Понятие корреляционного, регрессионного и кластерного анализа данных

Группы методов

Две группы методов: методы корреляционного анализа и методы регрессионного анализа

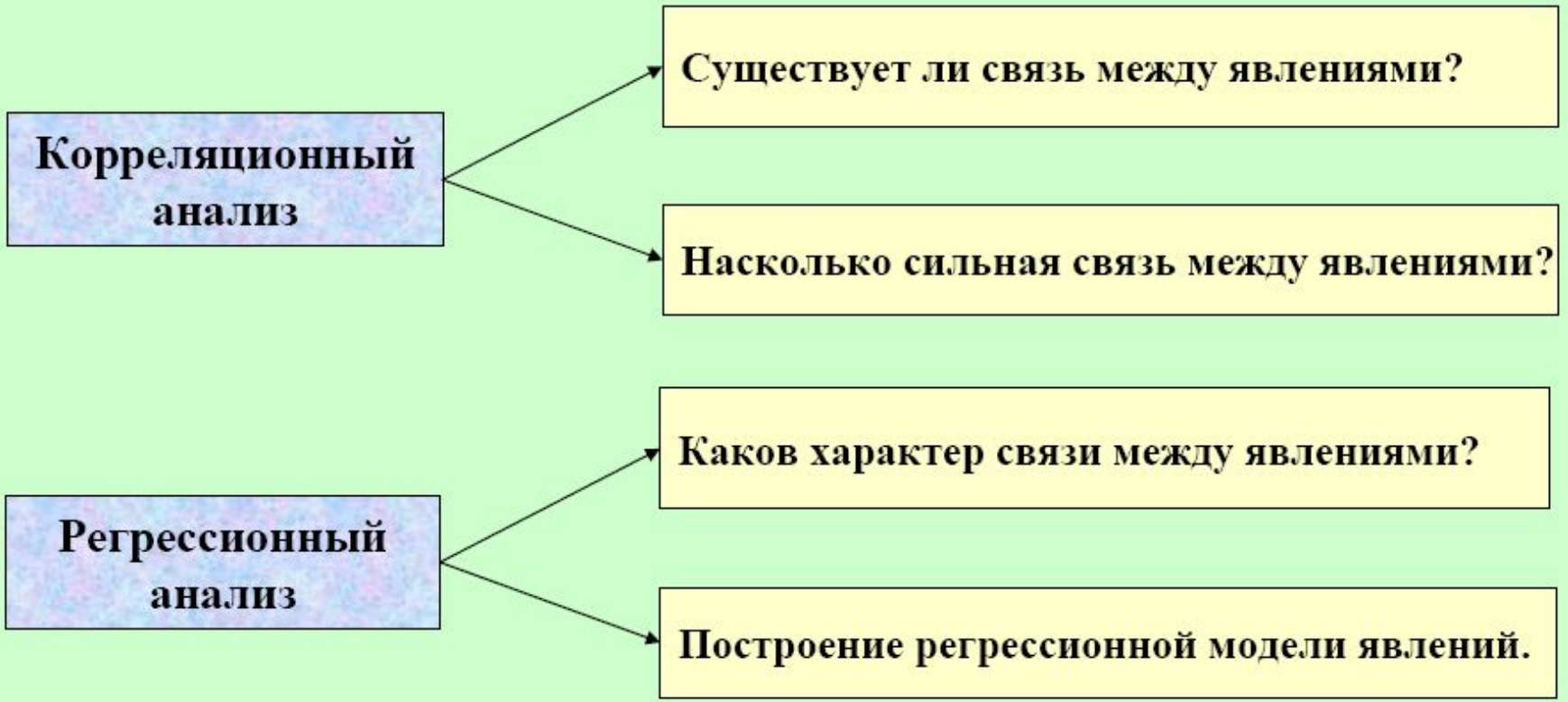
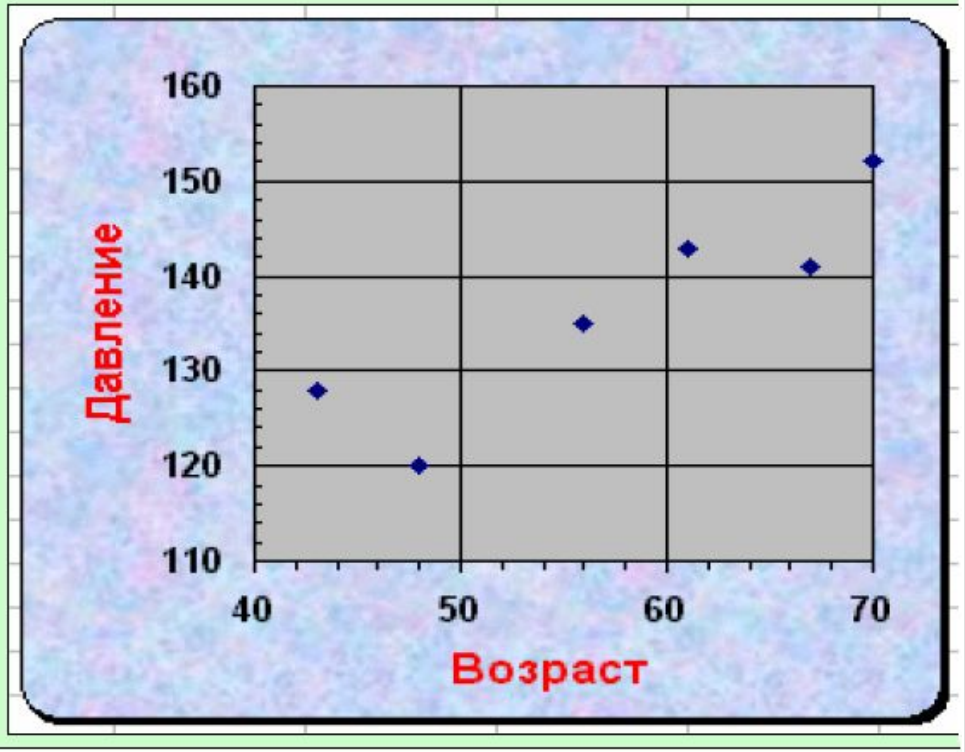


Диаграмма рассеяния

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



Свойства коэффициента корреляции

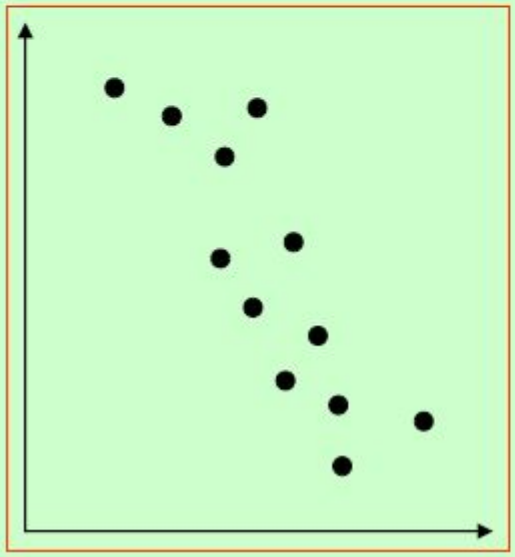
Основные свойства коэффициента корреляции:

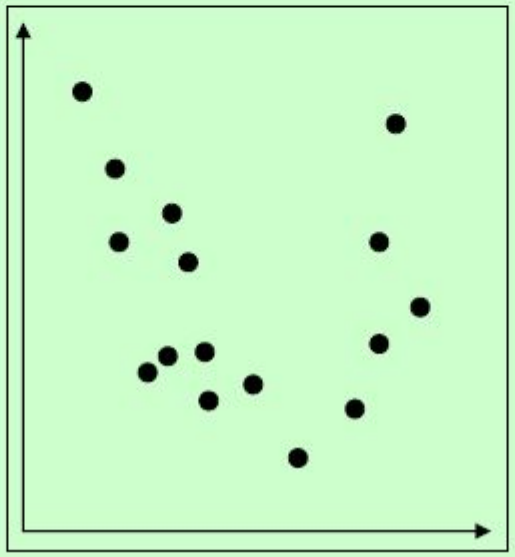
Сильная обратная связь

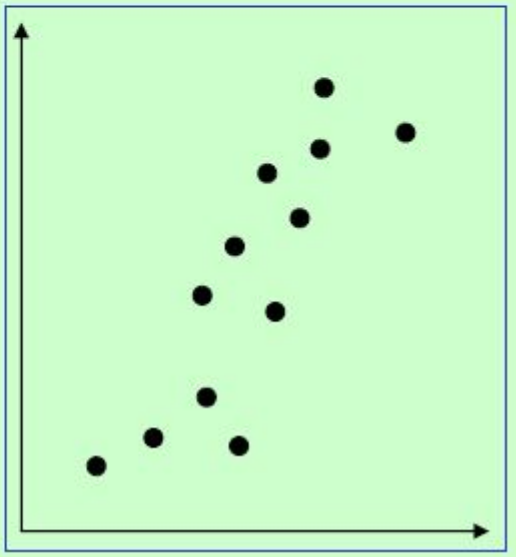
Нет линейной связи

Сильная прямая связь

-1 0 +1







Парная регрессия. Метод наименьших квадратов.

Построение уравнения регрессии

Задача построения уравнения регрессии для одного факторного и одного результативного признака формулируется следующим образом:

Пусть имеется набор значений двух переменных: результативного признака y_i и факторного признака x_i. Между этими переменными существует объективная связь вида: $y_i = f(x_i) + \varepsilon_i$.

Необходимо по данным наблюдения $y_i, x_i, i = \overline{1,n}$ подобрать функцию $\hat{y} = F(x)$, наилучшим образом описывающую существующую связь.

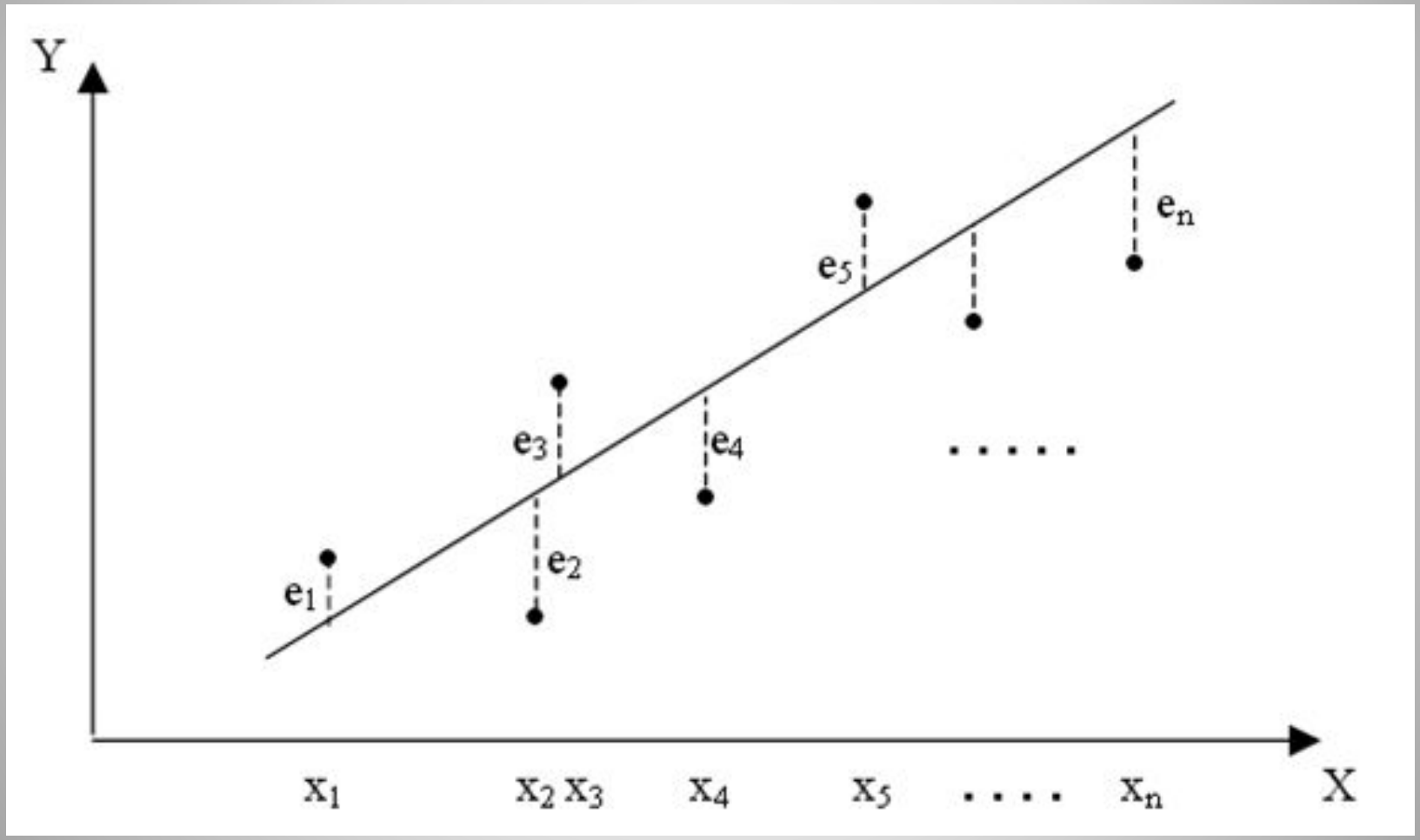
Чаще всего используются следующие зависимости :

* линейная $f(t) = a_0 + a_1 t$;
* параболическая $f(t) = a_0 + a_1 t + a_2 t^2$;
* экспоненциальная $f(t) = \exp(a_0 + a_1 t)$.

Оценка параметров осуществляется методом наименьших квадратов:

$$\sum_{t=1}^{n} (y_t - f(t))^2 \to \min$$

Графическая интерпретация



$$y_i = a + bx_i + e_i \qquad i = 1, 2, \ldots, n$$

Метод наименьших квадратов

$$S = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \to \min$$

Применение МНК для расчёта параметров регрессии рассмотрим на примере

$$\hat{y} = a + bx$$

$$S = \sum_{i=1}^{n} (y_i - a - bx_i)^2 \to \min$$

Вычисление коэффициентов

$$a = \frac{\sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i}{n} = \bar{y} - b\bar{x}$$

$$b = \frac{\frac{\sum x \sum y}{n} - \sum xy}{\frac{(\sum x)^2}{n} - \sum x^2} \quad \text{или} \quad b = r_{xy}\frac{\sigma_y}{\sigma_x}$$

Матричное представление

$$y_i = a + b\,x_i + e_i, i = 1, 2, \ldots, m$$

$$Y = X\,\beta + U$$

$$Y = \begin{pmatrix} y_1 \\ \ldots \\ y_m \end{pmatrix}$$

$$U = \begin{pmatrix} u_1 \\ \ldots \\ u_m \end{pmatrix}$$

$$\beta = \begin{pmatrix} a \\ b \end{pmatrix}$$

$$X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \ldots & \ldots \\ 1 & x_m \end{pmatrix}$$

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

Алгоритм кластерного анализа: K-means

Одной из широко используемых методик кластеризации является *разделительная кластеризация*, в соответствии с которой для выборки данных, содержащей *n* записей, задаётся число кластеров *k*, которое должно быть сформировано. Затем алгоритм разбивает все объекты выборки на *k* групп (*k*<*n*), которые и представляют собой кластеры.

К наиболее простым и эффективным алгоритмам кластеризации относится k-means (k-средних). Он состоит из четырёх шагов:

1. Задаётся число кластеров *k*, которое должно быть сформировано из объектов исходной выборки.
2. Случайным образом выбирается *k* записей, которые будут служить начальными центрами кластеров.
3. Для каждой записи исходной выборки определяется ближайший к ней центр кластера.
4. Производится вычисление *центроидов* - центров тяжести кластеров. Это делается путём определения среднего для значений **каждого** признака всех записей в кластере.

Шаги 3 и 4 повторяются до тех пор, пока не будет выполнено условие в соответствии с некоторым критерием сходимости (чаще всего используется сумма квадратов ошибок между центроидом кластера и всеми вошедшими в него записями).

Остановка алгоритма производится, когда на каждой итерации в каждом кластере остаётся один и тот же набор записей.

Пример

Заданы координаты пяти точек на плоскости:

(1, 5), (7, 4), (5, 2), (7, 5), (4, 7).

Задано число кластеров: 2.

Проведите кластеризацию методом K-means. Выполните две итерации.