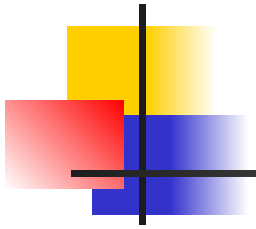
Корреляционный и регрессионный анализы

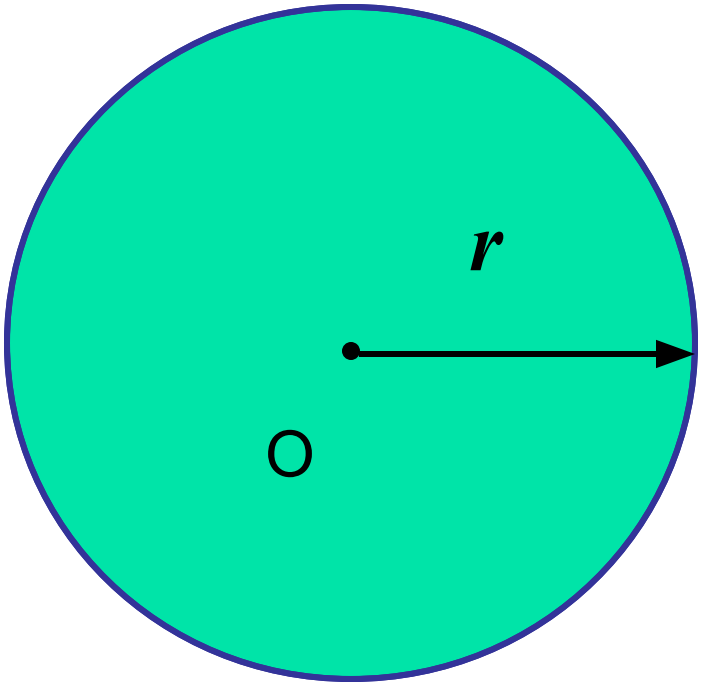


1. Основные задачи теории корреляции.
2. Корреляционный анализ.
3. Регрессионный анализ.

Функциональная зависимость:

каждому возможному значению переменной x ставится в соответствие единственное значение переменной y.

ФУНКЦИОНАЛЬНАЯ СВЯЗЬ



$$S = \pi \cdot r^2$$

ФУНКЦИОНАЛЬНАЯ СВЯЗЬ

$$S = v \cdot t$$

ФУНКЦИОНАЛЬНАЯ СВЯЗЬ

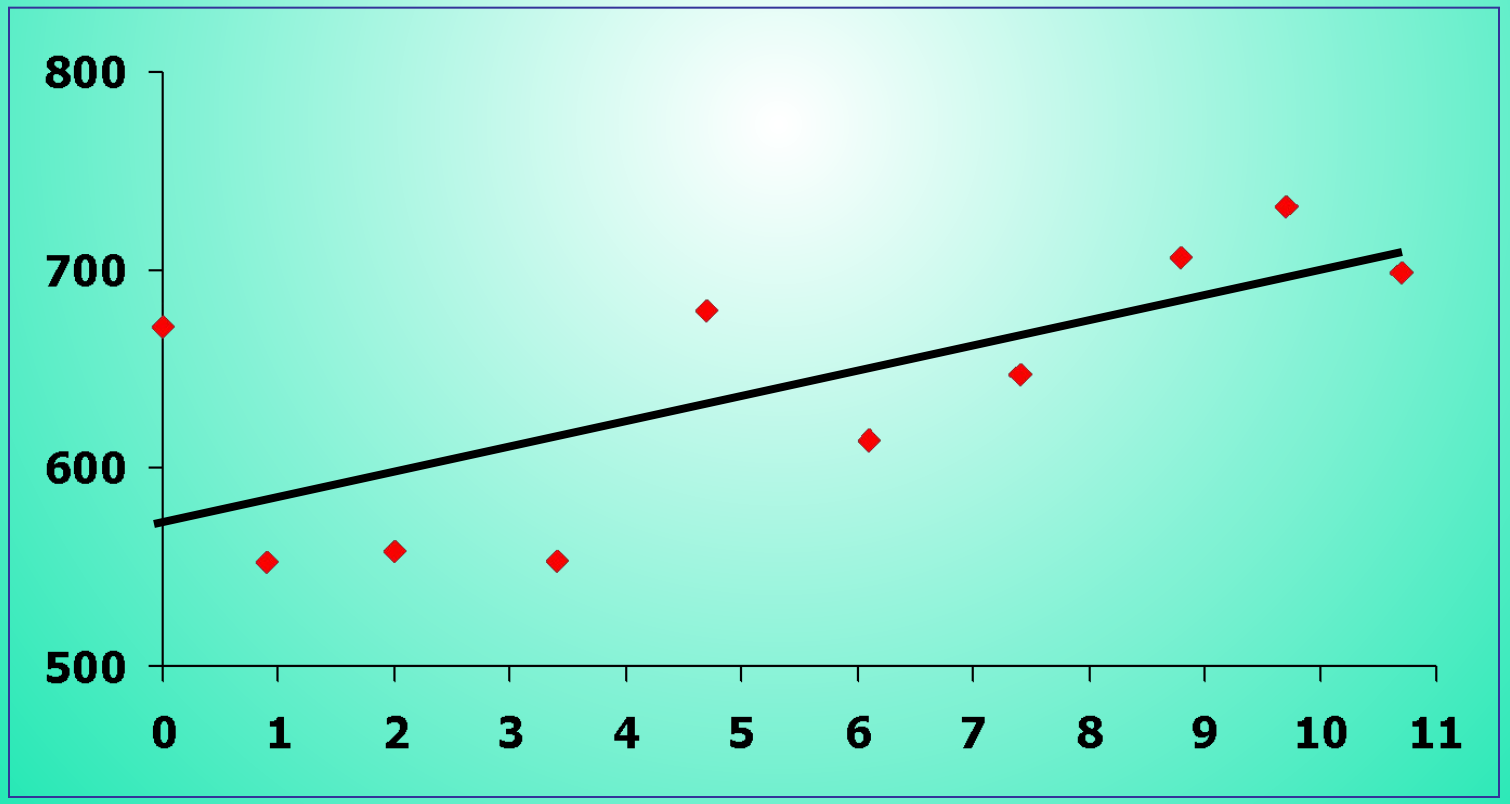
Стохастической зависимостью

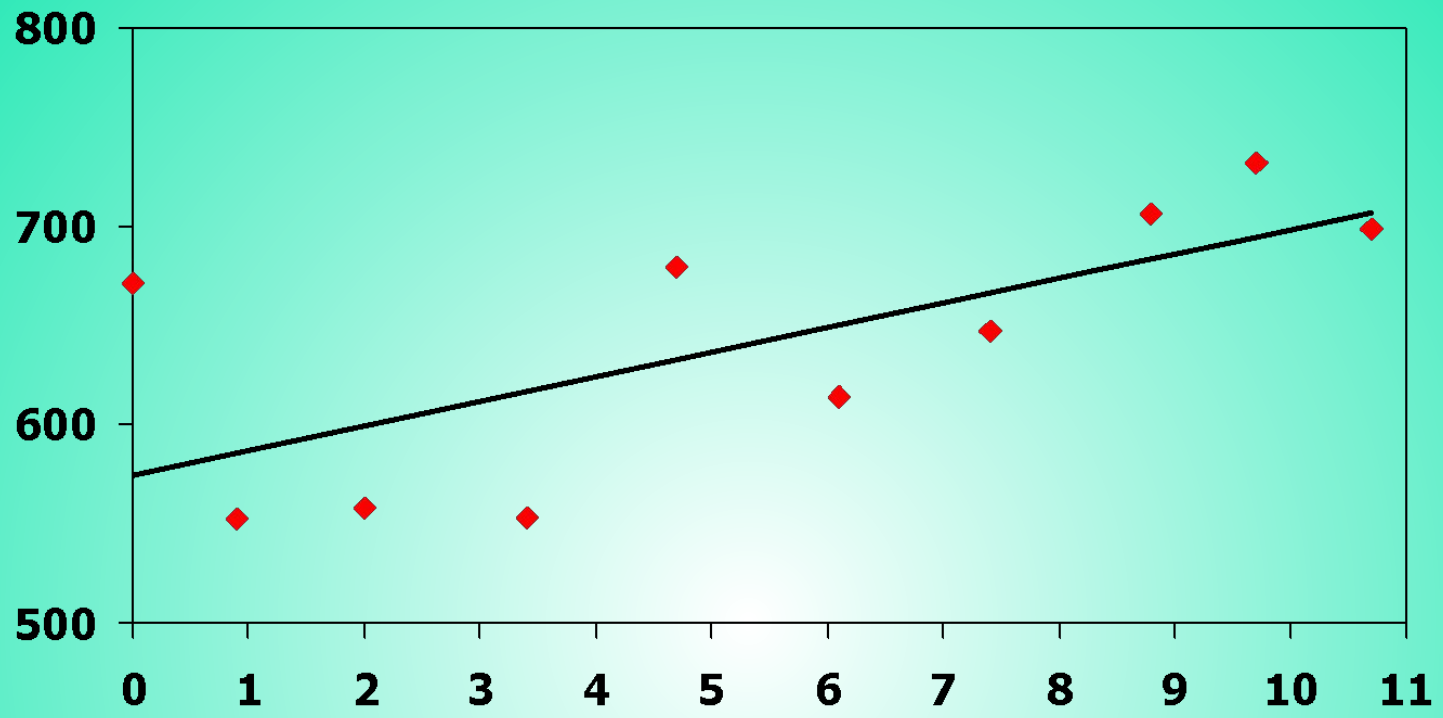
называют зависимость, при которой изменение одной из величин влечет изменение распределения другой.

Корреляционной зависимостью

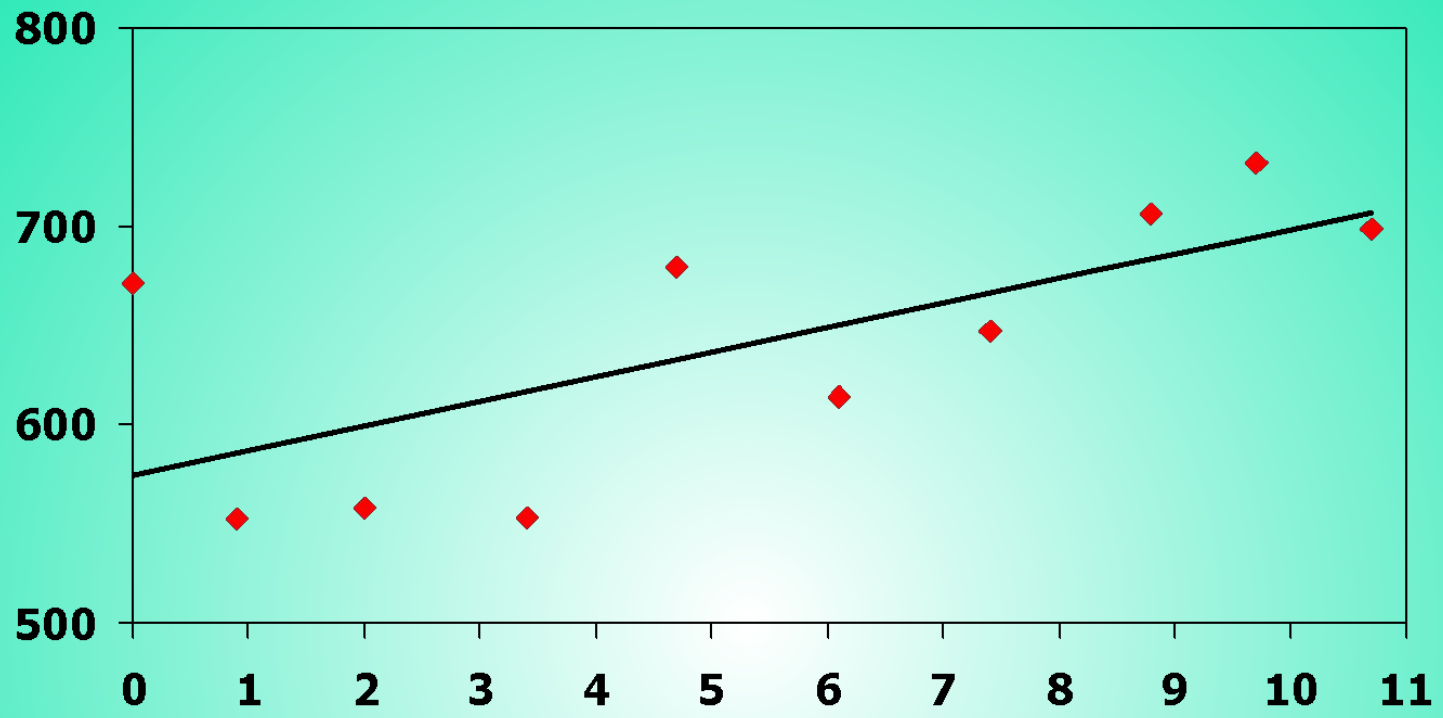
называют зависимость, при которой изменение одной из величин влечет изменение среднего значения другой.

Годы	1-й	2-й	3-й	4-й	5-й	6-й	7-й	8-й	9-й	10-й
Темп прироста населения	0	0,9	2	3,4	4,7	6,1	7,4	8,8	9,7	10,7
Число зарегист-рированных преступлений	671	552	558	553	679	614	647	706	732	699





Линия регрессии – это графическое представление ведущей тенденции связи между количественными признаками.



Чем ближе точки в поле диаграммы рассеяния к линии регрессии, тем сильнее воздействие независимой переменной на зависимую (тем сильнее корреляция между обеими переменными).

ТЕОРИЯ КОРРЕЛЯЦИИ

ЗАДАЧИ

Установить ФОРМУ корреляционной связи

решает

регрессионный анализ

Установить ТЕСНОТУ корреляционной связи

решает

корреляционный анализ

Корреляционный анализ

1. Коэффициент линейной корреляции Пирсона.
2. Свойства коэффициента корреляции.
3. Оценка значения коэффициента корреляции.

Простой (выборочный) коэффициент корреляции Пирсона

Номер наблюдения	X	Y
1	x_1	y_1
2	x_2	y_2
…	…	…
n	x_n	y_n

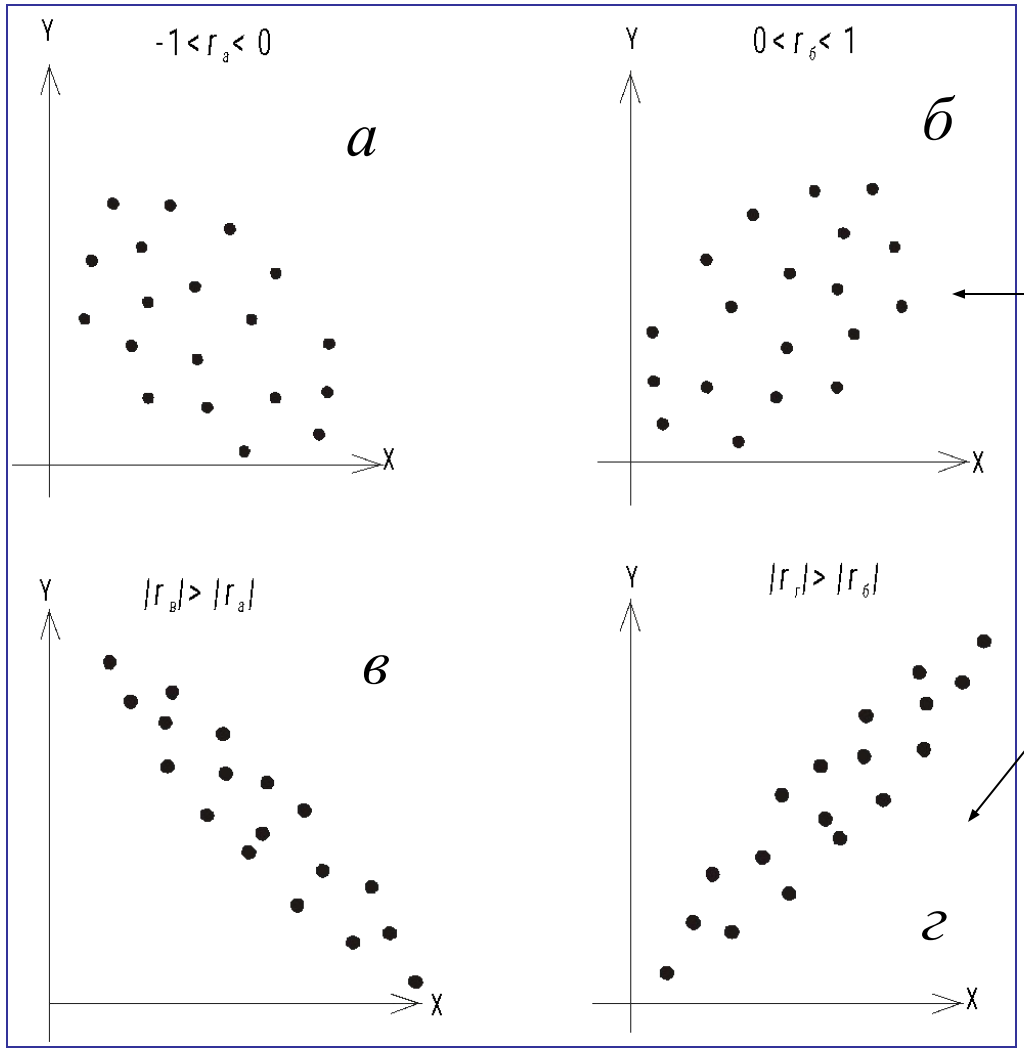
$$r_{X,Y} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{(n\sum_{i=1}^{n} (x_i)^2 - (\sum_{i=1}^{n} x_i)^2) \cdot (n\sum_{i=1}^{n} (y_i)^2 - (\sum_{i=1}^{n} y_i)^2)}};$$

Свойства коэффициента корреляции

1. Величина коэффициента корреляции заключена в пределах

$$-1 \leq r \leq 1,$$

Свойства коэффициента корреляции

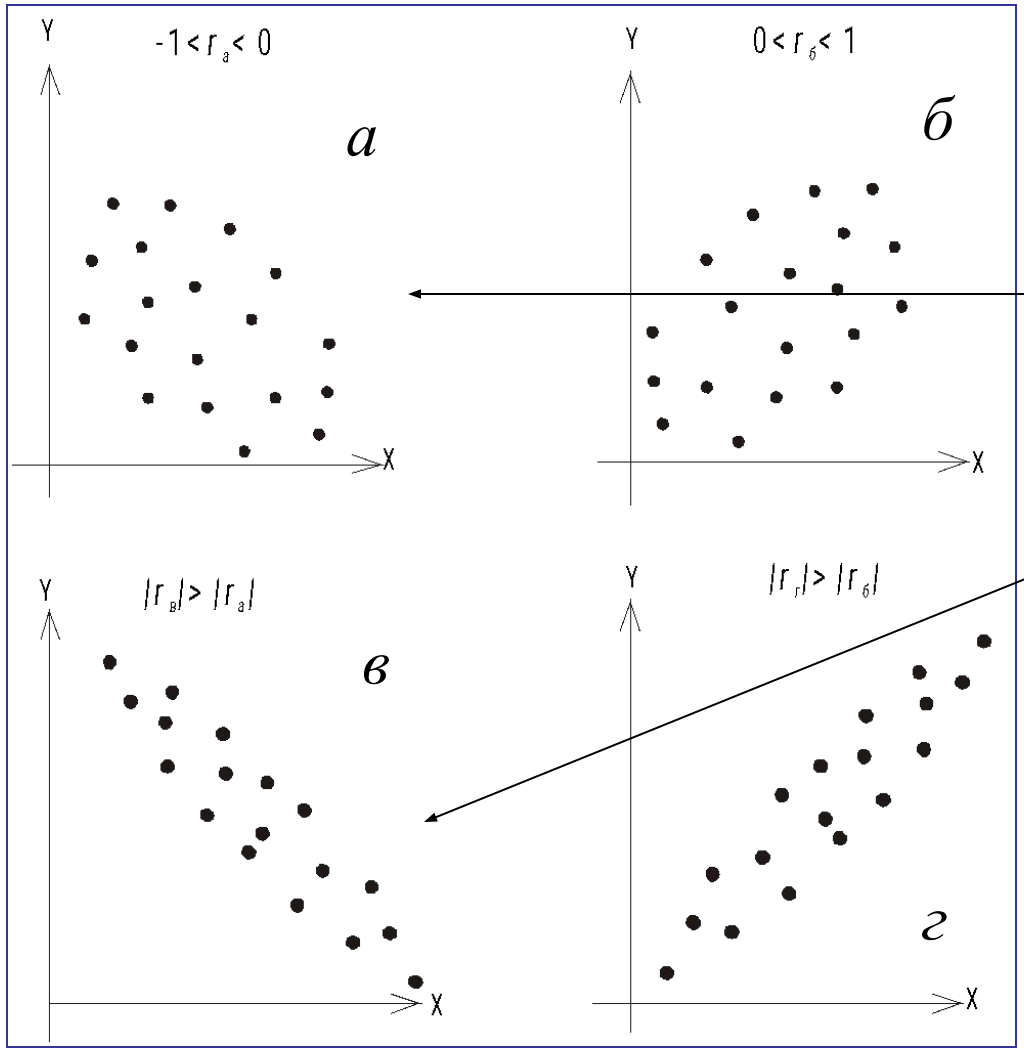


причем

$0 < r \leq 1$,

если при увеличении значений одной из величин значения другой имеют тенденцию к увеличению (*прямая связь*),

Свойства коэффициента корреляции

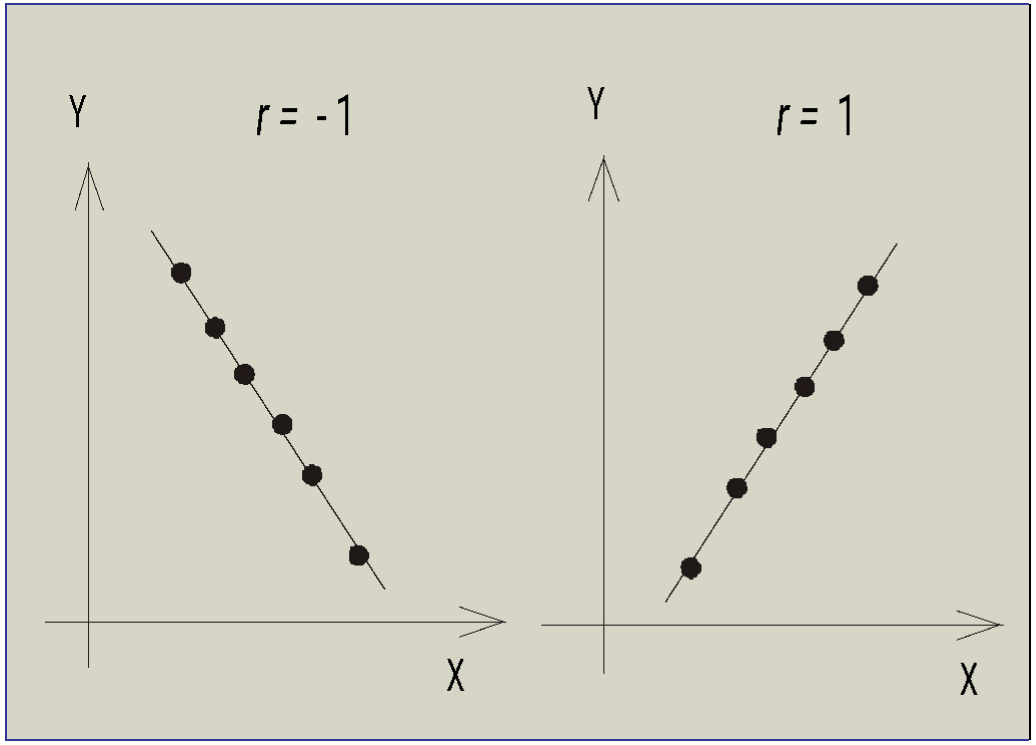


и

$$-1 \le r < 0,$$

если при увеличении значений одной из величин значения другой имеют тенденцию к уменьшению (*обратная связь*).

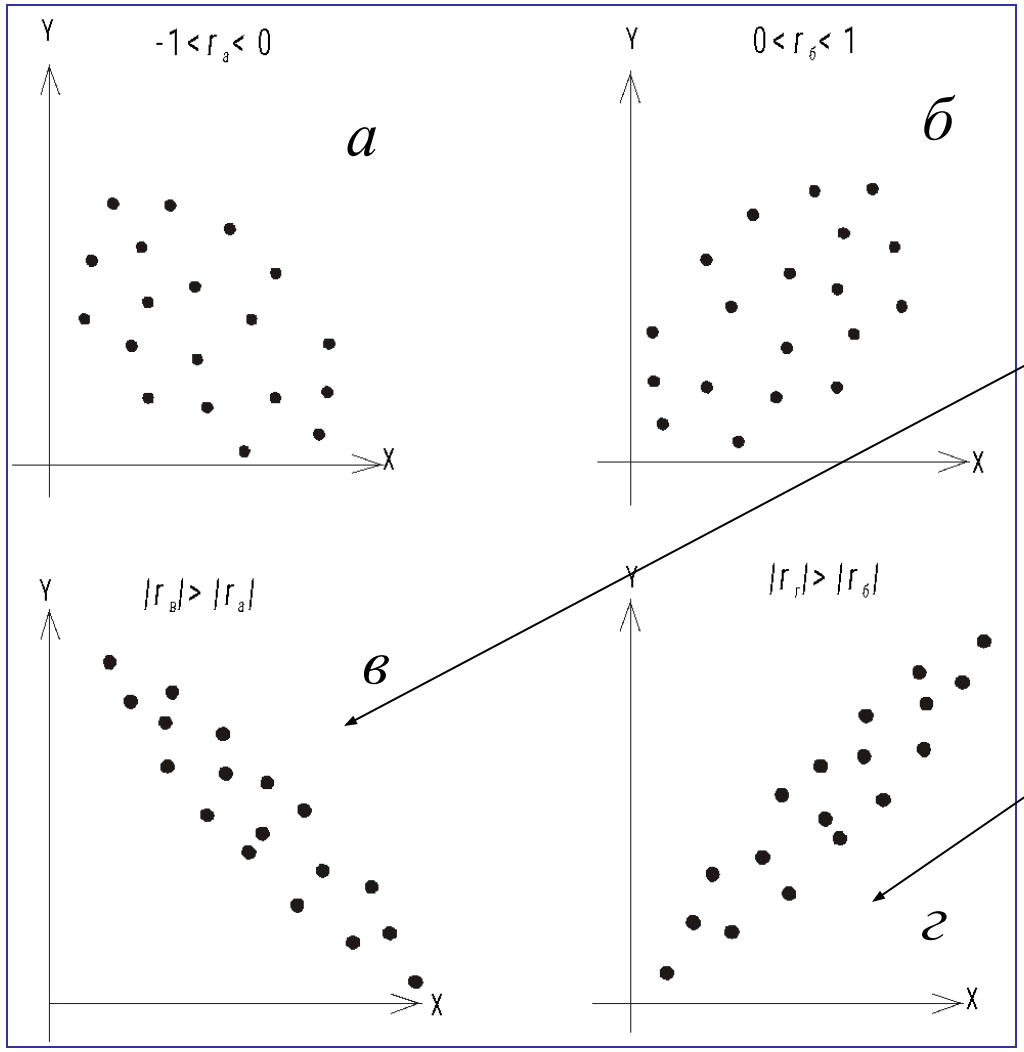
Свойства коэффициента корреляции



$$|r_{X,Y}| = 1$$

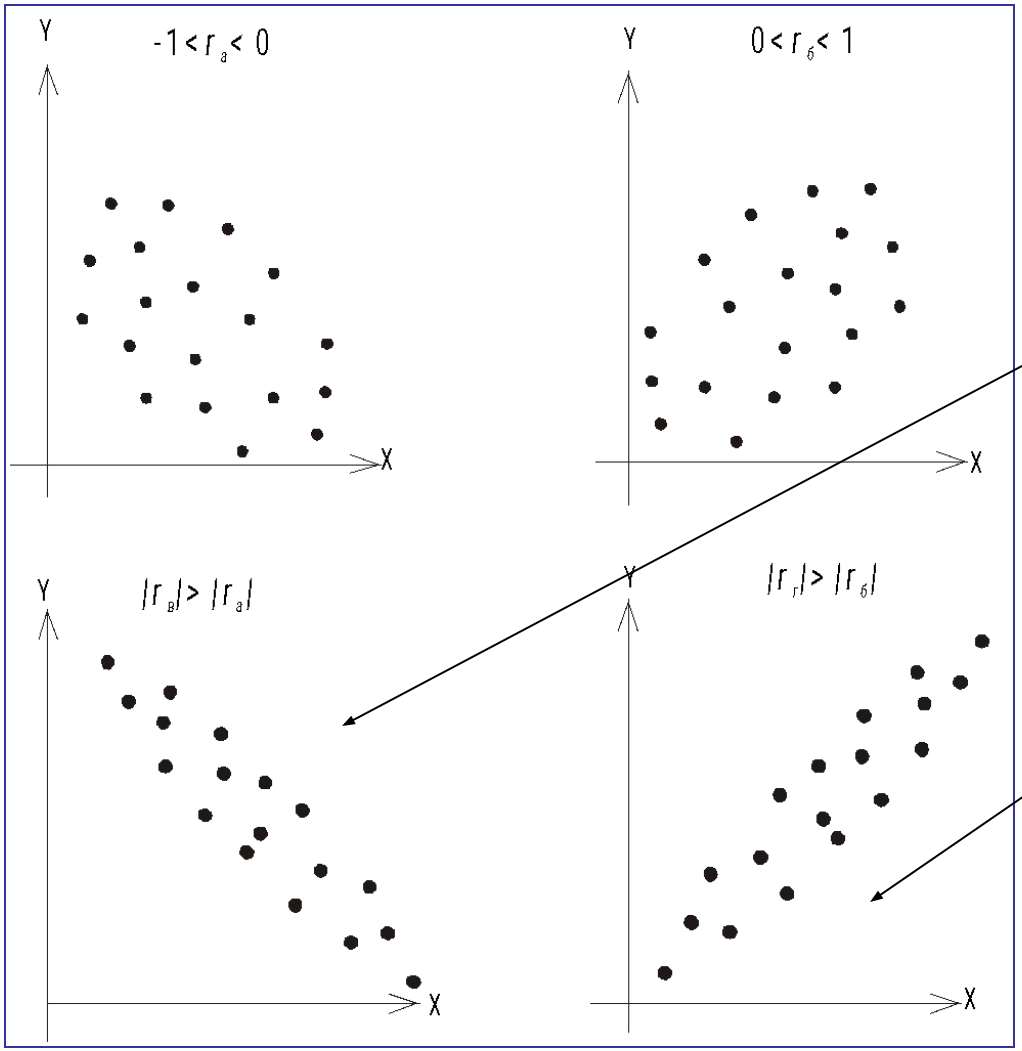
тогда и только тогда, когда случайные величины X и Y линейно связаны, т.е. точки с координатами (x_i, y_i) лежат на одной прямой.

Свойства коэффициента корреляции



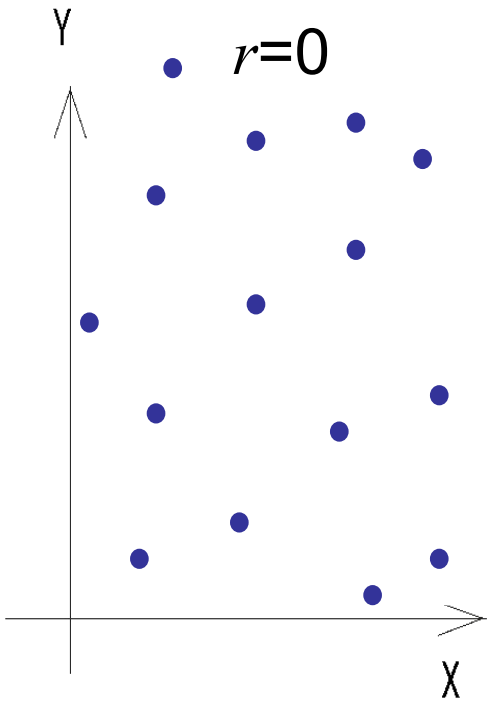
Чем ближе $|r_{X,Y}|$ к единице, тем сильнее линейная связь между случайными величинами, т.е. тем меньше точки с координатами (x_i, y_i) рассеяны около прямой.

Свойства коэффициента корреляции



Чем меньше точки с координатами (x_i, y_i) рассеяны около некоторой прямой, тем ближе $|r_{X,Y}|$ к единице.

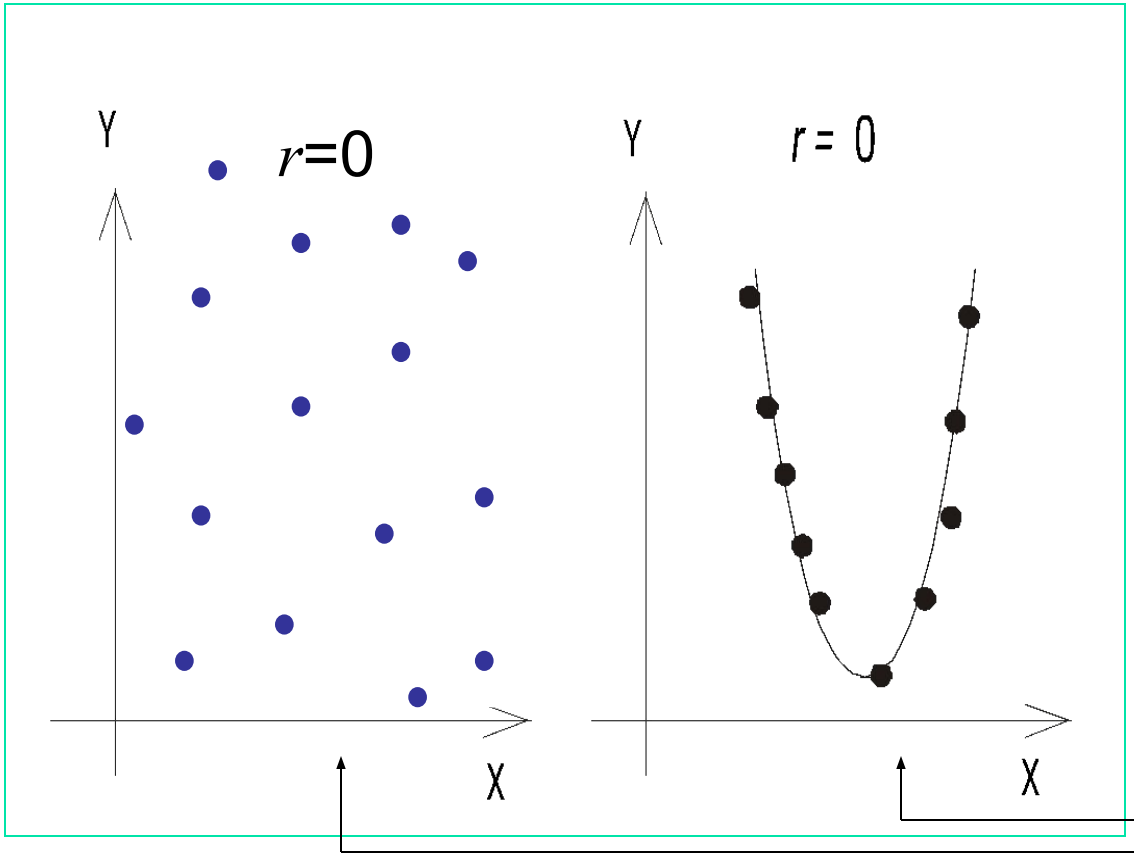
Свойства коэффициента корреляции



Если X и Y статистически независимы, то

$$|r_{X,Y}| = 0$$

Свойства коэффициента корреляции



Если

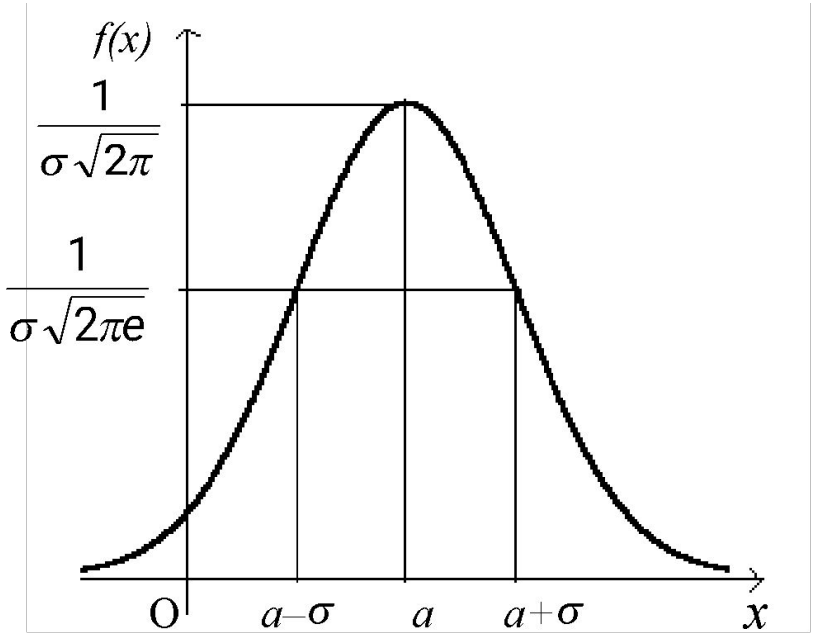
$$|r_{X,Y}| = 0$$

то связь между случайными величинами либо отсутствует,

либо не носит линейного характера.

Свойства коэффициента корреляции

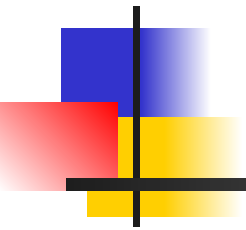
Для нормально распределенных X и Y из того, что



$$\left| r_{X,Y} \right| = 0$$

следует их независимость.

Оценка значения коэффициента корреляции



1) оценка тесноты статистической линейной связи по абсолютному значению *r*:

$|r| \approx 0$ – связь отсутствует;

$|r| \leq 0{,}3$ – связь слабая;

$0{,}3 < |r| \leq 0{,}5$ – связь умеренная;

$0{,}5 < |r| \leq 0{,}7$ – связь значительная;

$0{,}7 < |r| \leq 0{,}9$ – связь сильная;

$0{,}9 < |r|$ – очень сильная;

$|r| = 1$ – функциональная связь.

2) оценка направления статистической линейной связи по знаку r:

знак «+» – прямая связь,

знак «–» – обратная связь.

3) оценка значимости полученного результата:

Уровень значимости α, говорит о том, с какой надежностью γ=(1-α)×100% можно доверять полученному результату.

Если α близок к нулю, можно доверять вычисленному значению коэффициента корреляции;

когда α>0,2, к значению коэффициента корреляции следует относиться с большой осторожностью.

Расчетная таблица

№	x_i	y_i
1	0	671
2	0,9	552
3	2	558
4	3,4	553
5	4,7	679
6	6,1	614
7	7,4	647
8	8,8	706
9	9,7	732
10	10,7	699
Итого	53,7	6411

$$r_{X,Y} = \frac{10 \cdot 3601000 - 53,7 \cdot 6411}{\sqrt{(10 \cdot 416,45 - 53,7^2)(10 \cdot 4151625 - 6411^2)}} = 0,69$$

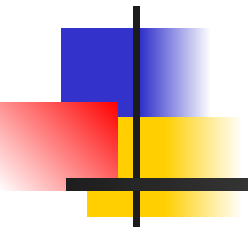
$$r_{X,Y} = \frac{10 \cdot 360100,00 - 537 \cdot 6411}{\sqrt{(10 \cdot 41645 - 537^2)(10 \cdot 41516255 - 6411^2)}} = 0,69$$

связь значительная ($|r| = 0,69$ $0,5 < |r| \le 0,7$),

прямая (знак «+»).

Регрессионный анализ

1. Классификация.
2. Основные задачи.
3. Анализ адекватности модели.



I. Классификация

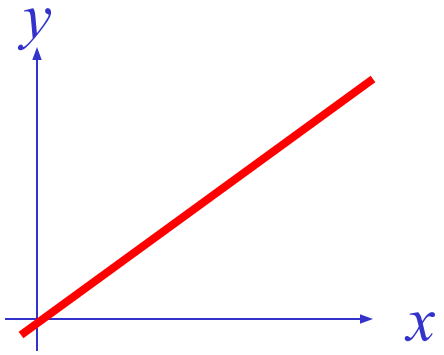
1.В зависимости от числа явлений

– простой (регрессия между двумя переменными);

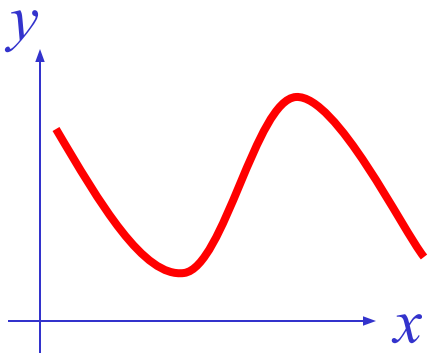
– множественной (регрессия между зависимой переменной Y и несколькими независимыми переменными (X_1, X_2, …, X_n)).

2.В зависимости от формы

– линейной (отображается линейной функцией, а между изучаемыми явлениями существуют линейные отношения);

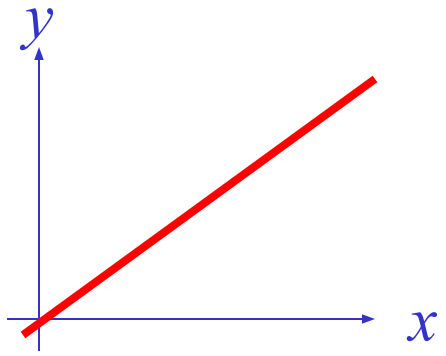


– нелинейной (отображается нелинейной функцией, между изучаемыми переменными связь носит нелинейный характер).

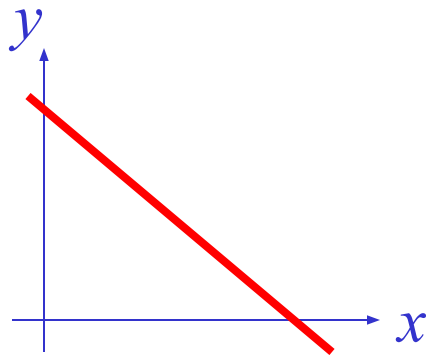


3. По характеру связи между включенными в рассмотрение переменными

– положительной (увеличение значения независимой переменной приводит к увеличению значения зависимой переменной и наоборот);



– отрицательной (с увеличением значения независимой переменной значение зависимой переменной уменьшается).

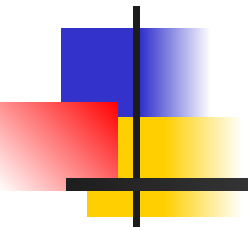


4. По типу

– непосредственной (в этом случае причина оказывает прямое воздействие на следствие, т.е. зависимая и независимая переменные связаны непосредственно друг с другом);

– косвенной (независимая переменная оказывает опосредованное действие через третью или ряд других переменных на зависимую переменную);

– ложной (нонсенс регрессия) – может возникнуть при поверхностном и формальном подходе к исследуемым процессам и явлениям.

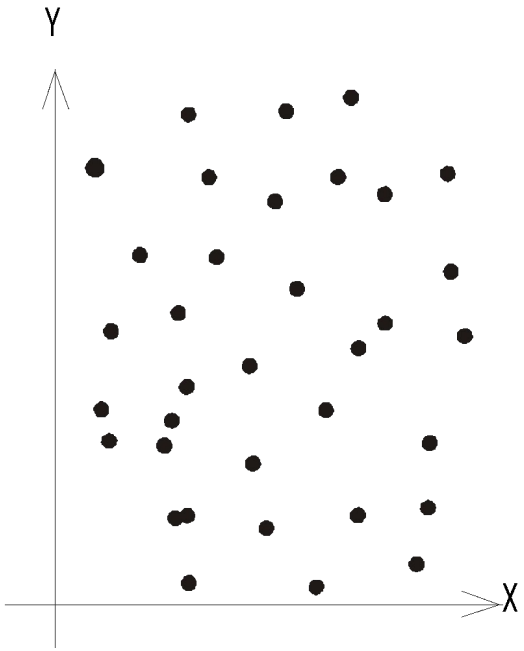


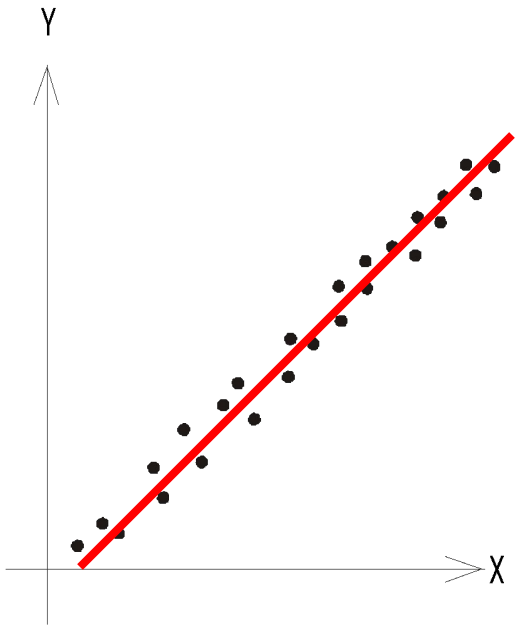
II. Основные задачи

Основные задачи

1. Определение формы зависимости.
2. Отыскание подходящих значений неизвестных параметров.
3. Оценка неизвестных значений зависимой переменной.

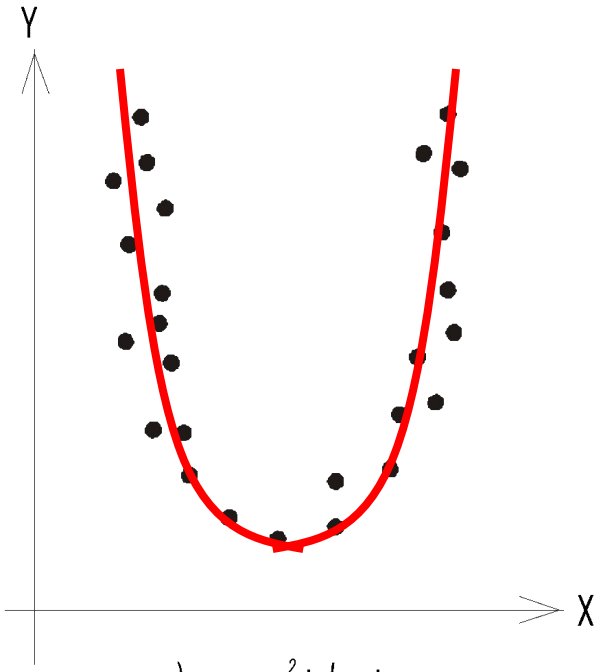
1. Определение формы зависимости



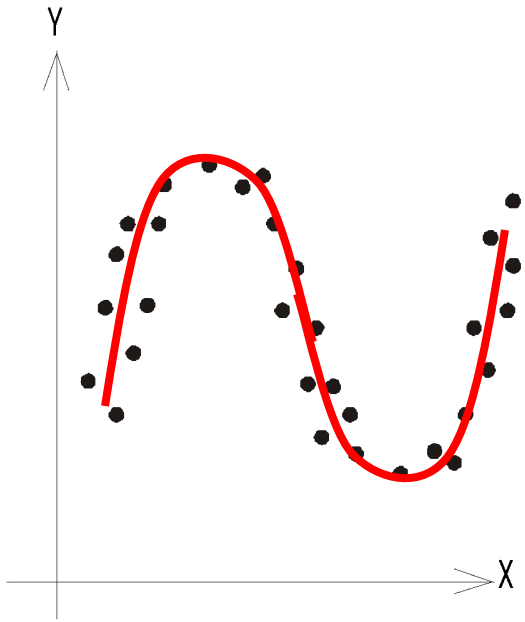


б) y = ax + b

1. Определение формы зависимости

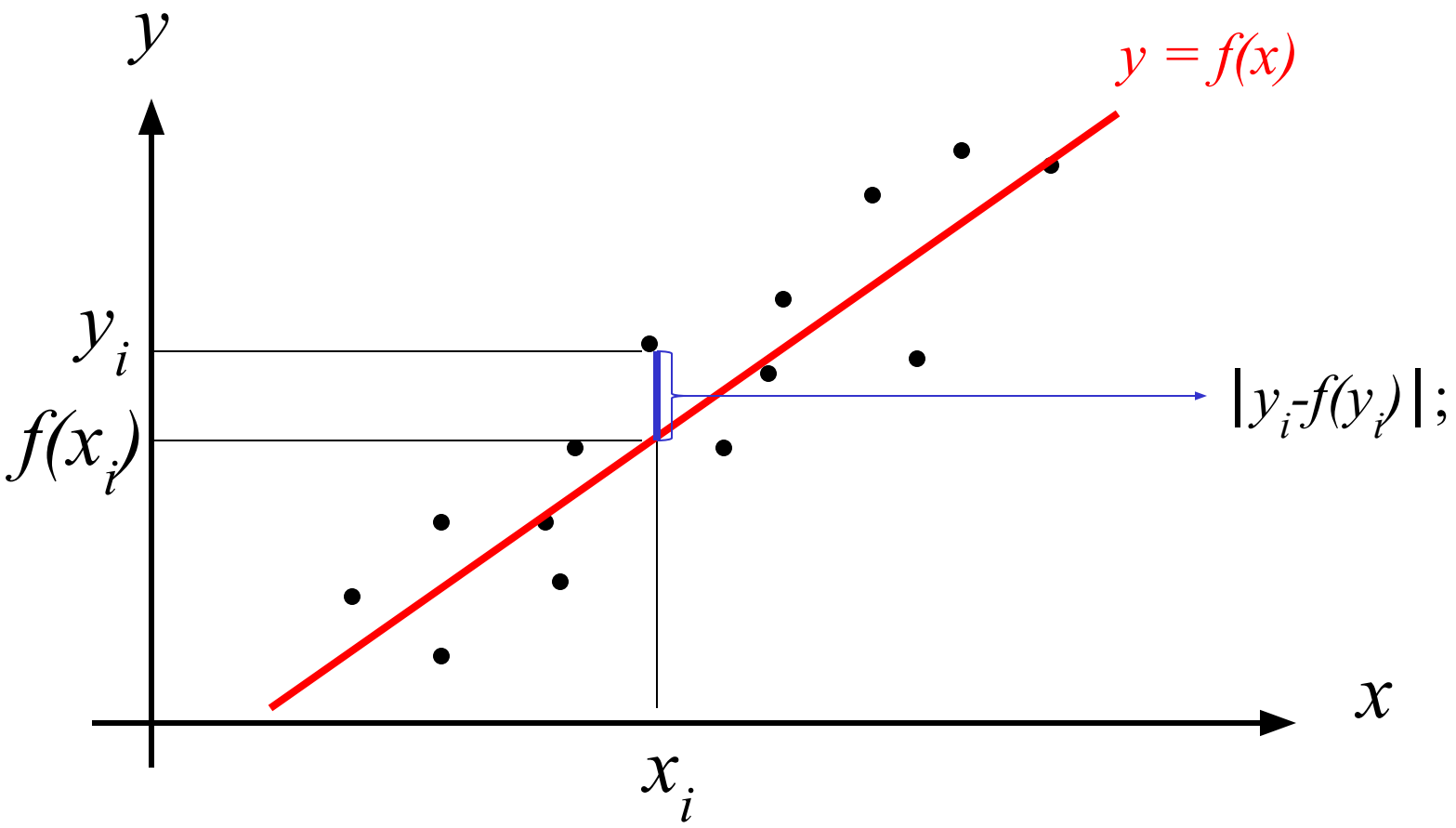


в) $y = ax^2 + bx + c$

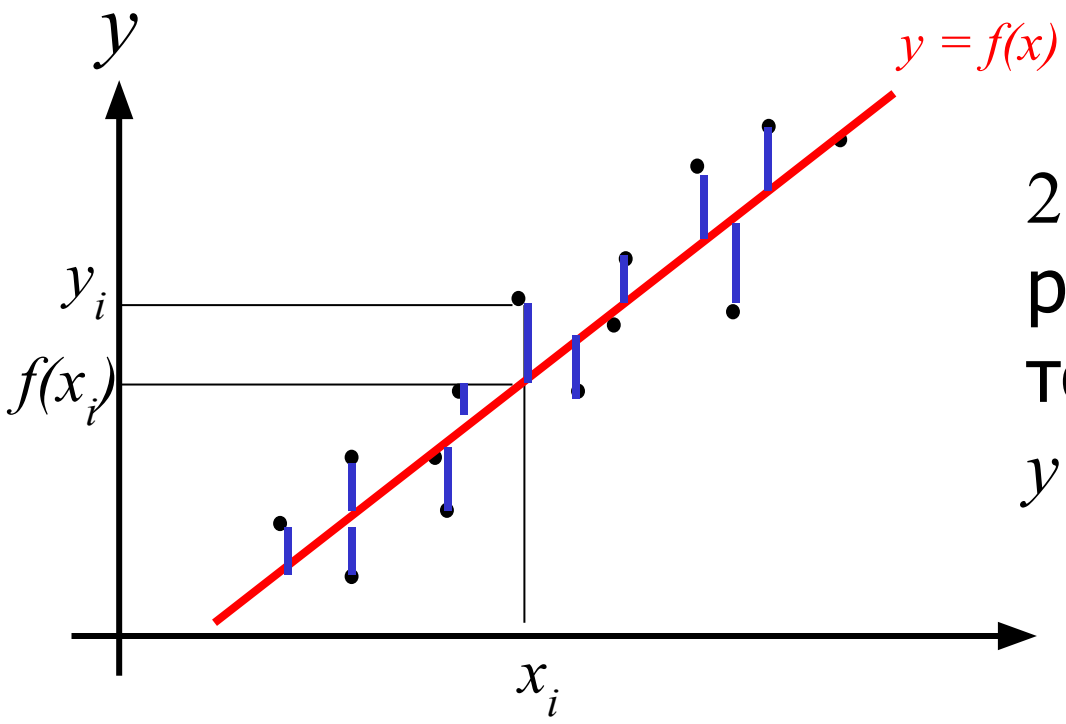


г) $y = a\sin(xb) + c$

2. Отыскание подходящих значений неизвестных параметров



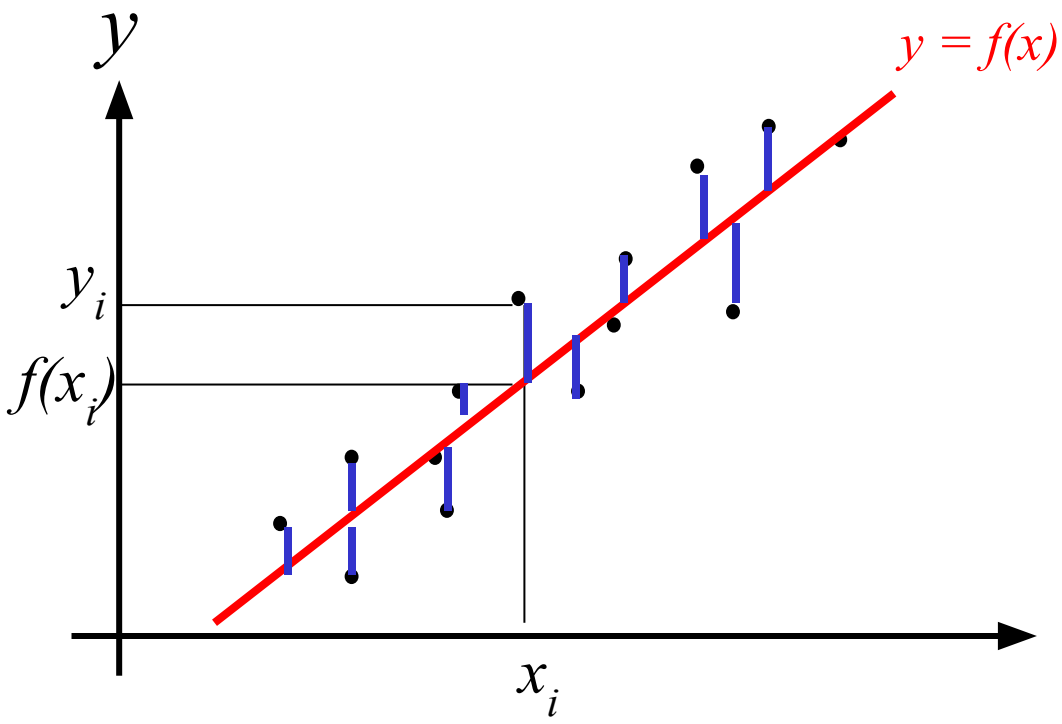
2. Отыскание подходящих значений неизвестных параметров



2.1 измеряем расстояние от каждой точки до прямой по оси y :

$$|y_i - f(y_i)|;$$

2. Отыскание подходящих значений неизвестных параметров



2.2 возводим эти расстояния в квадрат:

$$|y_i - f(x_i)|^2;$$

2. Отыскание подходящих значений неизвестных параметров

2.3 суммируем по всем точкам:

$$S=|y_1-f(x_1)|^2+|y_2-f(x_2)|^2+\ldots+|y_i-f(y_i)|^2;$$

2.4 требуем, чтобы полученная сумма квадратов расстояний была минимальной

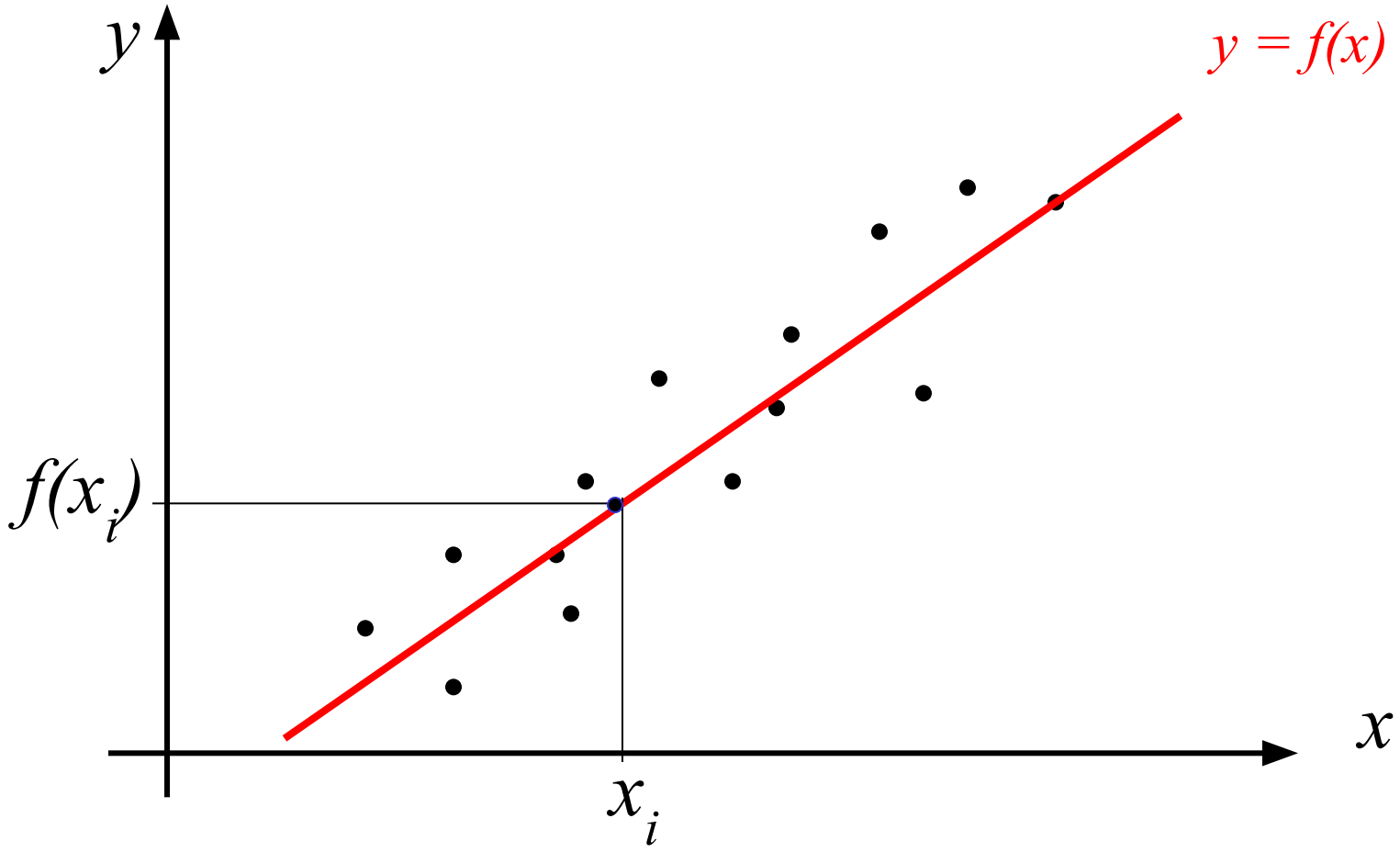
$$S \Rightarrow min$$

В случае линейной регрессии $y(x) = ax+b$

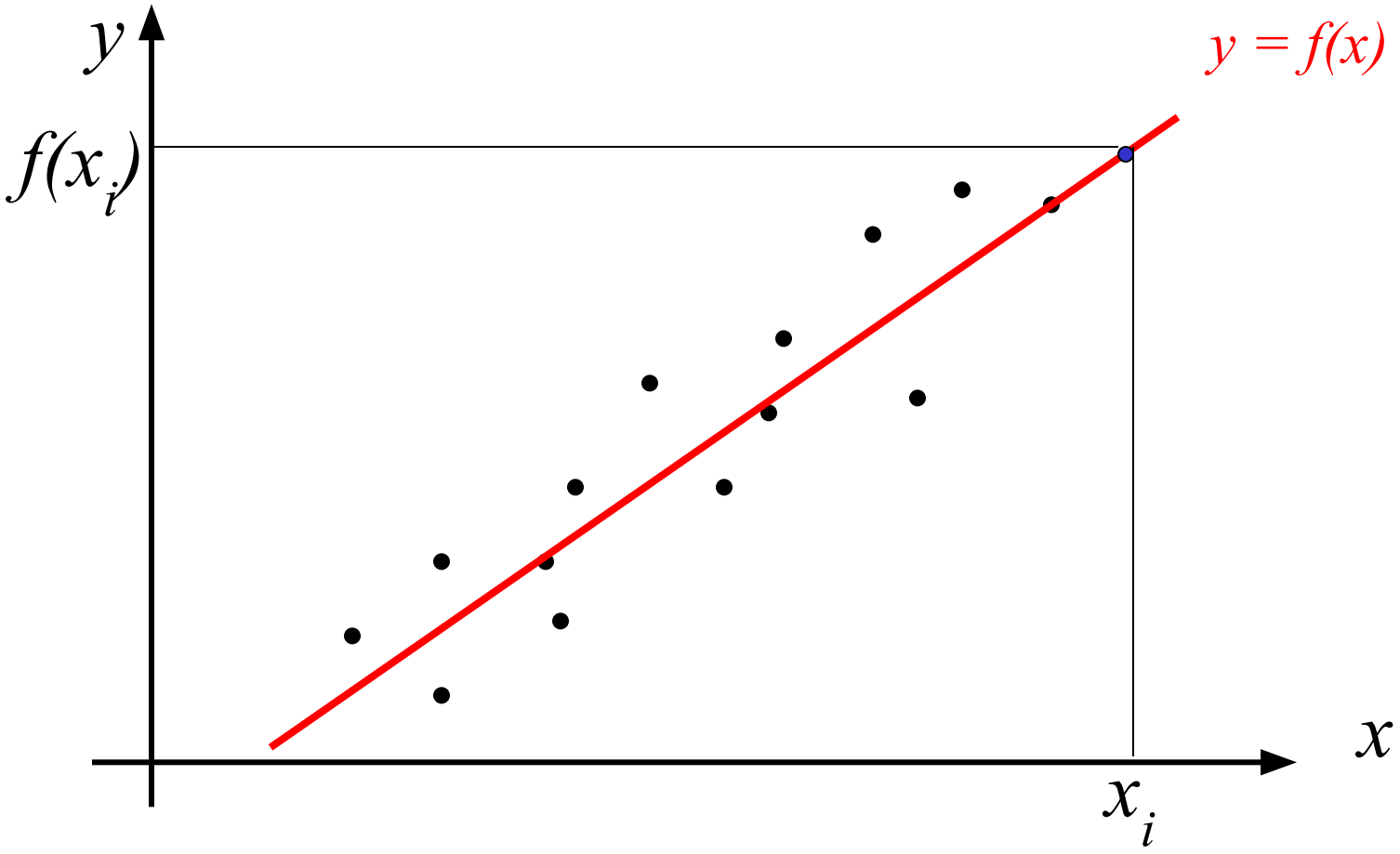
$$a = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

$$b = \frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

3. Оценка неизвестный значений зависимой переменной



3. Оценка неизвестный значений зависимой переменной



Анализ адекватности модели

- ***Предсказанные значения*** – значения, соответствующие наблюдаемым независимым значениям x_i, вычисленные согласно уравнению $y=f(x)$ (будем обозначать y_i^*).

- ***Остатки*** – разности между наблюдаемыми значениями и предсказанными: $y_i - f(x_i) = y_i - y_i^*$

$$\bar{y} = \frac{y_1 + y_2 + \ldots + y_n}{n}$$

$$SS = (y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \ldots + (y_n - \bar{y})^2$$

$$SS\,\mathrm{Pr} = (y_1^{*} - \bar{y})^2 + (y_2^{*} - \bar{y})^2 + \ldots + (y_n^{*} - \bar{y})^2$$

Коэффициент детерминации

$$RI = \frac{SS\,\mathrm{Pr}}{SS}$$

Коэффициент детерминации

Свойства:

а) $0 \leq RI \leq 1$;

б) Чем ближе коэффициент детерминации к 1, тем лучше регрессия «объясняет» зависимость данных;

в) В случае линейной регрессии $RI = r^2$

Средняя ошибка аппроксимации

$$\bar{\varepsilon} = \frac{1}{n}\sum_{i=1}^{n}\frac{\left|y_i - y_i^*\right|}{y_i}\cdot 100\%$$

Модель считается адекватной, если

$$\left|\bar{\varepsilon}\right| \le 15\%$$

Анализ остатков

Если модель подобрана правильно, то

- остатки будут вести себя достаточно хаотично,

- в остатках не будет систематической составляющей, резких выбросов,

- в чередовании знаков не будет никаких закономерностей.

Порядок действий

при использовании методов корреляционно-регрессионного анализа

1. Исследование природы рассматриваемых переменных для установления типа зависимости между переменными.

Порядок действий

2. Сбор экспериментальных данных, обсуждение вопроса об ограничениях:

2.1. Случайность выборки: несвязанность i-го наблюдения с предыдущими и отсутствие влияния на последующие.

2.2. Однородность дисперсий: рассеяния должны быть одинаковыми для всех значений независимого переменного.

2.3. Нормальность распределений.

Порядок действий

3. Построение диаграммы разброса.

4. Измерение тесноты связи, вычисление выборочного коэффициента корреляции.

5. Установление общего вида зависимости (линейная, параболическая и т.д.)

Порядок действий

6. Построение эмпирической линии регрессии методом наименьших квадратов.

7. Исследование статистических свойств регрессионной зависимости, оценка адекватности модели.

Спасибо за внимание!