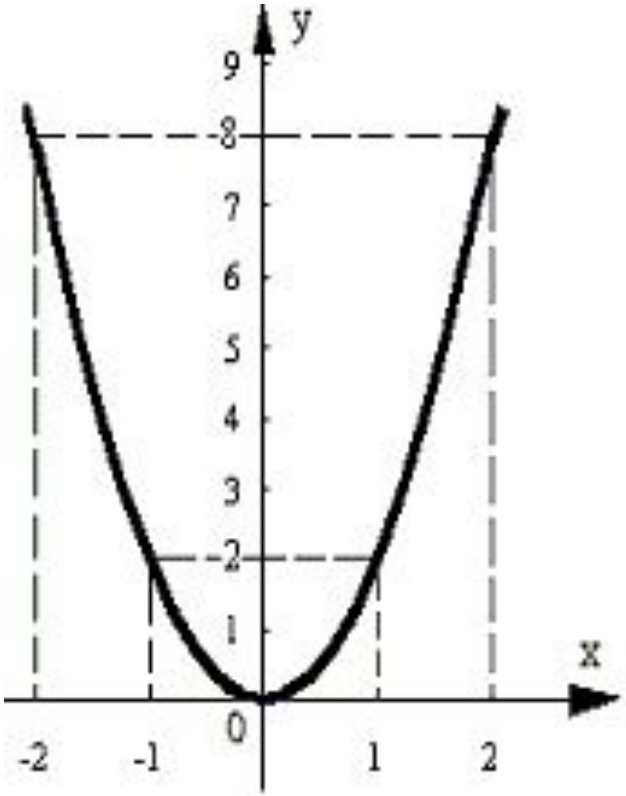
Корреляционный и регрессионный анализ

- Жорж Кювье, XYIII в., «Закон корреляции».
- Фрэнсис Гальтон, конце XIX в., понятие «корреляция» в статистике, «corelation» (соответствие).

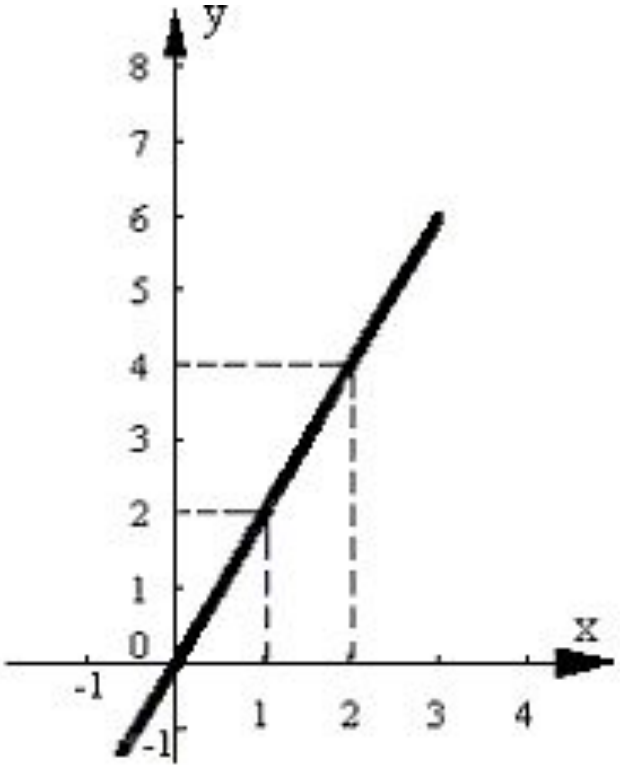
- Различают два типа связей между различными явлениями и их признаками: **функциональную и статистическую**.

- ***Статистической*** называют зависимость, при которой изменение одной из величин влечет изменение распределения других (другой), и эти другие величины принимают некоторые значения с определенными вероятностями.
- ***Функциональной*** называют зависимость, в которой значению одной переменной обязательно соответствует одно или несколько точно заданных значений другой переменной.
- В общем виде $y = f(x)$, *где y – зависимая переменная, или функция от независимой переменной x*

Примеры функциональной зависимости



$y = x^2$



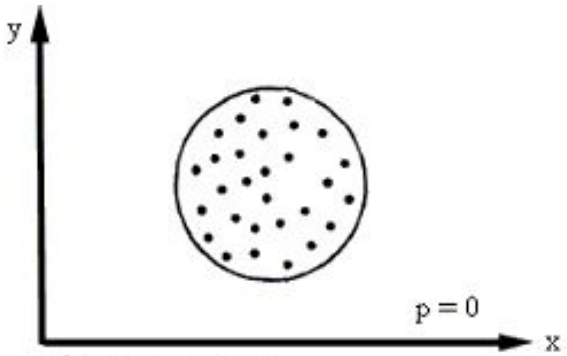
$y = a+bx^2$

- ***Корреляционная*** зависимость, характеризующая взаимосвязь значений одних случайных величин со средним значением других, хотя в каждом отдельном случае любая взаимосвязанная величина может принимать различные значения.
- Если же у взаимосвязанных величин вариацию имеет только одна переменная, а другая является детерминированной (т.е. строго определенной), то такую связь называют не корреляционной, а ***регрессионной*.**

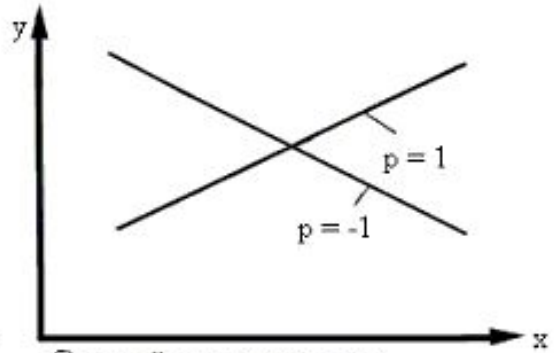
- Задачи корреляционного анализа:
- 1) измерение параметров уравнения, выражающего связь средних значений зависимой переменной со значениями независимой переменной;
- 2) измерение тесноты связи двух (или большего числа) признаков между собой.
- Вторая задача специфична для статистических связей (корреляционный анализ), а первая разработана для функциональных связей и является общей (корреляционный и регрессионный анализ).

- Для измерения тесноты связи применяется несколько показателей, например коэффициент корреляции.
- Корреляционная связь между признаками может быть *линейной и нелинейной, положительной и отрицательной.*

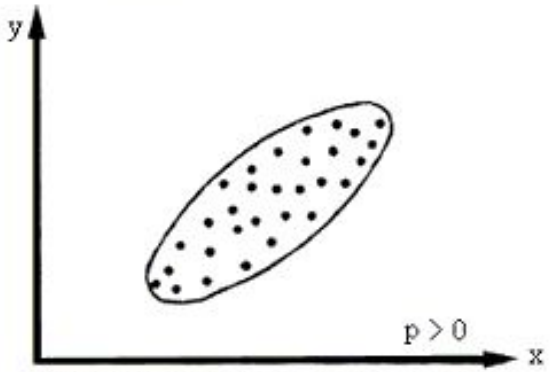
Графическая интерпретация взаимосвязи между показателями



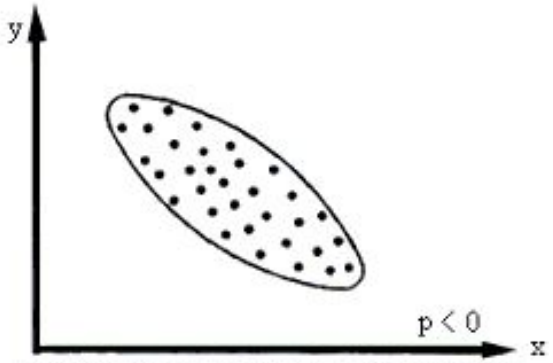
а) независимость



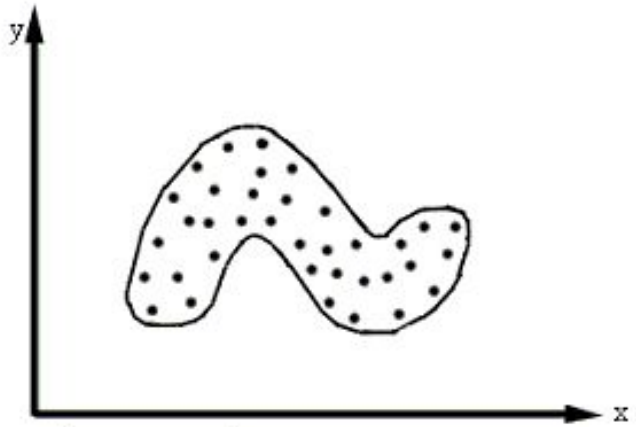
б) линейная зависимость



в) прямая положительная зависимость



г) прямая отрицательная зависимость



д) криволинейная зависимость

Регрессионный анализ

- Задачей регрессионного анализа является нахождение функциональной зависимости между зависимой *y* и независимой *x* переменными *y* = f(*x*), которую называют **регрессией (или функцией регрессии)**. График функции называют **линией или кривой регрессии**.
- На практике *x* задается, а *y* - это наблюдение какой-либо величины на опыте, в эксперименте.

Зависимость может быть линейной

$$y = \beta_0 + \beta_1 x$$

$$y = \beta_0 + \beta_1 \ln x$$

и нелинейной.

$$y = \beta_0 + \beta_1 \exp(\beta_2 x)$$

Если одна независимая переменная, то регрессию называют простой.

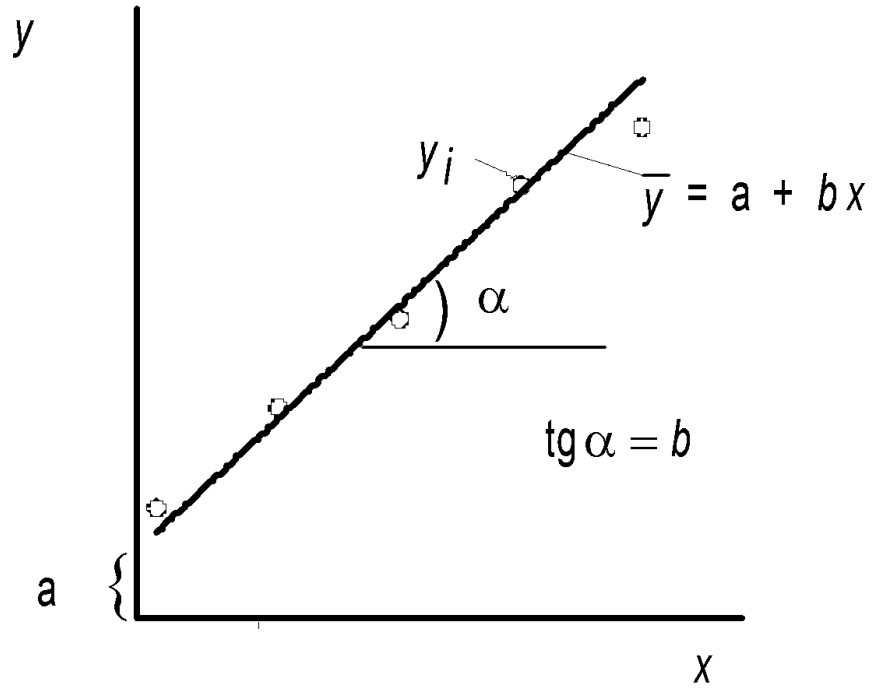
Если две независимых переменных или более, то регрессию называют множественной.

- Задачи линейного регрессионного анализа:

1. Оценка параметров линейной модели.
2. Оценка адекватности линейной модели (или тесноты линейной связи между переменными).

(Простой) линейный регрессионный анализ

Рассмотрим простую линейную модель: $y = a + bx$.



- Основным методом решения задачи нахождения параметров уравнения является метод наименьших квадратов (МНК), разработанный К. Ф. Гауссом.

Суть МНК:

Если все измерения y_i выполняются с одинаковой точностью при значениях x_i, то оценки параметров a, b определяются из условия, чтобы сумма квадратов отклонений измеренных значений y_i от рассчитанных была минимальной.

$$S = \sum_{i=1}^{n} \left[y_i - \hat{y} \right]^2 = \min$$

или

$$S = \sum_{i=1}^{n} \left[y_i - a - bx_i \right]^2 = \min$$

Параметры модели находят из условия

$$\frac{\partial S}{\partial a} = 0 \text{ и } \frac{\partial S}{\partial b} = 0 .$$

В точке минимума эти производные должны превратиться в ноль.

Продифференцируем уравнение прямой по параметрам a, b. Получим систему уравнений, решение которой дает выражение для параметров.

$$\begin{cases} -2\sum_{i=1}^{n}\left(y_i - a - bx_i\right) = 0 \\ -2\sum_{i=1}^{n}\left(y_i - a - bx_i\right)x_i = 0 \end{cases}$$

После преобразований получим:

$$\begin{cases} na + \sum_{i=1}^{n} bx_i = \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} ax_i + \sum_{i=1}^{n} bx_i^2 = \sum_{i=1}^{n} x_i y_i \end{cases}$$

После решения системы уравнений получим

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

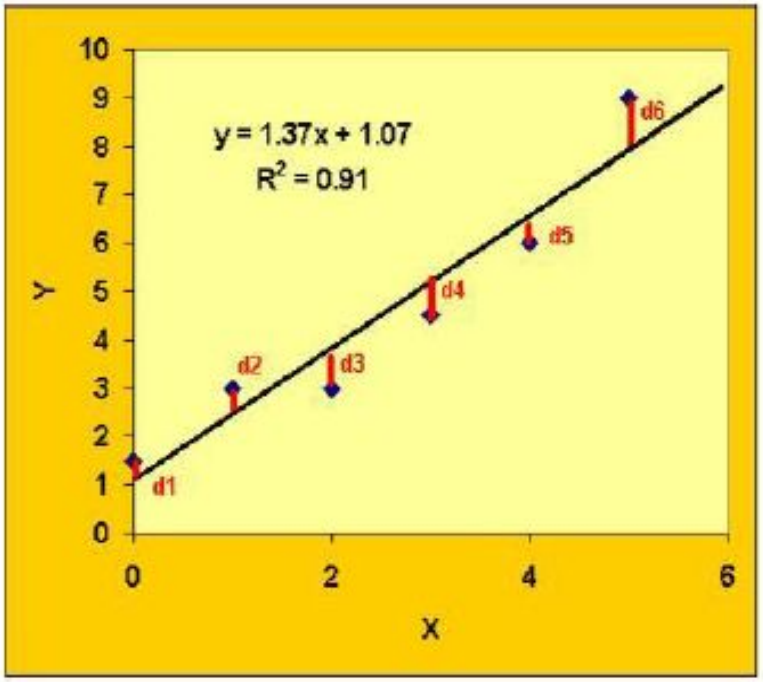
$$b = \frac{\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

- Для определения степени тесноты парной линейной зависимости (адекватности) служит **коэффициент корреляции** *r*:
- $$r = \frac{\sum_{i=1}^{n} xy - \frac{\sum_{i=1}^{n} x - \sum_{i=1}^{n} y}{n}}{\sqrt{\left[\sum_{i=1}^{n} x^2 - \left(\sum_{i=1}^{n} x\right)^2 \Big/ n\right]\left[\sum_{i=1}^{n} y^2 - \left(\sum_{i=1}^{n} y\right)^2 \Big/ n\right]}}$$
- Коэффициент корреляции может принимать значения в пределах от 0 до 1. Чем ближе он по абсолютной величине к 1, тем теснее связь.

Линейная регрессия

$$\overline{y_x} = \rho_{yx}x + b \qquad \overline{x_y} = \rho_{xy}y + c$$

Метод наименьших квадратов



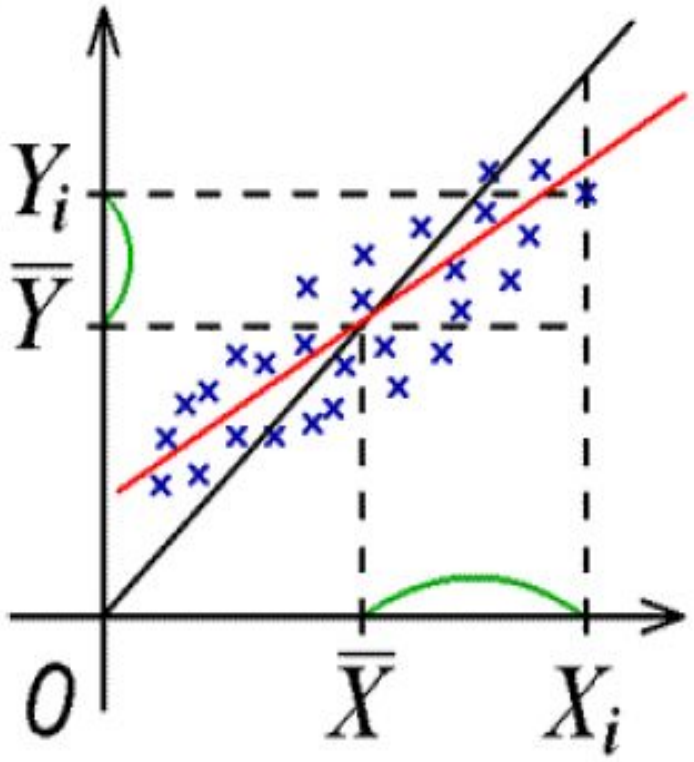
$$\sum_{i=1}^{n} \left(y_{\text{теор}_i} - y_{\text{факт}_i}\right)^2 = \sum_{i=1}^{n} d_i^2 \to min$$

Линейная корреляция

$$r_{\text{в}} = \pm\sqrt{\rho_{yx} \cdot \rho_{xy}}$$

$$\overline{y_x} - \bar{y} = r_{\text{в}} \frac{\sigma_y}{\sigma_x}(x - \bar{x})$$

$$\overline{x_y} - \bar{x} = r_{\text{в}} \frac{\sigma_x}{\sigma_y}(y - \bar{y})$$



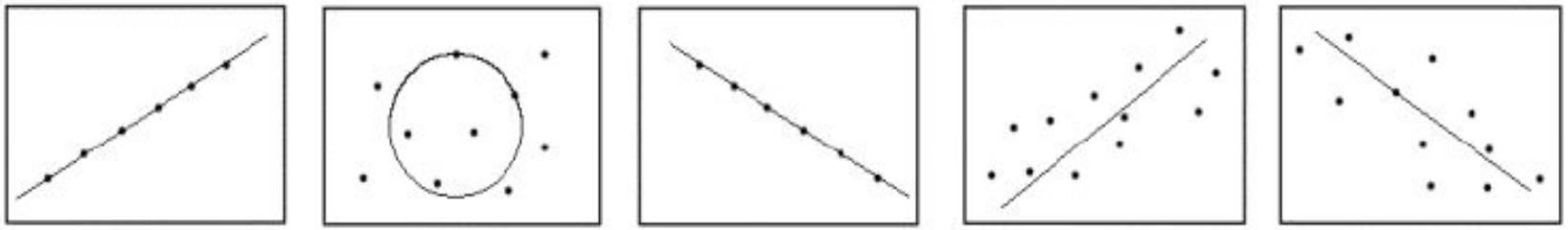


Таблица 1

$\overline{y_x} = \rho_{yx}x + b$	$\rho_{yx} = \frac{n\sum xy - \sum x \cdot \sum y}{n\sum x^2 - (\sum x)^2}$	$b = \frac{\sum x^2 \cdot \sum y - \sum x \cdot \sum xy}{n\sum x^2 - (\sum x)^2}$
$\overline{x_y} = \rho_{xy}y + c$	$\rho_{xy} = \frac{n\sum xy - \sum x \cdot \sum y}{n\sum y^2 - (\sum y)^2}$	$c = \frac{\sum y^2 \cdot \sum x - \sum y \cdot \sum xy}{n\sum y^2 - (\sum y)^2}$

Пример 3. В результате исследований зависимости двух случайных величин X и Y были получены следующие экспериментальные данные:

x_i	1,00	1,50	3,00	4,50	5,00
y_i	1,25	1,40	1,50	1,75	2,25

Определить параметры выборочного уравнения линейной регрессии Y на X и X на Y.