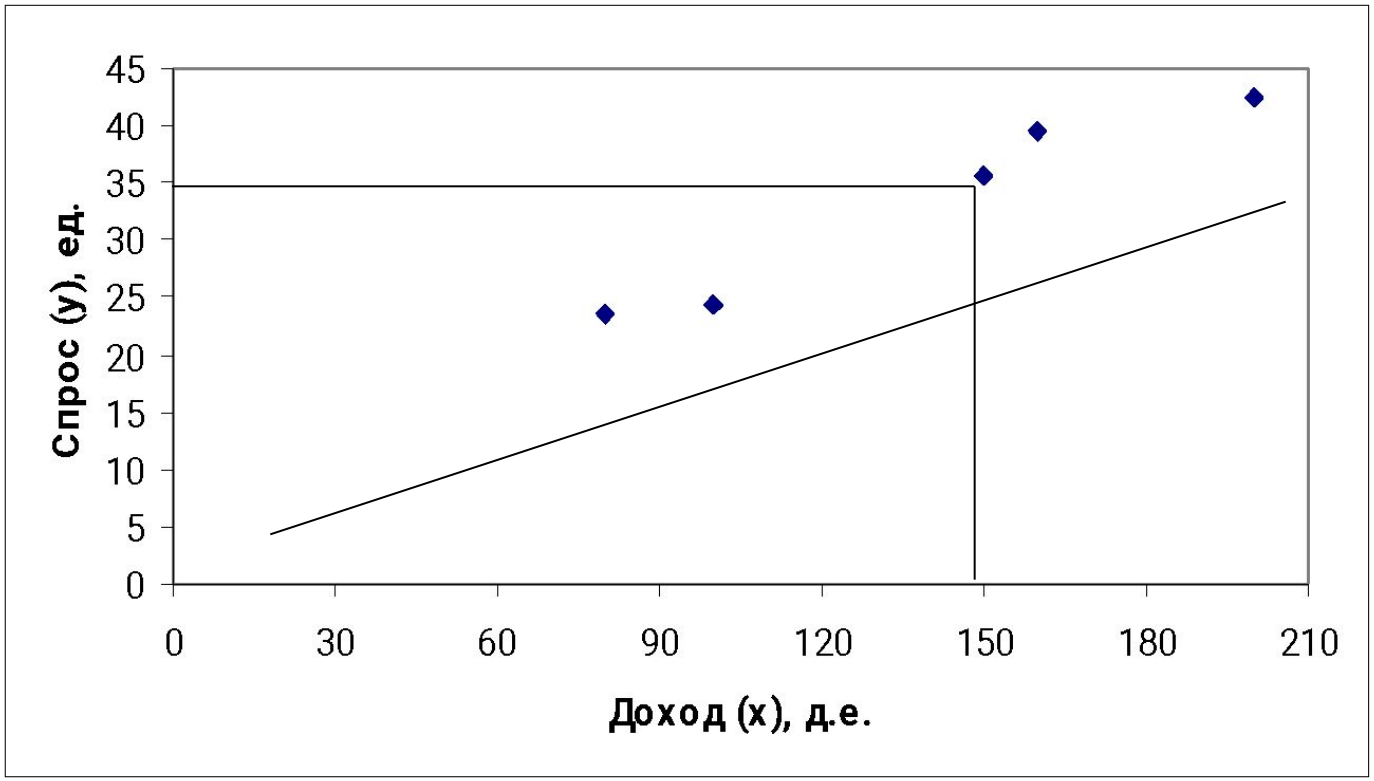
Нанесем точки на график

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



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

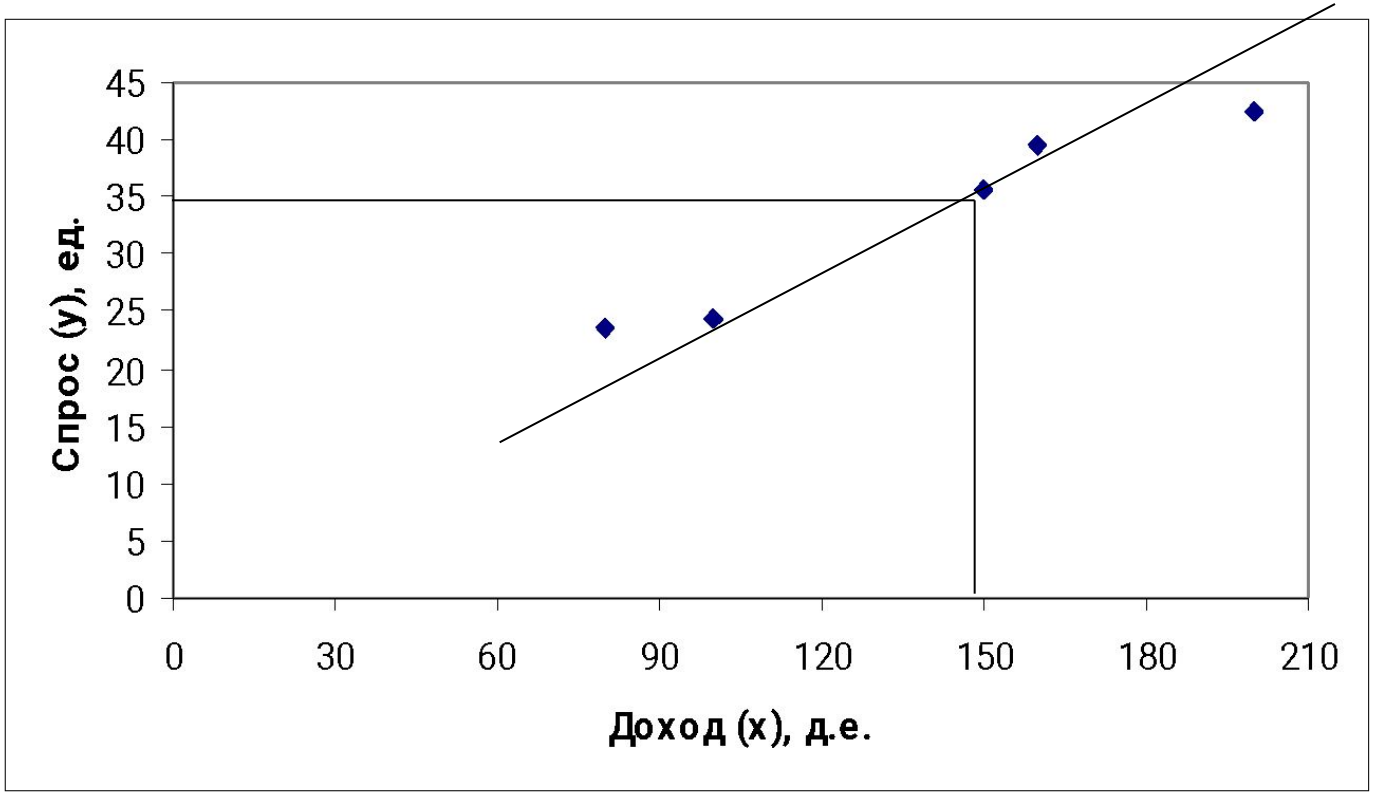
Нанесем точки на график



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

Плохая прямая!

Нанесем точки на график

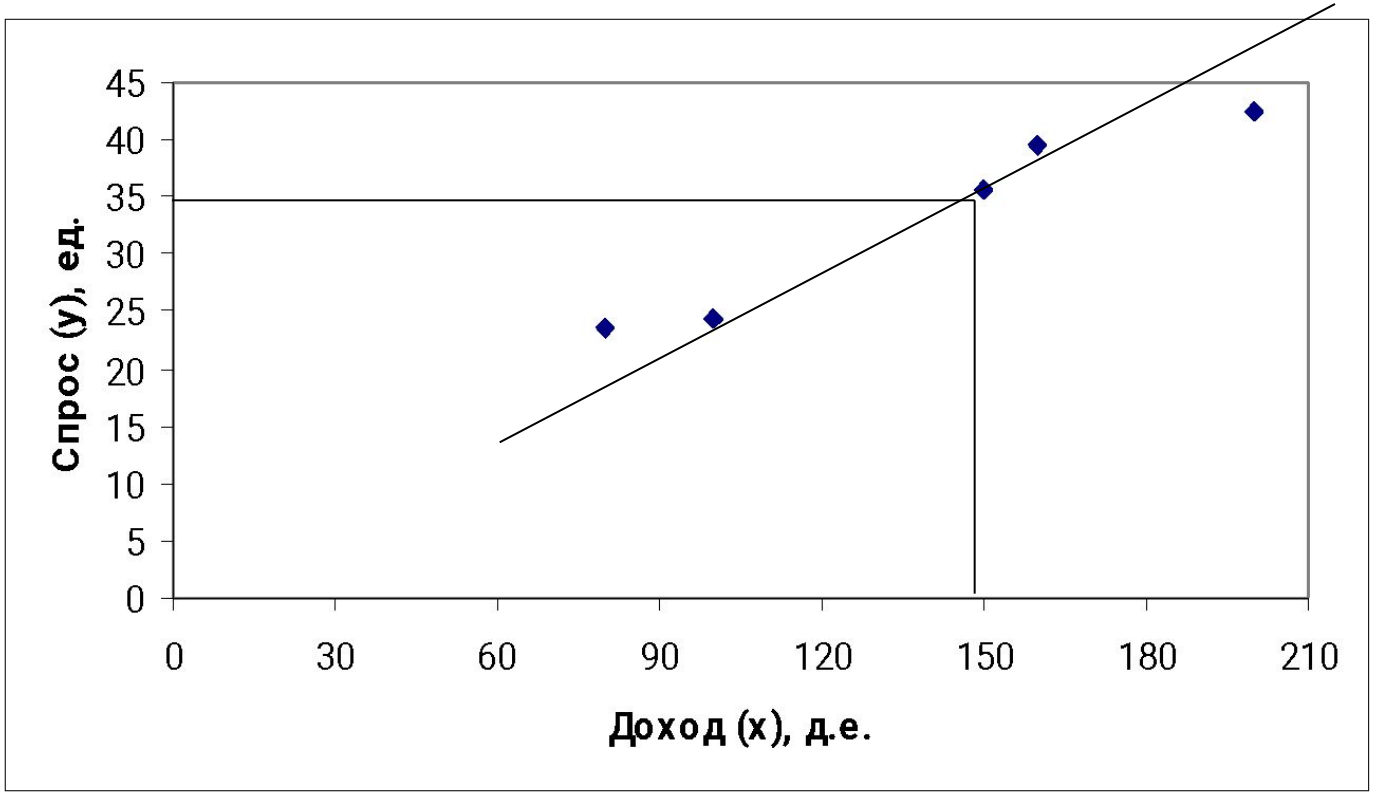


x	y
100	24
200	42
150	35
80	24
160	39

Хорошая прямая! Но может быть есть еще лучше?

Нанесем точки на график

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

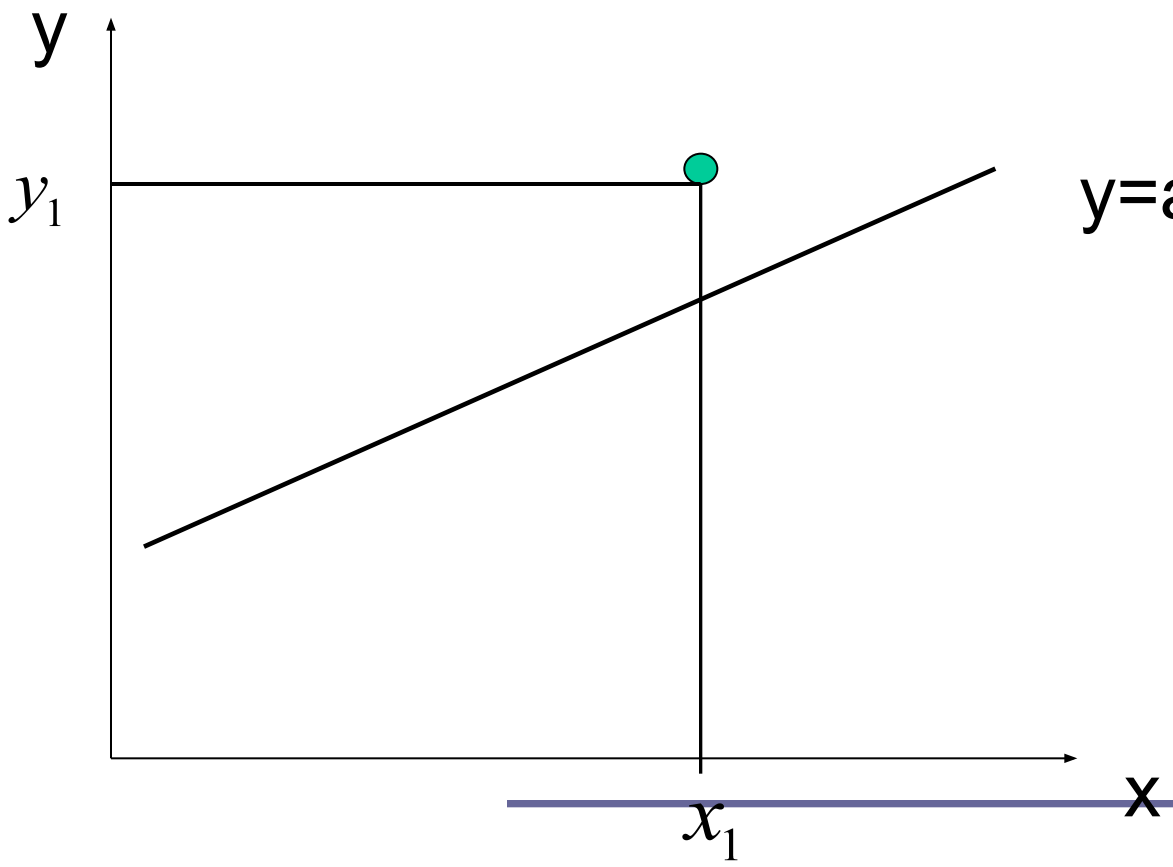


Уравнение прямой в общем виде y=ax+b. Надо найти наиболее подходящие a и b.

Обозначим

x_1 доход 1-го домохозяйства

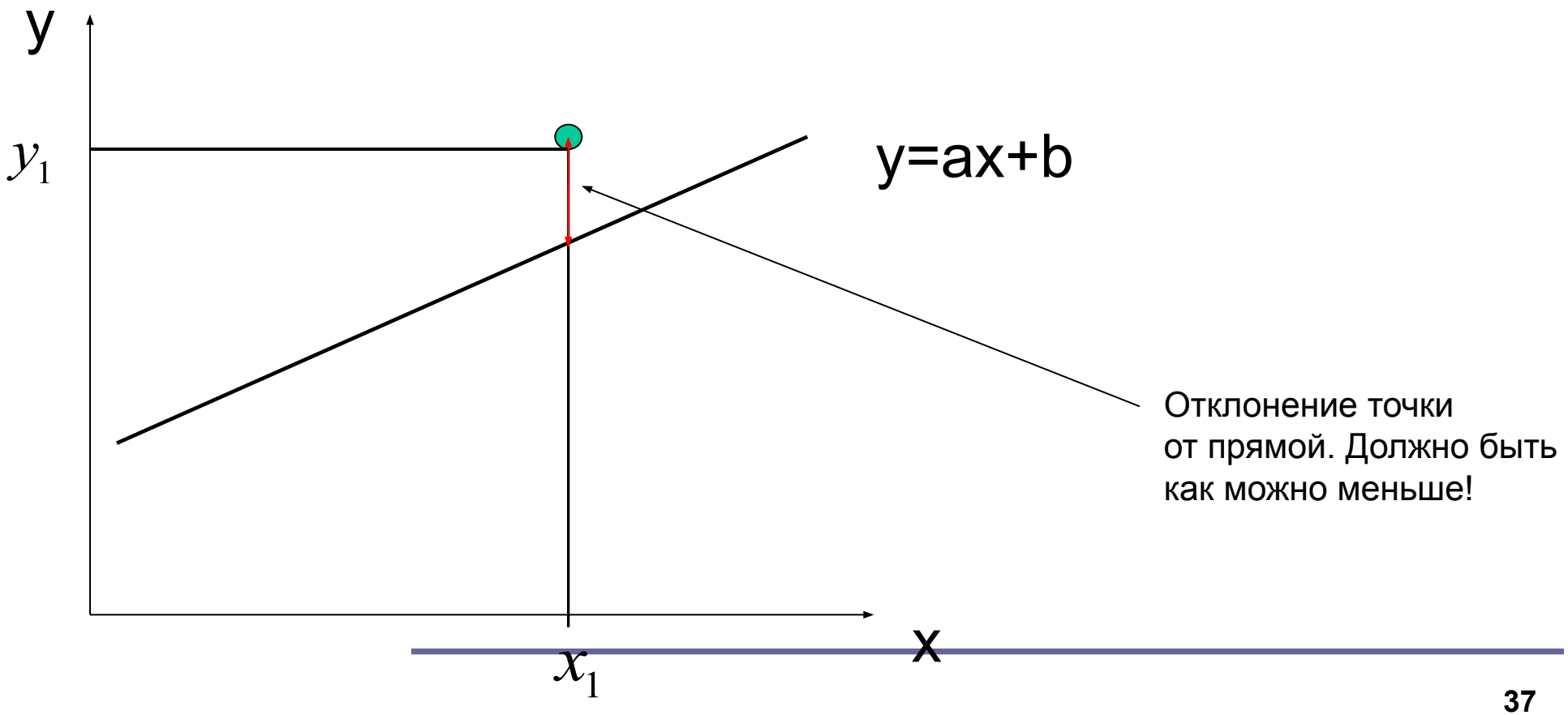
y_1 спрос 1-го домохозяйства на продукт



Обозначим

x_1 доход 1-го домохозяйства

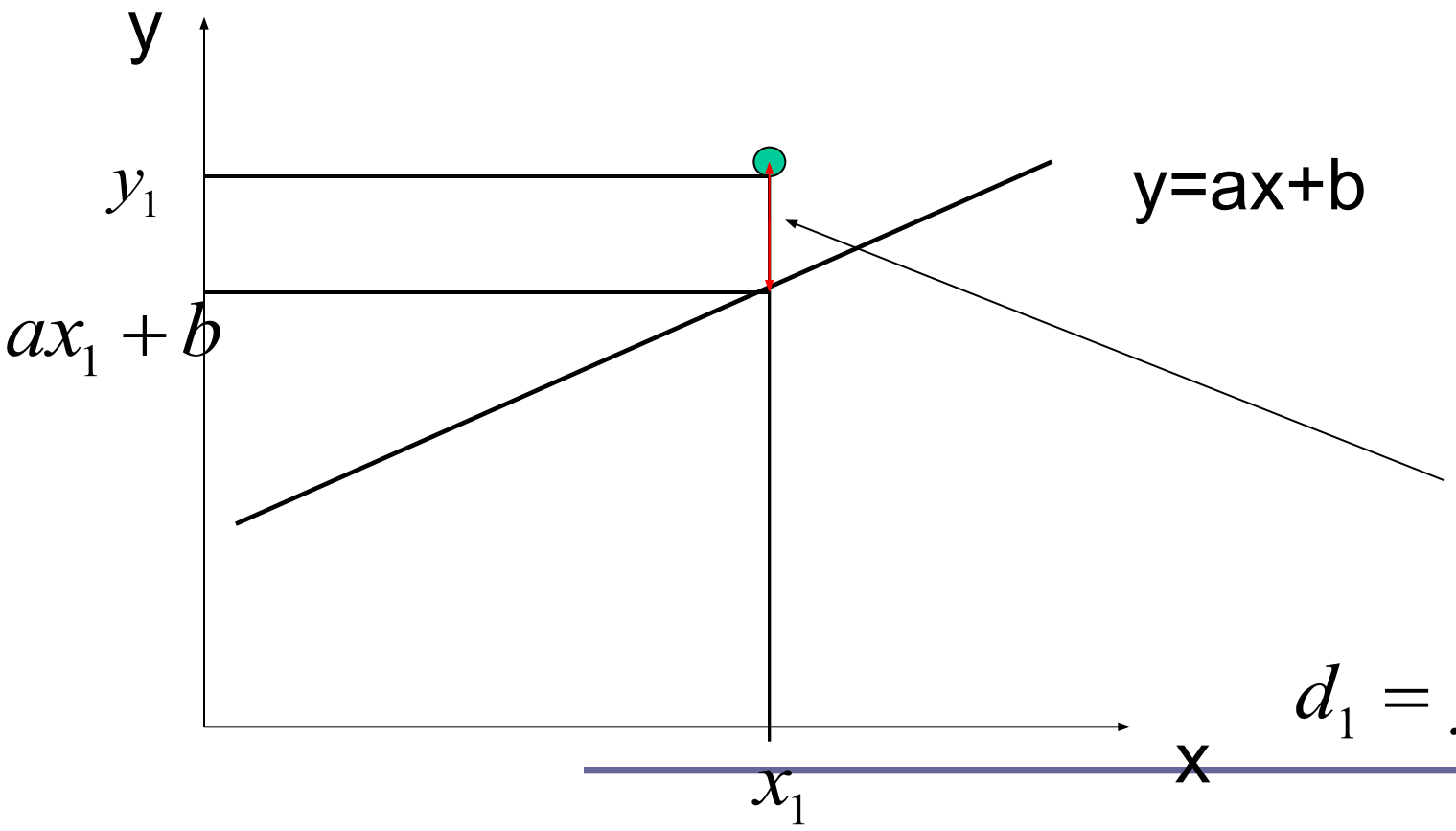
y_1 спрос 1-го домохозяйства на продукт



Обозначим

x_1 доход 1-го домохозяйства

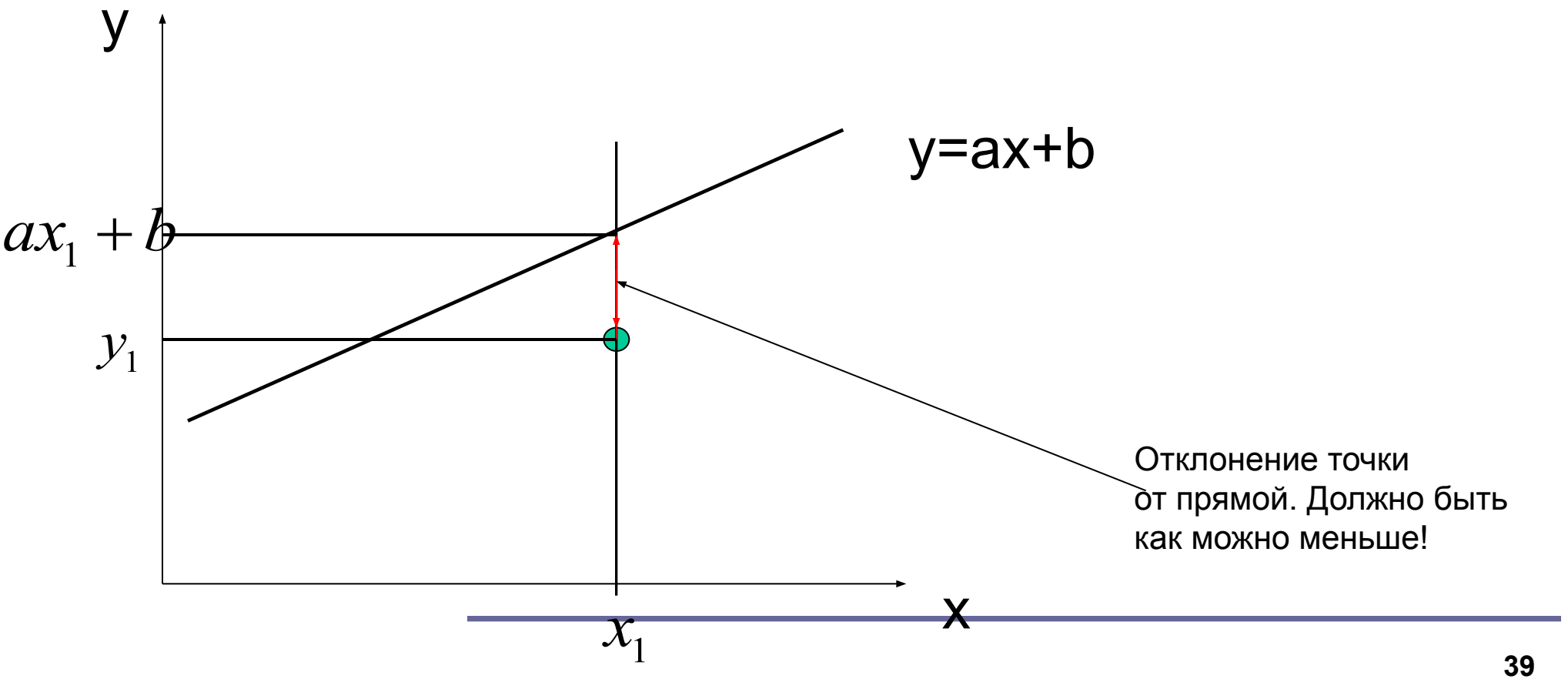
y_1 спрос 1-го домохозяйства на продукт



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

А если точка лежит ниже прямой?

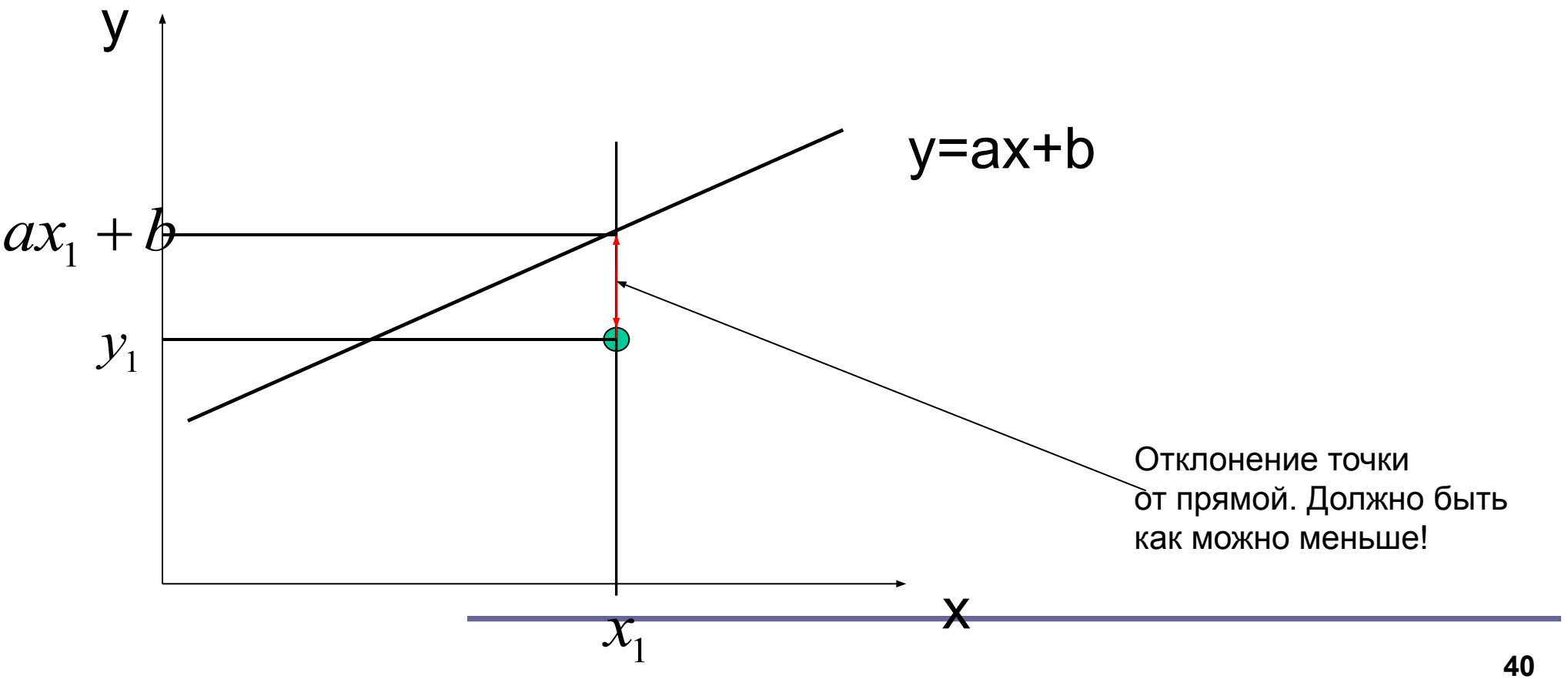
Тогда отклонение $d_1 = (ax_1 + b) - y_1$



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?

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

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

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

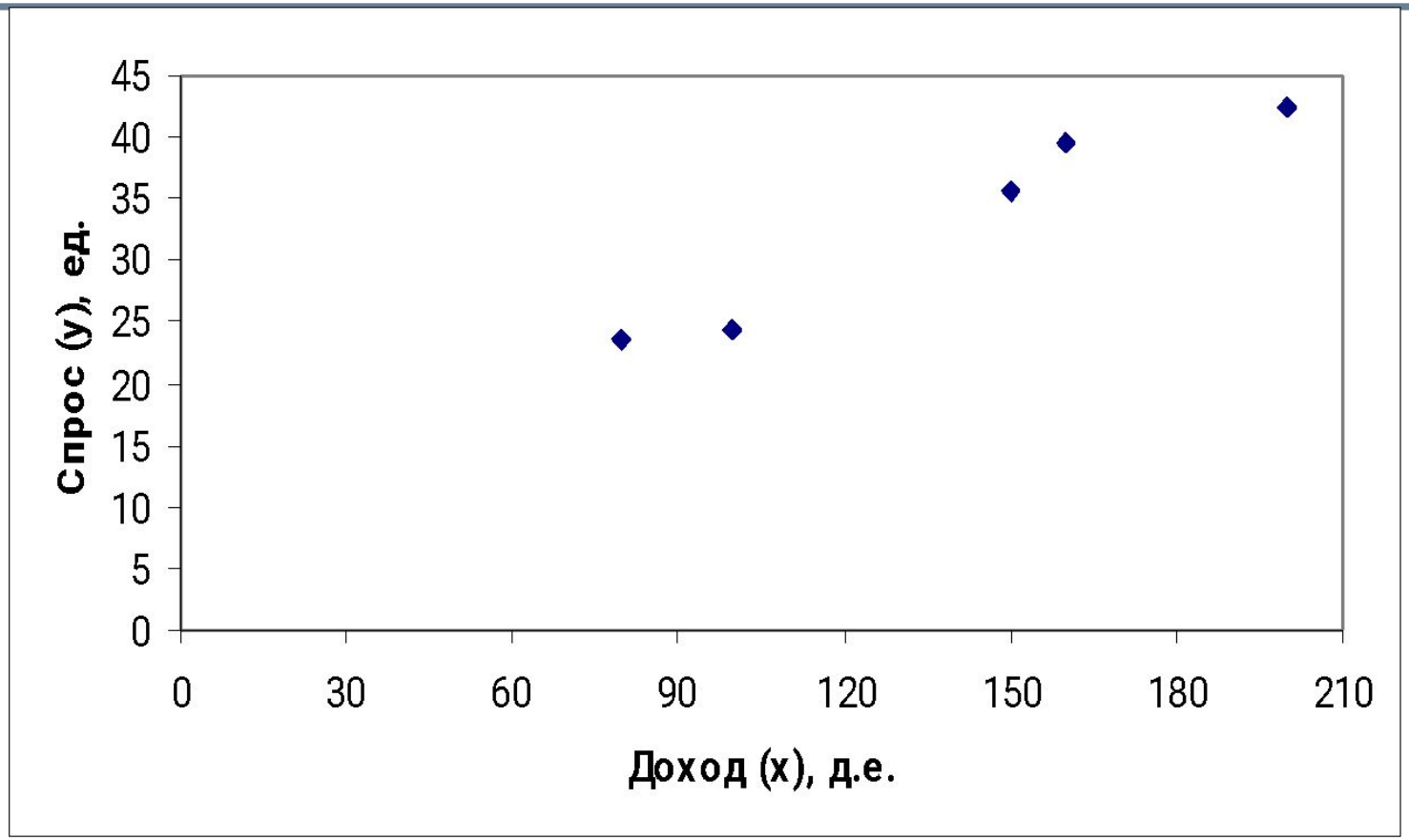
Получили функцию двух переменных, для которой надо найти минимум, т.е. надо исследовать на экстремум.

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

$$a = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - (\bar{x})^2} \qquad a = \frac{\mathrm{cov}(x,y)}{s_x^2}$$

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

Вернемся к примеру



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

Вернемся к примеру

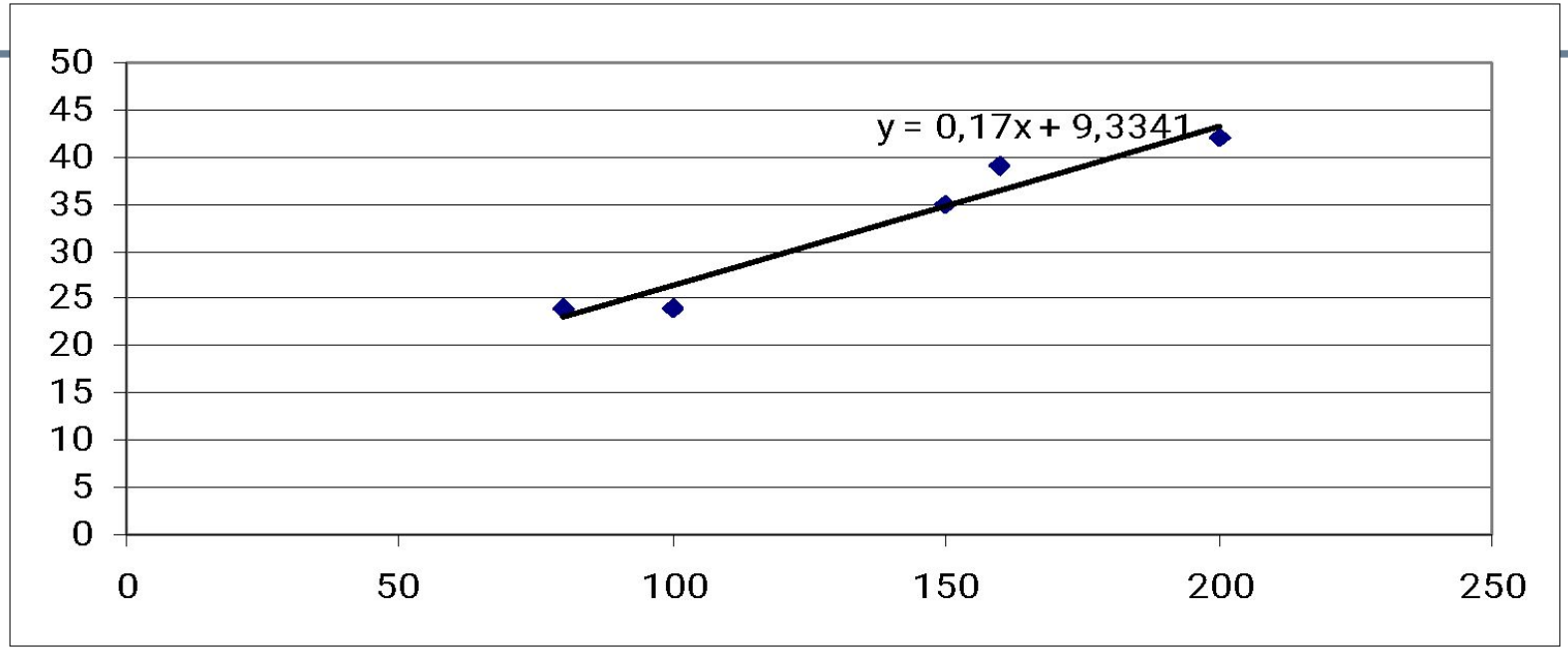
	A	B	C	D	E	F
1	N	X	Y	X^2	Y^2	X*Y
2	1	100	24	10000	576	2400
3	2	200	42	40000	1764	8400
4	3	150	35	22500	1225	5250
5	4	80	24	6400	576	1920
6	5	160	39	25600	1521	6240
7	Сумма	690	164	104500	5662	24210
8						
9	Среднее по X				138	
10	Среднее по Y				32,8	
11	Выборочная дисперсия по X				1856	
12	Выборочная дисперсия по Y				56,56	
13	Выборочная ковариация				315,6	
14	Коэффициент a				0,17	
15	Коэффициент b				9,33	

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

$$\operatorname{cov}(x, y) = \overline{x \cdot y} - \overline{x} \cdot \overline{y}$$

$$a = \frac{\operatorname{cov}(x, y)}{s_x^2}$$

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



y=0,17x+9,33 - 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=0,17x+9,33 - функция спроса в зависимости от дохода.

Пример вычисления

Найдем линейное уравнение регрессии для нашего примера.

Студент	Часы x	Оценка y
A	6	82
B	2	63
C	1	57
D	5	88
E	2	68
F	3	75

Шаг 1. Достроим таблицу

Проводим необходимые вычисления.

$$a = \frac{\operatorname{cov}(x, y)}{s_x^2}$$

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

Ковариация	17,47
Выборочная дисперсия по x	3,14
Выборочная дисперсия по y	114,47
Коэффициент корреляции	0,92

Коэффициент a	5,57
Коэффициент b	54,54

Ответ. Получили уравнение «наилучшей прямой»:

y = 5,57 x + 54,54

Интерпретация

1. Увеличение времени подготовки на 1 час приводит к улучшению результата на 5,57 балла.

2. Если не заниматься вообще – получишь 54,5 балла.

Интерпретация некорректна, выходим за границы анализируемой области!

$$y = 5{,}57\,x + 54{,}54$$

Отчет из Excel

$$y = 5{,}57\,x + 54{,}54$$

Отчет о расчете коэффициентов регрессии, полученный из Excel.

www.zeallsoft.com	B	C	D	E	F	G	H	I
ВЫВОД ИТОГОВ								
рессионная статистика								
Множеств	0,92							
R-квадрат	0,85							
Нормиров	0,81							
Стандартн	5,08							
Наблюден	6							
Дисперсионный анализ								
	df	*SS*	*MS*	*F*	*начимость F*			
Регрессия	1	583,5413	583,5413	22,59773	0,008946			
Остаток	4	103,292	25,82301					
Итого	5	686,8333						
	Коэффициенгдартная остатистиP-Значение				*ижние 95%*	*ерхние 95%*	*ижние 95,0'*	*Верхние 95,0%*
Y-пересеч	54,540	4,249	12,836	0,000	42,743	66,337	42,743	66,337
Переменн	5,566	1,171	4,754	0,009	2,315	8,817	2,315	8,817

Будьте осторожны с прогнозами!

Когда прогнозы распространяются за пределы исследуемых данных, интерпретировать результаты необходимо с особой осторожностью.

Помните, что, когда делаются прогнозы, они основываются на текущих условиях или на предположении, что существующие ныне тенденции продолжатся в будущем. Это предположение может оправдаться или не оправдаться.