

# СТАТИСТИКА



---

---

## Аналитическая статистика.

---

---

### Лекция 3. Статистическое изучение взаимосвязи социально-экономических явлений.

---

---

Автор: Равичев Л.В.

РХТУ им. Д.И.Менделеева

Кафедра управления технологическими инновациями

Москва - 2013

# Корреляционный и регрессионный анализ

Основная задача статистики – обнаружить связь между явлениями, её вид и дать количественную характеристику этой связи.

**Вид связи между явлениями**

```
graph TD; A[Вид связи между явлениями] --> B[Функциональная]; A --> C[Статистическая]
```

**Функциональная**

**Статистическая**

# Корреляционный и регрессионный анализ

Предмет корреляционно-регрессионного анализа составляет исследование статистических зависимостей между явлениями.

**Корреляционный анализ**

Существует ли связь между явлениями?

Насколько сильная связь между явлениями?

**Регрессионный анализ**

Каков характер связи между явлениями?

Построение регрессионной модели явлений.

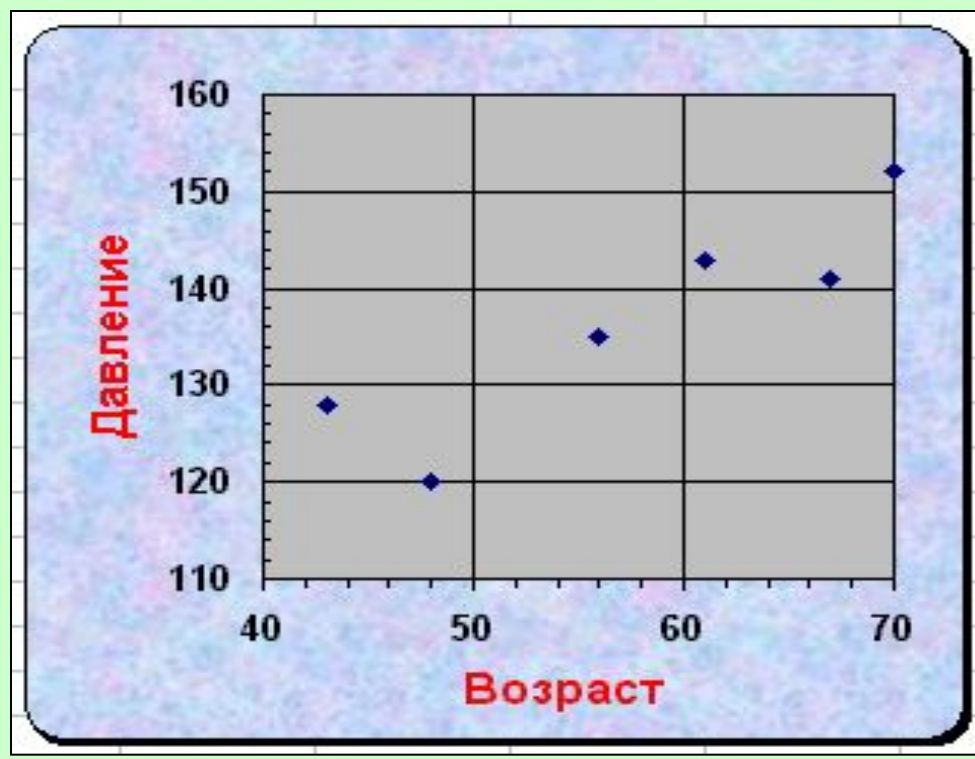
# Корреляционный анализ

## Диаграмма рассеяния

Простейшим приемом при исследовании зависимости между двумя количественными признаками является построение **диаграммы рассеяния**.

**Пример 1.** Построить диаграмму рассеяния для результатов наблюдения за возрастом и артериальным давлением группы людей, приведенных в таблице.

№	Возраст, лет (x)	Давление, мм.рт.ст. (y)
1	43	128
2	48	120
3	56	135
4	61	143
5	67	141
6	70	152



# Корреляционный анализ

## Линейный коэффициент корреляции Пирсона

Наиболее часто употребляемой количественной характеристикой линейных зависимостей между признаками является *линейный коэффициент корреляции Пирсона*:

$$r = \frac{(x - \tilde{x}) \cdot (y - \tilde{y})}{\sigma_{\tilde{x}} \cdot \sigma_{\tilde{y}}}$$



$$r = \frac{\overline{xy} - \tilde{x} \cdot \tilde{y}}{\sigma_{\tilde{x}} \cdot \sigma_{\tilde{y}}}$$



$$r = \frac{\sum (x_i - \tilde{x}) \cdot (y_i - \tilde{y})}{n \cdot \sigma_{\tilde{x}} \cdot \sigma_{\tilde{y}}}$$

# Корреляционный анализ

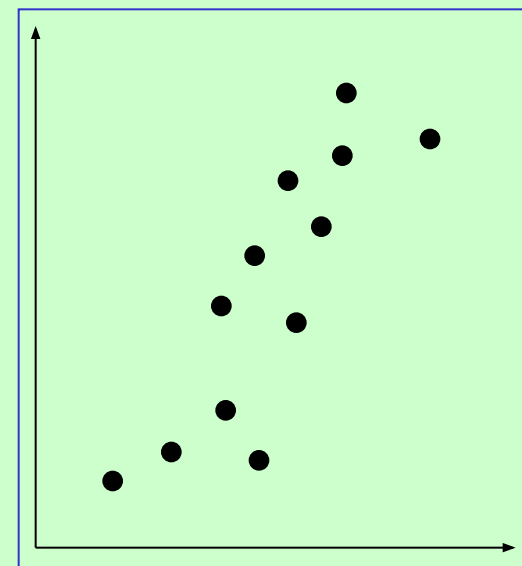
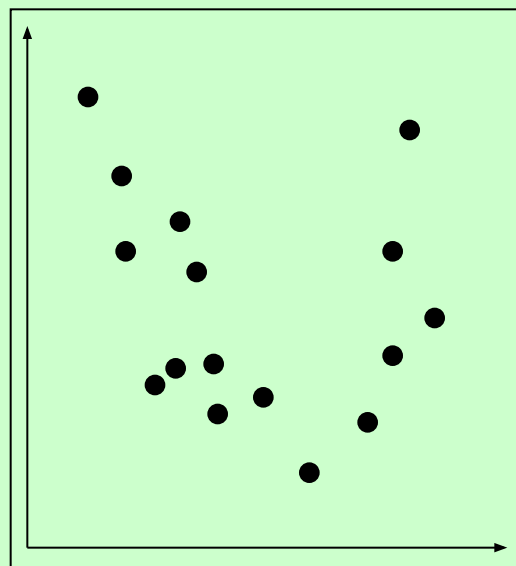
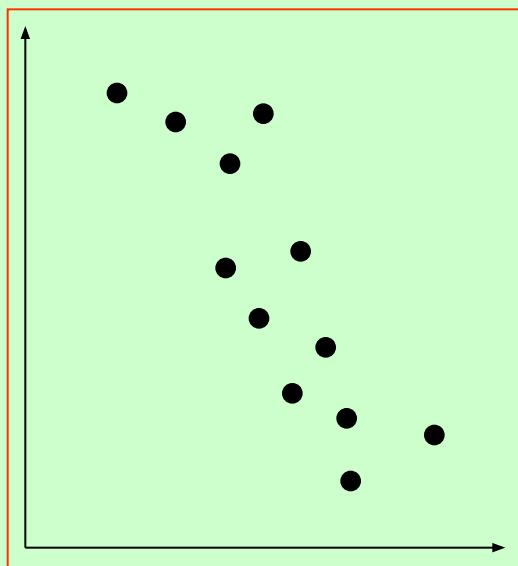
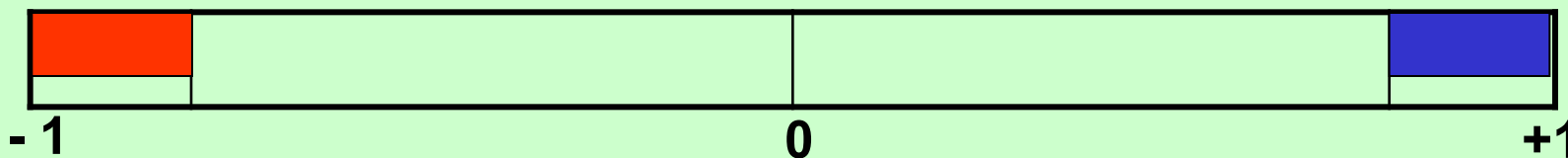
## Линейный коэффициент корреляции Пирсона

Основные свойства коэффициента корреляции:

**Сильная  
обратная  
связь**

Нет  
линейной  
связи

**Сильная  
прямая  
связь**



# Корреляционный анализ

## Линейный коэффициент корреляции Пирсона

**Пример 2.** Для данных, приведенных в примере 1 вычислить линейный коэффициент корреляции Пирсона и оценить тип связи между величинами.

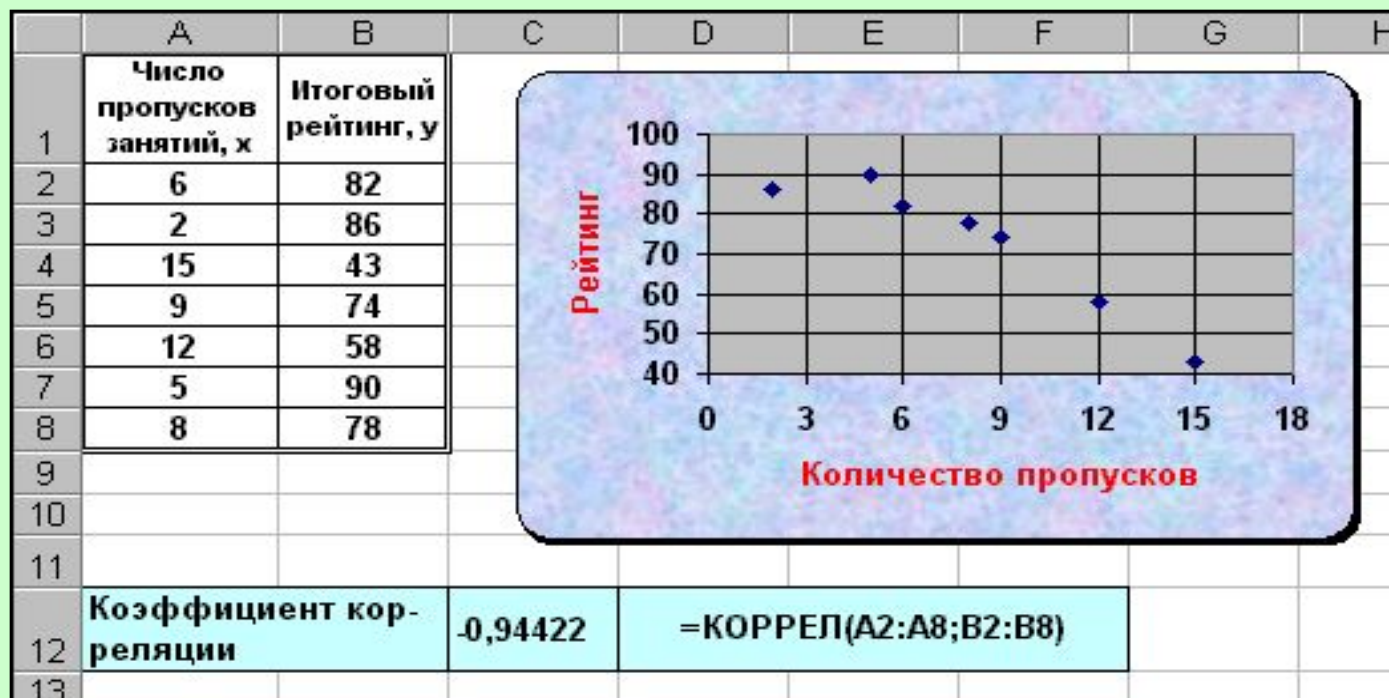
	A	B	C	D	E	F
	<b>№</b>	<b>Возраст, лет (x)</b>	<b>Давление, мм.рт.ст. (y)</b>	$(x_i - x_{cp})/ст.х$	$(y_i - y_{cp})/ст.у$	
1						
2	1	43	128	-1,50	-0,82	
3	2	48	120	-0,98	-1,59	
4	3	56	135	-0,16	-0,14	
5	4	61	143	0,36	0,62	
6	5	67	141	0,98	0,43	
7	6	70	152	1,29	1,49	
8	Среднее x		57,50	=СРЗНАЧ(B2:B7)		
9	Среднее y		136,50	=СРЗНАЧ(C2:C7)		
10	Стандарт x		9,67	=СТАНДОТКЛОНП(B2:B7)		
11	Стандарт y		10,40	=СТАНДОТКЛОНП(C2:C7)		
12	Коэффициент корреляции		0,897	=СУММПРОИЗВ(D2:D7;E2:E7)/6		
13	Коэффициент корреляции, рассчитанный через <b>КОРРЕЛ</b>		0,897	=КОРРЕЛ(B2:B7;C2:C7)		
14						

# Корреляционный анализ

## Линейный коэффициент корреляции Пирсона

**Пример 3.** Для данных, приведенных в таблице построить диаграмму рассеяния и вычислить коэффициент корреляции для группы студентов (7 человек).

<b>Число пропусков занятий, x</b>	<b>6</b>	<b>2</b>	<b>15</b>	<b>9</b>	<b>12</b>	<b>5</b>	<b>8</b>
<b>Итоговый рейтинг, y</b>	<b>82</b>	<b>86</b>	<b>43</b>	<b>74</b>	<b>58</b>	<b>90</b>	<b>78</b>



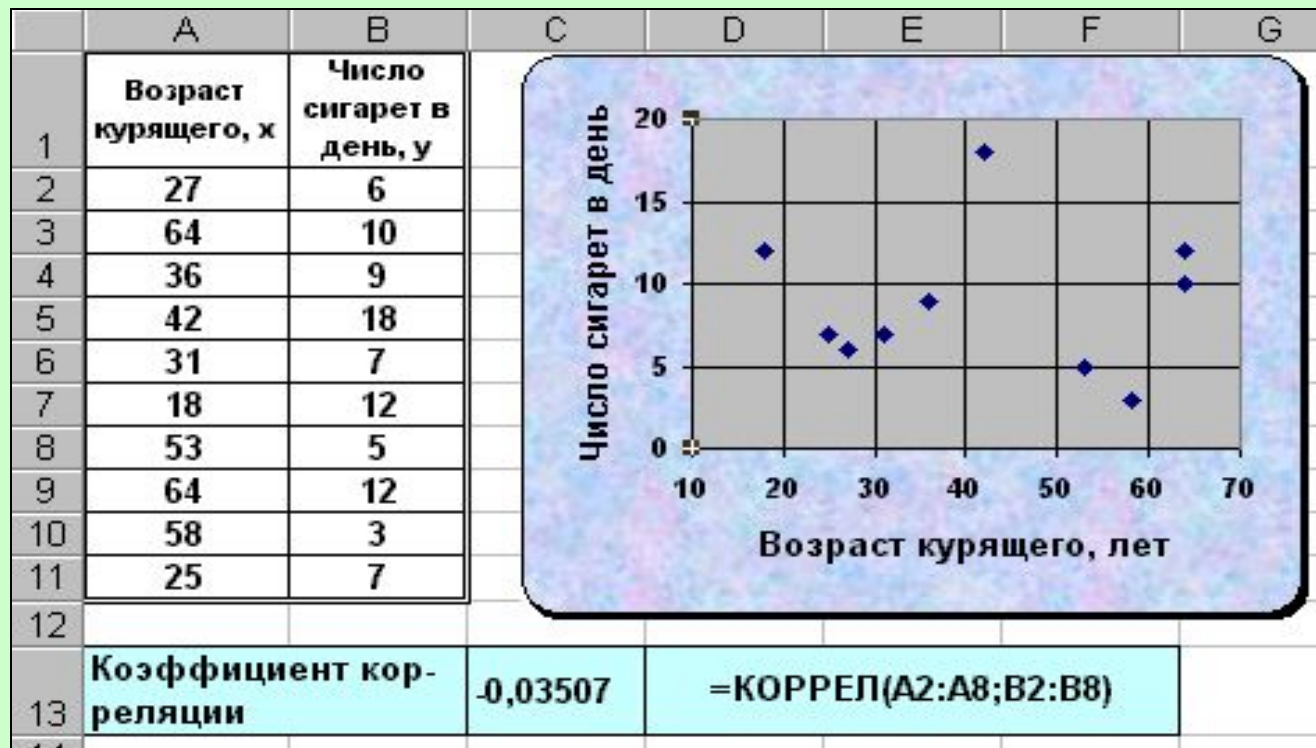


# Корреляционный анализ

## Линейный коэффициент корреляции Пирсона

**Пример 4.** В таблице приведены данные для группы курящих людей. Построить диаграмму рассеяния и вычислить коэффициент корреляции.

Возраст курящего, x	27	64	36	42	31	18	53	64	58	25
Число сигарет в день, y	6	10	9	18	7	12	5	12	3	7



## Проверка значимости коэффициента корреляции

Линейный коэффициент корреляции для генеральной совокупности:

$$\rho = \frac{\sum (x_i - \bar{x}) \cdot (y_i - \bar{y})}{N \cdot \sigma_{\bar{x}} \cdot \sigma_{\bar{y}}}$$

Критерий Стьюдента для коэффициента корреляции:

$$t_p = |r| \sqrt{\frac{n-2}{1-r^2}}$$

При большом числе наблюдений ( $n > 100$ ):

$$t_p = |r| \sqrt{\frac{n}{1-r^2}}$$

## Проверка значимости коэффициента корреляции

Оценка значимости коэффициента корреляции проводится с помощью *аппарата проверки гипотез*.

Относительно *генерального* коэффициента корреляции можно выдвинуть две гипотезы:

- генеральный коэффициент корреляции равен 0 (основная гипотеза);
- генеральный коэффициент корреляции отличен от 0.

**Сформировав выборку и рассчитав её коэффициент корреляции  $r$ , необходимо решить – является ли его значение настолько большим, чтобы вероятность (по различным выборкам) выпадения такого значения при нулевом генеральном коэффициенте корреляции  $\rho$  была бы мала (меньше уровня значимости). Если является, то в этом случае основная гипотеза отвергается, а коэффициент корреляции и установленная зависимость между величинами полагаются значимыми.**

## Проверка значимости коэффициента корреляции

**Пример 5.** Исследовать значимость коэффициента корреляции, рассчитанного в примере 2.

1) Сформулируем проверяемые утверждения:

$H_0: \rho=0$  (в генеральной совокупности нет зависимости, найденная зависимость случайна);

$H_1: \rho \neq 0$  (найденная зависимость справедлива для генеральной совокупности).

2) Находим критическое значение критерия Стьюдента:

$$\text{при } p=0,05 \text{ и } k=6-1=5 \quad t_{\text{кр}}=2,571$$

3) Находим расчетное значение критерия Стьюдента:

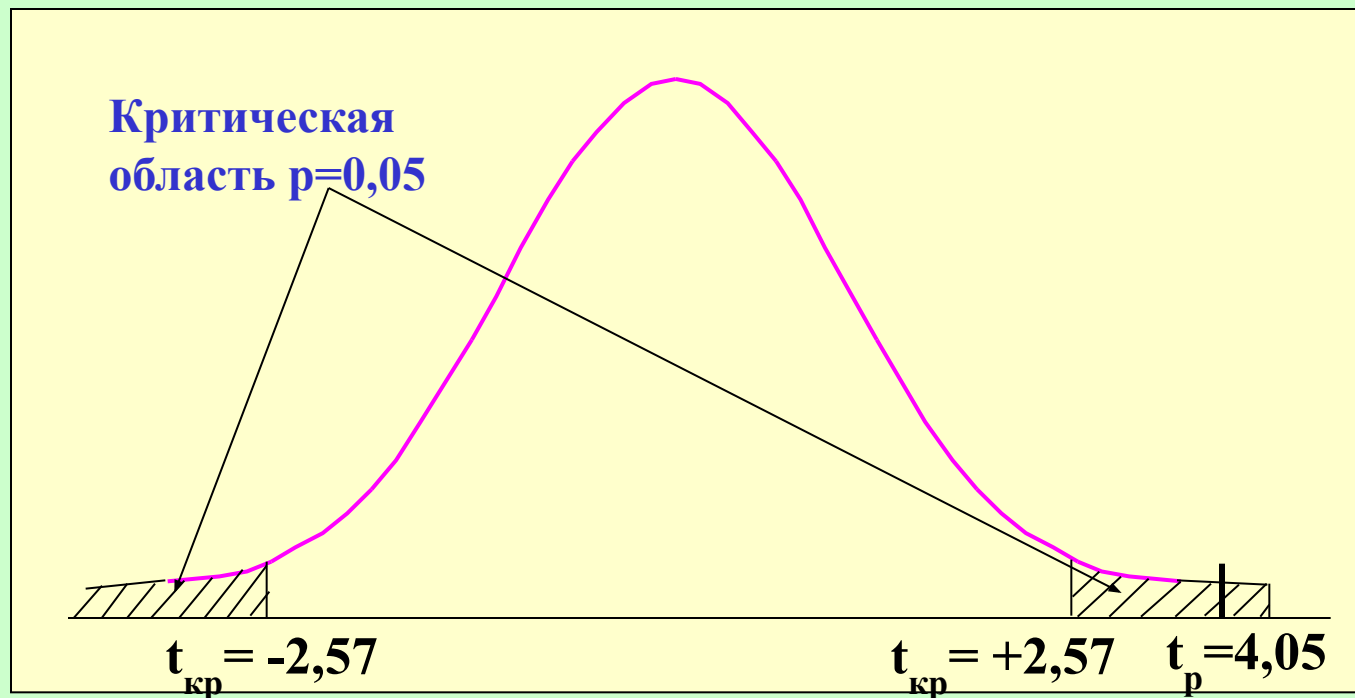
$$t_p=4,059$$

4) Находим критическую область значения критерия Стьюдента:

$$|t_p| \geq t_{\text{кр}}$$

## Проверка значимости коэффициента корреляции

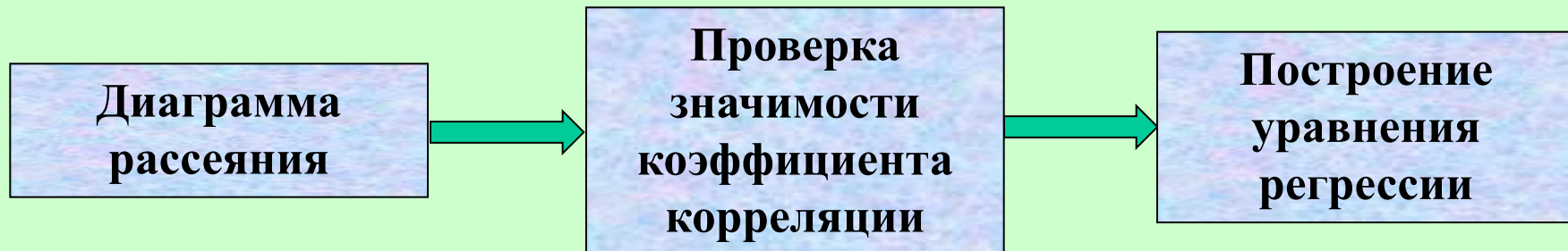
5) Принятие решения. Значение критерия попадает в критическую область:



основная гипотеза отклоняется.

Вывод: прямая зависимость между возрастом человека и артериальным давлением является значимой и её можно распространить на всю совокупность пациентов.

# Регрессионный анализ



Наиболее распространенным способом построения уравнения регрессии является *метод наименьших квадратов (МНК)*.

Метод МНК для получения уравнения регрессии основан на минимизации суммы квадратов остатков:

$$S = \sum_{i=1}^n [y_i - (a_0 + \varphi(a_i, x_i))]^2 \Rightarrow \min$$

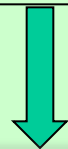
Уравнение регрессии является линейным относительно коэффициентов  $a_j$  ( $j=0,1,\dots,n$ ).

# Регрессионный анализ

## Парная линейная регрессия

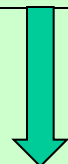
Для уравнения линейной регрессии:

$$S = \sum_{i=1}^n [y_i - (a_0 + a_1 x_i)]^2 \Rightarrow \min$$




$$\frac{\partial S}{\partial a_0} = 2 \sum_{i=1}^n [(y_i - a_0 - a_1 x_i)(-1)] = 0$$

$$\frac{\partial S}{\partial a_1} = 2 \sum_{i=1}^n [(y_i - a_0 - a_1 x_i)(-x_i)] = 0$$




# Регрессионный анализ

## Парная линейная регрессия


$$na_0 + a_1 \sum_{i=1}^n (x_i) = \sum_{i=1}^n (y_i)$$

$$a_0 \sum_{i=1}^n (x_i) + a_1 \sum_{i=1}^n (x_i^2) = \sum_{i=1}^n (x_i y_i)$$

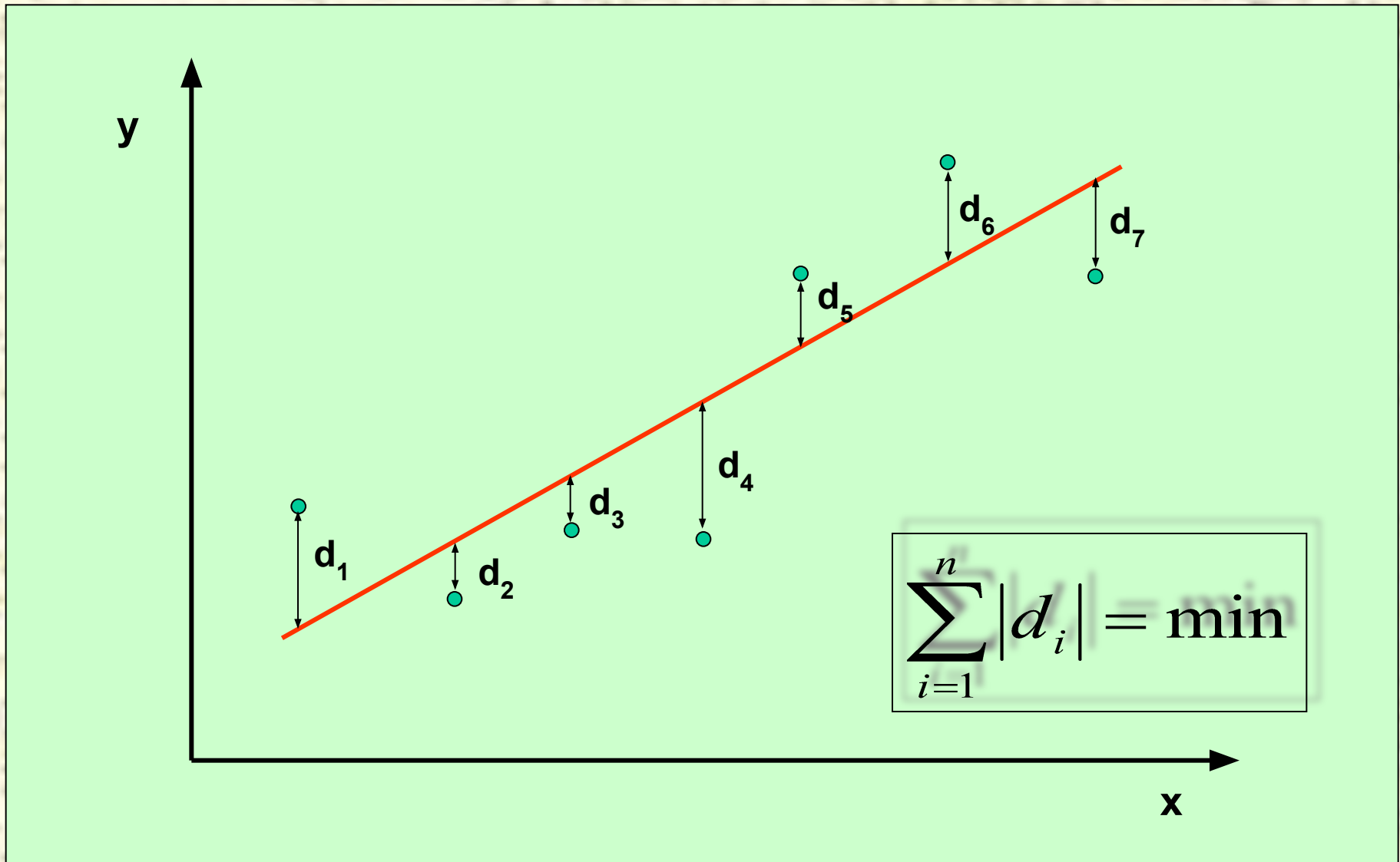

$$a_0 = \frac{(\sum y_i)(\sum x_i^2) - (\sum x_i)(\sum x_i y_i)}{n(\sum x_i^2) - (\sum x_i)^2}$$

$$a_1 = \frac{n(\sum x_i y_i) - (\sum x_i)(\sum y_i)}{n(\sum x_i^2) - (\sum x_i)^2}$$



# Регрессионный анализ

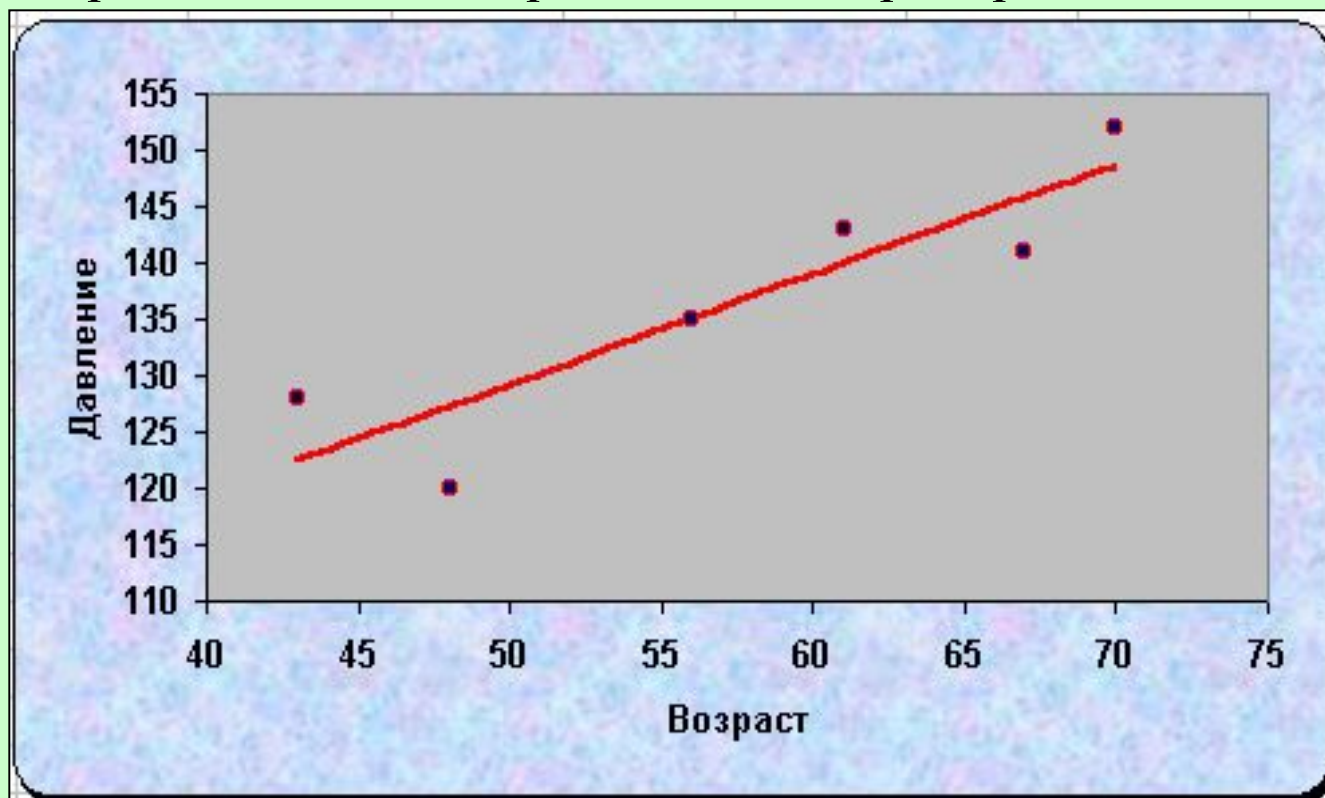
## Парная линейная регрессия



# Регрессионный анализ

## Парная линейная регрессия

**Пример 6.** Построить уравнение линейной регрессии для зависимости величин возраста и давления, приведенных в примере 1.

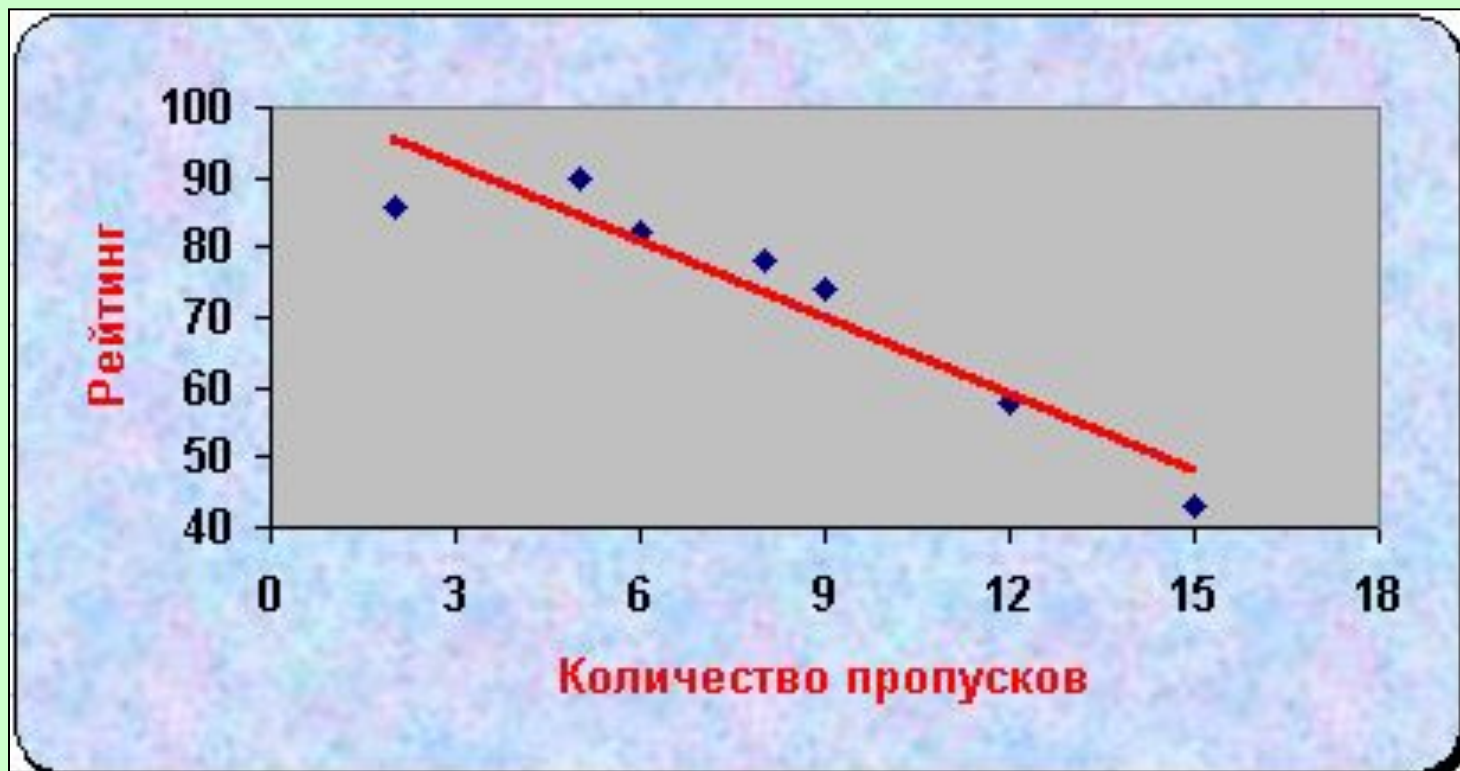


$$\hat{y} = 81,048 + 0,964x$$

# Регрессионный анализ

## Парная линейная регрессия

**Пример 7.** Построить уравнение линейной регрессии для зависимости количества пропущенных занятий и рейтинга, приведенных в примере 3.

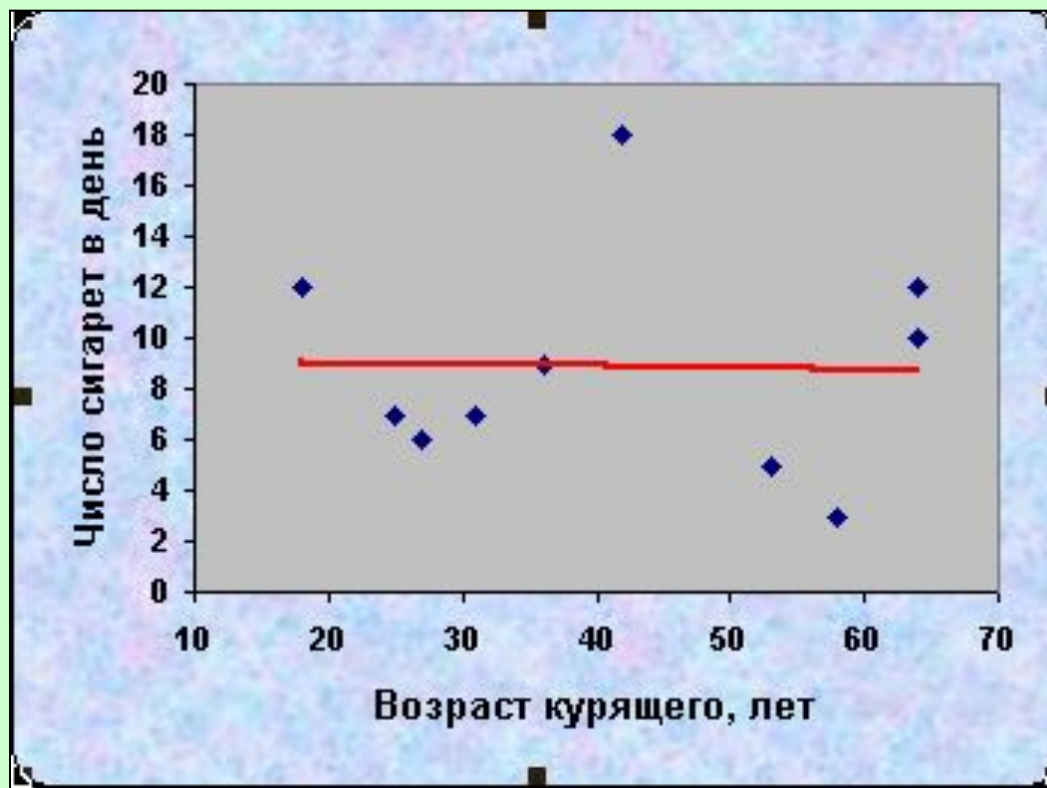


$$\hat{y} = 102,49 - 3,62x$$

# Регрессионный анализ

## Парная линейная регрессия

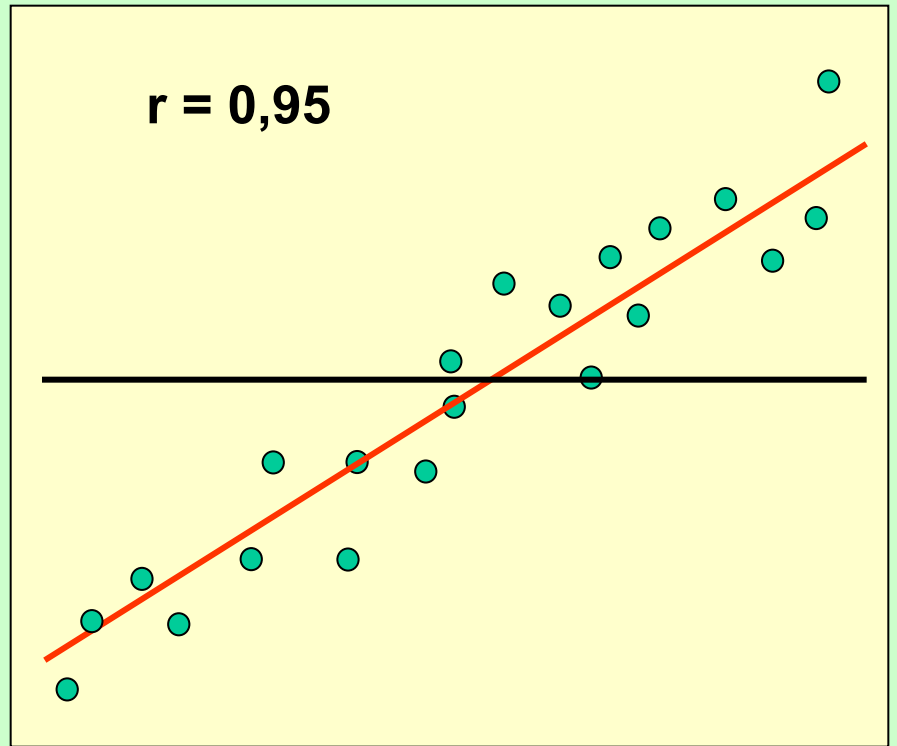
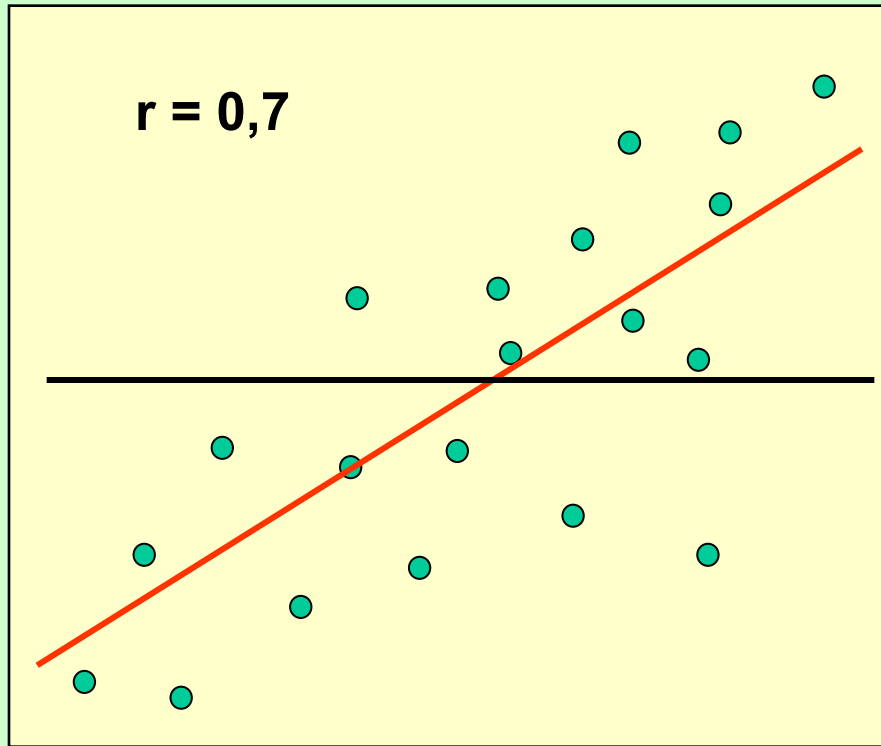
**Пример 8.** Построить уравнение линейной регрессии для данных, приведенных в примере 4.



$$\hat{y} = 9,27 - 0,009x$$

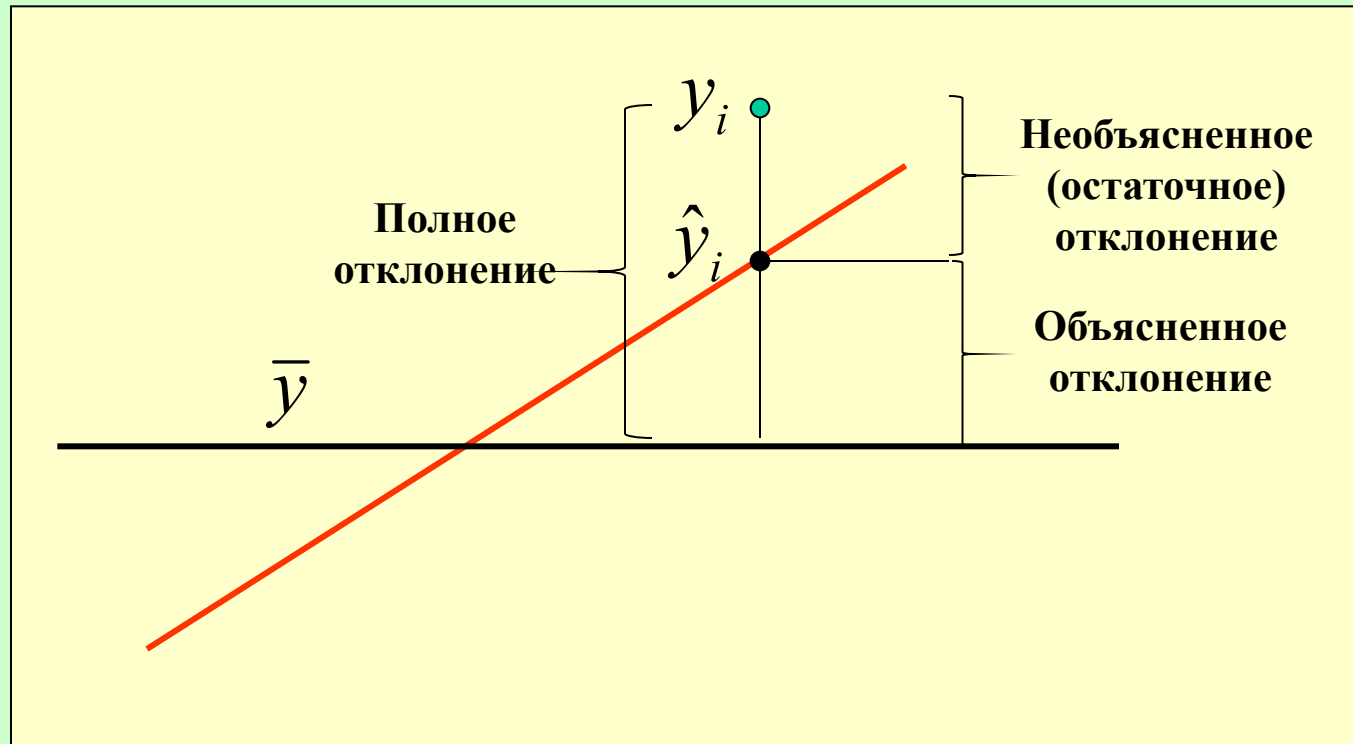
# Регрессионный анализ

## Анализ точности модели.



# Регрессионный анализ

## Анализ точности модели.



Для  $i$ -ой точки:

$$(y_i - \bar{y}) = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$$

# Регрессионный анализ

## Анализ точности модели.

$$(y_i - \bar{y}) = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$$



$$\frac{1}{n} \sum (y_i - \bar{y})^2 = \frac{1}{n} \sum (\hat{y}_i - \bar{y})^2 + \frac{1}{n} \sum (y_i - \hat{y}_i)^2$$



$$\sigma_y^2 = \sigma_{\hat{y}}^2 + \sigma_{ост}^2$$

# Регрессионный анализ

## Анализ точности модели.

Коэффициент детерминации:

$$r^2 = \frac{S_{\hat{y}}^2}{S_y^2}$$

Коэффициент детерминации является основной характеристикой регрессионной модели и показывает, какую долю вариации (изменчивости) результативного признака можно объяснить изменением факторного признака.

Одним из практических применений коэффициента детерминации является оценка качества и сравнение между собой различных моделей (линейной и нелинейных) парной регрессии.



# Регрессионный анализ

## Стандартные ошибки.

Помимо коэффициента детерминации, качество регрессионной модели характеризуют стандартные ошибки коэффициентов:

$$S_{ост}(a_0) = \frac{S_{ост}}{\sqrt{n-2}}$$

$$S_{ост}(a_1) = \frac{S_{ост}}{\sqrt{n-2} \cdot S_x}$$

и стандартная ошибка модели:

$$S_{ост}(\hat{y}) = \sqrt{\left(1 + \frac{2}{n-2}\right)} \cdot S_{ост}$$

где:

$$S_x^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

дисперсия независимой  
величины  $x$

# Регрессионный анализ

## Схема проверки гипотез о значимости коэффициентов.

**Пример 9.** На основании данных наблюдений в США за 25 – летний период (1959 – 1983 годы) построена зависимость суммарных расходов на питание ( $y$ ) от располагаемых доходов ( $x$ ):

$$\hat{y} = 55,3 + 0,093x$$

$$S_{ост}(a_0) = 2,4; \quad S_{ост}(a_1) = 0,003$$

При уровне значимости 5% проверить гипотезы о значимости коэффициентов.

# Регрессионный анализ

## Схема проверки гипотез о значимости коэффициентов.

1) Гипотезы для обоих коэффициентов формулируются одинаково:

$$H_0: a_0=0; H_1: a_0 \neq 0.$$

$$H_0: a_1=0; H_1: a_1 \neq 0.$$

2) Находим критическое значение критерия Стьюдента:

$$\text{при } p=0,05 \text{ и } k=25-2=23, t_{\text{кр}}=2,069$$

3) Находим расчетные значения критерия Стьюдента:

$$t_p(a_0) = a_0/S_{\text{осм}}(a_0) = 55,3/2,4 = 23,04$$

$$t_p(a_1) = a_1/S_{\text{осм}}(a_1) = 0,093/0,003 = 31$$

4) Принятие решения. Основные гипотезы отклоняются, коэффициенты значимы.

$$a_0 - S_{\text{осм}}(a_0) * t_{\text{кр}} < a_0 < a_0 + S_{\text{осм}}(a_0) * t_{\text{кр}}$$

$$a_1 - S_{\text{осм}}(a_1) * t_{\text{кр}} < a_1 < a_1 + S_{\text{осм}}(a_1) * t_{\text{кр}}$$

# Регрессионный анализ

## Проверка гипотезы о значимости модели.

Для решения вопроса действительно ли полученное при оценке регрессии значение  $r^2$  отражает истинную зависимость или оно получено случайно, применяется процедура проверки гипотез, основанная на анализе *F-критерия* (критерия Фишера):

$$F_p = \frac{S_y^2}{S_{ост}^2} \text{ или } F_p = \frac{S_{ост}^2}{S_y^2} \Rightarrow \frac{S_{большая}^2}{S_{меньшая}^2}$$

$$S_{ост}^2 = \frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{n - m}; \quad S_y^2 = \frac{\sum_{i=1}^n (y - y_i)^2}{n - 1}$$

где  $m$  – число параметров уравнения регрессии ( включая свободный ):

# Регрессионный анализ

## Проверка гипотезы о значимости модели.

### Способы нахождения критерия Фишера.

1) С помощью таблиц распределения ( $k_1$  – число степеней свободы числителя,  $k_2$  – число степеней свободы знаменателя):

Уровень значимости $p=0,05$							
$k_2$	$k_1$						
	1	2	...	6	...	24	...
1	161	200	...	234	...	249	...
2	18,51	19,00	...	19,33	...	19,45	...
...	...	...	...	...	...	...	...
23	4,28	3,42	...	2,53	...	2,00	...
...	...	...	...	...	...	...	...

# Регрессионный анализ

## Проверка гипотезы о значимости модели

2) С помощью стандартной функции Excel *FRASПОБР*.

**FRASПОБР**( $p; k_1; k_2$ )

	A	B	C
1	Число степеней свободы числителя	24	
2	Число степеней свободы знаменателя	23	
3	Уровень значимости ( $p$ )	0,050	
4	Критическое значение F	2,005	=FRASПОБР(B3;B1;B2)

# Регрессионный анализ

## Нелинейная парная регрессия

**Пример 10.** В таблице приведены данные количества покупаемых бананов в месяц (кг) от годового дохода (в тыс. условных единиц) для десяти семей.

Годовой доход, $x_i$	1	2	3	4	5	6	7	8	9	10
Количество бананов, $y_i$	1,93	7,13	8,78	9,69	10,09	10,42	10,62	10,71	10,79	11,13

Построить уравнения линейной и нелинейной регрессии и оценить качество полученных моделей.

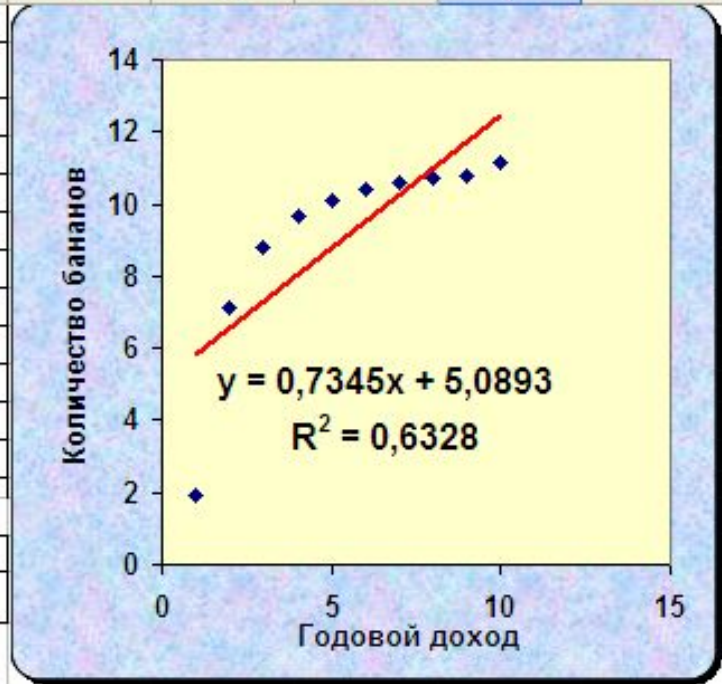
# Регрессионный анализ

## Нелинейная парная регрессия

1. Уравнение линейной регрессии:

$$\hat{y} = 5,0893 + 0,7345x$$

	A	B	C	D	E	F	G	H	I	J
1	Годовой доход	Количество	Линейная регрессия							
2	$x_i$ (в 1000 у.е.)	бананов $y_i$ (в кг)	Урасч	$(y_i - U_{расч})^2$	$(U_{сред} - y_i)^2$					
3	1	1,93	5,82	15,16	51,83					
4	2	7,13	6,56	0,33	4,00					
5	3	8,78	7,29	2,21	0,12					
6	4	9,69	8,03	2,76	0,31					
7	5	10,09	8,76	1,76	0,92					
8	6	10,42	9,50	0,85	1,67					
9	7	10,62	10,23	0,15	2,22					
10	8	10,71	10,97	0,07	2,50					
11	9	10,79	11,70	0,83	2,76					
12	10	11,13	12,43	1,70	4,00					
13										
14	Среднее значение $y$		9,13							
15	Дисперсия воспроизводимости		7,81	$n-1=$	9					
16	Остаточная дисперсия		3,2285	$n-m=$	8					
17	Расчетное значение критерия Фишера		2,4206	$=C15/C16$						
18	Критическое значение критерия Фишера		3,3881	$=FРАСПОБР(0,05;E15;E16)$						
19										



$F_p < F_{кр}$  - модель неадекватна



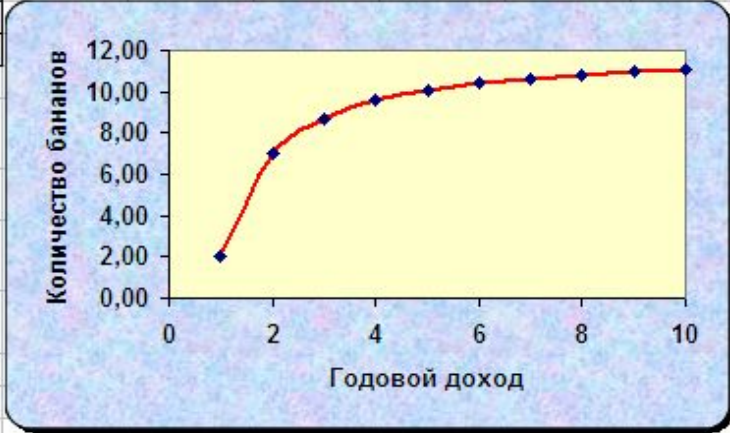
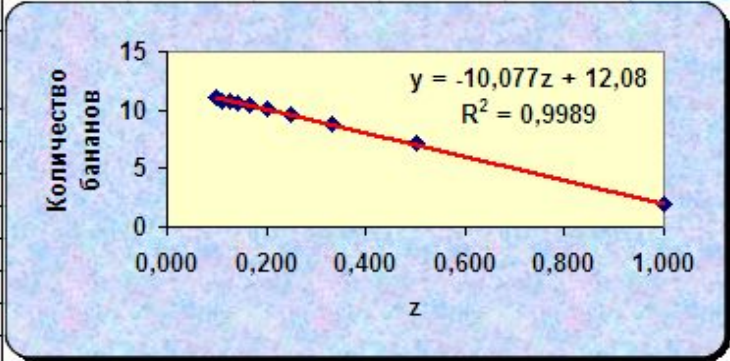
# Регрессионный анализ

## Нелинейная парная регрессия

2. Уравнение нелинейной регрессии:

$$\hat{y} = 12,08 - 10,077 \cdot \frac{1}{x}$$

	A	B	C	D	E	F
1	Годовой доход $x_i$ (в 1000 у.е.)	Количество бананов $y_i$ (в кг)	Нелинейная регрессия			
2			$z=1/x$	Урасч	$(y_i - y_{i\text{расч}})^2$	$(y_{\text{сред}} - y_i)^2$
3	1	1,93	1,000	2,00	0,0053	51,8256
4	2	7,13	0,500	7,04	0,0078	3,996001
5	3	8,78	0,333	8,72	0,0035	0,121801
6	4	9,69	0,250	9,56	0,0167	0,314721
7	5	10,09	0,200	10,06	0,0006	0,923521
8	6	10,42	0,167	10,40	0,0004	1,666681
9	7	10,62	0,143	10,64	0,0004	2,223081
10	8	10,71	0,125	10,82	0,0122	2,499561
11	9	10,79	0,111	10,96	0,0290	2,758921
12	10	11,13	0,100	11,07	0,0033	4,004001
13	Среднее значение $y$		9,13			
14	Дисперсия воспроизводимости		7,81	$n-1=$	9	
15	Остаточная дисперсия		0,0099	$n-m=$	8	
16	Расчетное значение критерия Фишера		788,2278	$=C14/C15$		
17	Критическое значение критерия Фишера		3,3881	$=FPASPOBR(0,05;E14;E15)$		
18						
19						
20						



$F_p > F_{кр}$  - модель адекватна

# Регрессионный анализ

## Нелинейная парная регрессия

Нелинейные модели парной регрессии и преобразование переменных.

Тип модели	Связь	Преобразования	Линейное уравнение
Экспоненциальная	$y = \exp(a_0 + a_1 x)$	$\ln(y) = u$	$u = a_0 + a_1 x$
Обратная по y	$y = 1/(a_0 + a_1 x)$	$1/y = u$	$u = a_0 + a_1 x$
Обратная по x	$y = a_0 + a_1/x$	$1/x = z$	$y = a_0 + a_1 z$
Дважды обратная	$y = 1/(a_0 + a_1/x)$	$1/x = z; 1/y = u$	$u = a_0 + a_1 z$
Логарифм по x	$y = a_0 + a_1 \ln(x)$	$\ln(x) = z$	$y = a_0 + a_1 z$
Мультипликативная	$y = a_0 x^{a_1}$	$\ln(x) = z; \ln(y) = u; \ln(a_0) = b$	$u = b + a_1 x$
Квадратный корень по x	$y = a_0 + a_1 x^{1/2}$	$x^{1/2} = z$	$y = a_0 + a_1 z$
Квадратный корень по y	$y = (a_0 + a_1 x)^{1/2}$	$y^2 = u$	$u = a_0 + a_1 x$
S-кривая	$y = \exp(a_0 + a_1/x)$	$\ln(y) = u; 1/x = z$	$u = a_0 + a_1 z$