Тема 2. **Однофакторная линейная регрессия**

Постановка задачи

В самых разных областях знания возникает *задача определения зависимости между случайными величинами,* являющимися признаками одних и тех же объектов. Например, это может быть зависимость

- между ростом и весом человека;
- между силой сигнала на входе и выходе технического устройства;
- между затратами компании на рекламу и доходом от продаж;
- между уровнем инфляции и безработицей;
- между содержанием радиоактивного вещества в растениях-медоносах и в мёде, полученном от этих растений.

На практике на значение исследуемой величины влияет множество *факторов,* но для простоты мы будем считать, что основное влияние оказывает один из них, потому и анализ будем называть ***однофакторным***.

Будем считать, что оба признака, зависимость между которыми мы стараемся выявить, представимы как значения вещественных переменных. Предположим, что нам известны результаты n измерений. Каждое измерение $i(i=1,\ldots,n)$ даёт пару чисел (x_i, y_i) – значения двух признаков измеряемого объекта, (например, рост – вес, затраты на рекламу – доход и т.д.), т.е. сырые данные представимы как таблица:

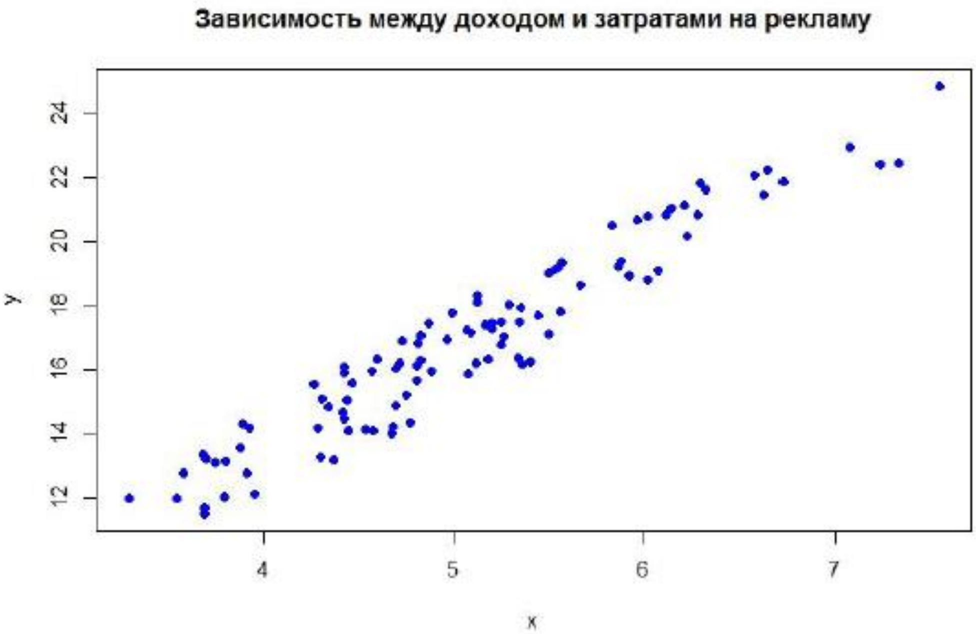
№ наблюдения, i	Значения фактора, x_i	Значения переменной отклика, y_i
1	x_1	y_1
…		
i	x_i	y_i
…		
n	x_n	y_n

Здесь каждая строка соответствует одному *объекту* (*наблюдению*). Признак, который может быть непосредственно измерен (x), является ***фактором*** (***предиктором***), прогнозируемая переменная (y) – ***переменная отклика***.

Цель исследования – построить (линейную) функцию (регрессионную модель), которая позволит прогнозировать значение переменной отклика (y) *по известному значению фактора* (x)

Визуализация сырых данных

Построим систему координат, где по оси абсцисс будем откладывать значения фактора (x), по оси ординат – значения переменной отклика (y). Таким образом, каждому наблюдению (т.е. каждой паре) (x_i, y_i) $\left(i=\overline{1,n}\right)$ соответствует точка на координатной плоскости. Если зависимость между изучаемыми признаками была бы *линейной* и отсутствовала бы случайная компонента, то все эти точки лежали бы на одной прямой. Однако из-за наличия случайного «шума» точки оказываются разбросанными по координатной плоскости в виде так называемого «облака». Пример такого облака показан на приведённом ниже рисунке.



Здесь ось абсцисс соответствует затратам на рекламу, ось ординат – объёму продаж (зафиксированному через заданное время после проведения рекламной кампании).

Поставим задачу: *найти такую линейную функцию, которая наилучшим образом отражает зависимость переменной отклика y (объёма продаж) от фактора x* (затрат на рекламу).

Эта задача называется задачей ***однофакторной линейной регрессии***.

Приведём математическую формулировку задачи.

Математическая постановка задачи нахождения уравнения регрессии

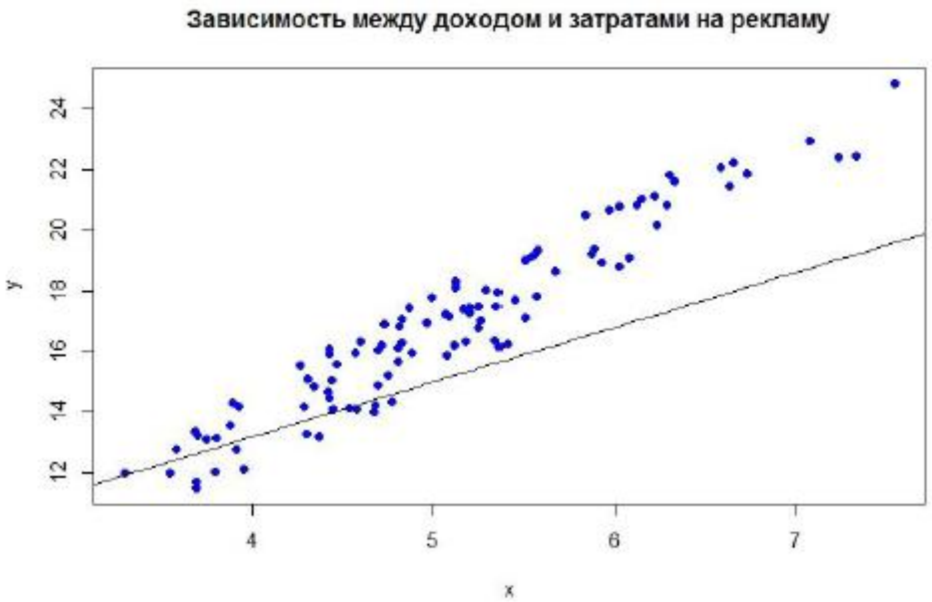
Необходимо найти зависимость $y = f(x)$. Значения y и x измеряют в процессе эксперимента и при анализе они уже известны. Однако вид их функции связи (модель) до опыта не известен и должен быть найден по опытным данным. При этом имеется в виду, что на то, какое значение примет y , влияет не только значение x, но также ряд мешающих неуправляемых факторов, к которым относятся погрешности измерения, неконтролируемые изменения окружающей среды и другие. Поэтому даже при фиксированных значениях x функция $y = f(x)$ ведет себя случайным образом, в связи с чем ставится задача нахождения ее математического ожидания и дисперсии или доверительных интервалов. Каждому значению x соответствует некоторое вероятностное распределение случайной величины (СВ) Y . Предположим, что СВ Y «в среднем» линейно зависит от значений переменной x . Это означает, что условное математическое ожидание случайной величины Y при заданном значении переменной x имеет вид (Y среднее значение всех значений y для данного x)

$$M(Y|x) = \beta_0 + \beta_1 x \quad (1)$$

Функция переменной x , определяемая правой частью этой формулы, называется линейной регрессией Y на x, а параметры β_0, β_1 - параметрами линейной регрессии.

β_1 – тангенс угла наклона графика этой функции к оси OX (английский термин: «*slope*»), β_0– ордината точки пересечения этой прямой с осью *OY* (англ.: «*intercept*»). Задача состоит в том, чтобы *найти такие значения переменных* β_0, β_1, *при которых прямая (1) наилучшим образом проходит через облако точек* (x_i, y_i), $i = 1,..n$.

Поясним смысл задачи геометрически. Зафиксируем произвольные значения β_0 и β_1 и построим соответствующую прямую:



Очевидно, построенная «наугад» прямая является не самой лучшей для данного облака точек. Формализуем понятие «качества» модели. При фиксированных β_0 и β_1 «ожидаемое» (согласно (1)) значение y при $x = x_i$ составляет $\beta_0 + \beta_1 \cdot x_i$, $i = 1,...n$ (т.е. точка $(x_i, \beta_0 + \beta_1 \cdot x_i)$ лежит на построенной прямой). Но фактическое значение переменной y при $x = x_i$ составляет y_i, т.е. «ошибка» составляет $((\beta_0 + \beta_1 \cdot x_i) - y_i)$.

На практике параметры линейной регрессии неизвестны и их оценки определяются по результатам наблюдений переменных y_i и x_i . Пусть проведено n независимых наблюдений случайной величины Y при значениях переменной X: $x_1, x_2, ..., x_n$. При этом измерения величины Y дали следующие результаты: $y_1, y_2, ..., y_n$. Так как эти значения имеют разброс относительно линейной регрессии, то связь между переменными Y и X можно записать в виде линейной (по параметрам β_0, β_1) регрессионной модели:

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

где ε - случайная ошибка наблюдений, причем $M(\varepsilon) = 0$, $D(\varepsilon) = \sigma^2$. Значение дисперсии ошибок σ^2 неизвестно, и оценка ее определяется по результатам наблюдений.

Задача линейного регрессионного анализа состоит в том, чтобы по результатам наблюдений (x_i, y_i), $i = \overline{1, n}$:

1.Получить наилучшие точечные и интервальные оценки неизвестных параметров β_0, β_1 и σ^2 линейной регрессионной модели.

2) Проверить статистические гипотезы о параметрах модели.

3) Проверить, достаточно ли хорошо модель согласуется с результатами наблюдений (адекватность модели результатам наблюдений).

В соответствии с моделью результаты наблюдений зависимой переменной Y : $y_1, y_2, ..., y_n$ являются реализациями случайных величин $\beta_0 + \beta_1 x_i + \varepsilon_i$, обозначаемых y_i, $i = \overline{1, n}$.

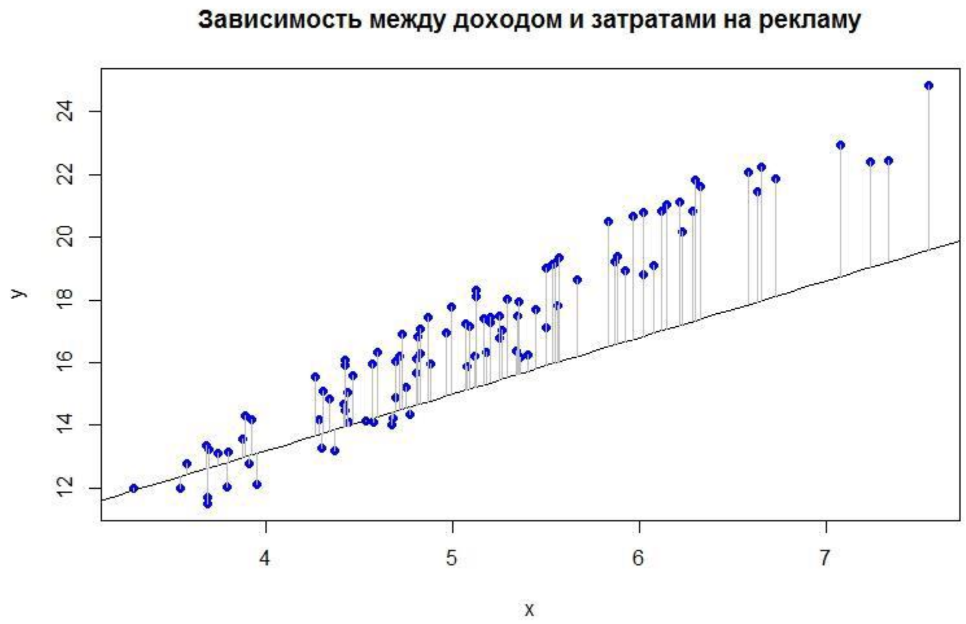
Задача линейного регрессионного анализа решается в предположении, что случайные ошибки наблюдений ε_i и ε_j не коррелированы, имеют математические ожидания равные нулю, и одну и ту же дисперсию, равную σ^2, т.е. $M(\varepsilon_i) = 0$, $K_{\varepsilon_i \varepsilon_j} = \begin{cases} 0, i \neq j \\ \sigma^2, i = j \end{cases}$, $i = \overline{1, n}$. При статистическом анализе регрессионной модели предполагается также, что случайные ошибки наблюдений ε_i , $i = \overline{1, n}$ имеют нормальное распределение, т.е. $\varepsilon_i \sim N(0, \sigma^2)$, $i = \overline{1, n}$. В этом случае ошибки наблюдений также являются независимыми случайными величинами. Можно определить величину ошибки для всех отмеченных точек. Линейная модель, которая наилучшим образом аппроксимирует данные – одна из тех, для которых общая ошибка выборки имеет наименьшее значение. Чтобы рассчитать ее нужно избежать позитивных и негативных значений. Это можно сделать, возведя все ошибки в квадрат и делая их положительными величинами. Линия наилучшего подбора – та, которая минимизирует квадраты разниц между рассматриваемыми значениями y и соответствующими значениями x, рассчитанными с помощью линии наилучшего подбора. Эта линия называется **линией регрессии, полученной методом наименьших квадратов**. Для нахождения оценок параметров модели по результатам наблюдений используется **метод наименьших квадратов**. По этому методу выбирают такие оценки β_0, β_1, которые минимизируют сумму квадратов отклонений наблюдаемых значений случайных величин y_i от их математических ожиданий, т.е.

$$\sum_{i=1}^{n}\left(\left(\beta_0 + \beta_1 \cdot x_i\right) - y_i\right)^2 \to \min \qquad (2)$$

Метод наименьших квадратов

Принцип поиска коэффициентов регрессии путём минимизации суммы квадратов отклонений между реальными значениями признака и прогнозируемыми согласно предполагаемой форме зависимости (в нашем случае – линейной) называется *методом наименьших квадратов* (англ.: Least Square Method, LSM).

Проиллюстрируем целевую функцию задачи (2) на следующем рисунке.



Значение целевой функции задачи (2) при фиксированных значениях β_0 и β_1 равно сумме квадратов длин построенных отрезков. Из рисунка видно, что построенная прямая – не лучшая, так как можно провести прямую, обеспечивающую меньшее значение целевой функции задачи (2).

Найдём решение задачи (2). Целевую функцию задачи (2) обозначим через $\varphi(\beta_0, \beta_1)$. Очевидно, $\varphi(\beta_0, \beta_1)$ – дифференцируемая функция *двух* переменных. Найдём её частные производные:

$$\frac{\partial \varphi}{\partial \beta_0} = 2\sum_{i=1}^{n}\left(\left(\beta_0 + \beta_1 \cdot x_i\right) - y_i\right) \qquad (3)$$

$$\frac{\partial \varphi}{\partial \beta_1} = 2\sum_{i=1}^{n}\left(\left(\beta_0 + \beta_1 \cdot x_i\right) - y_i\right)x_i$$

и запишем систему для поиска стационарной точки:

$$\begin{cases} \sum_{i=1}^{n}((\beta_0 + \beta_1 \cdot x_i) - y_i) = 0 \\ \sum_{i=1}^{n}((\beta_0 + \beta_1 \cdot x_i) - y_i)x_i = 0 \end{cases} \qquad (4)$$

После несложных преобразований системы (4) получим

$$\begin{cases} n\beta_0 + \beta_1 \cdot \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i = 0 \\ \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \cdot \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i y_i = 0 \end{cases} \qquad (5)$$

Введём обозначения: $\overline{x} = \frac{1}{n}\sum_{n=1}^{n} x_i$, $\overline{y} = \frac{1}{n}\sum_{n=1}^{n} y_i$, $\overline{x \cdot y} = \frac{1}{n}\sum_{n=1}^{n} x_i y_i$ и $\overline{x^2} = \frac{1}{n}\sum_{n=1}^{n} x_i^2$.

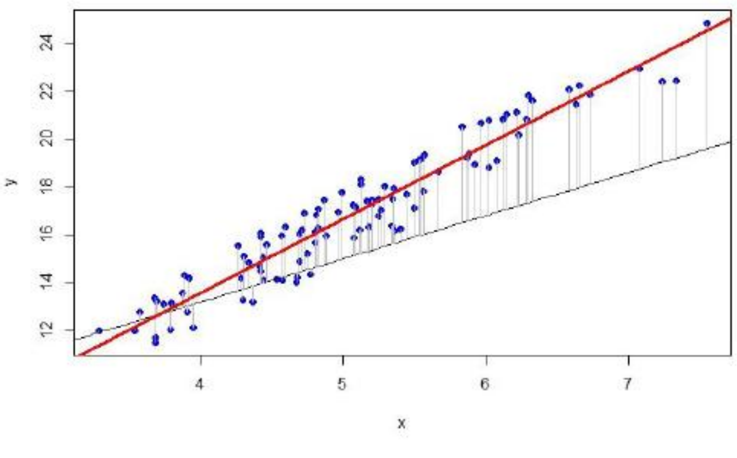
Поделив оба уравнения системы (5) на n , получим

$$\begin{cases} \beta_0 + \beta_1 \cdot \overline{x} - \overline{y} = 0 \\ \beta_0 \overline{x} + \beta_1 \cdot \overline{x^2} - \overline{x \cdot y} = 0 \end{cases} \qquad (6)$$

Нетрудно показать, что система (5) имеет единственное решение (β_0^*, β_1^*), где

$$\beta_1^* = \frac{\overline{x} \cdot \overline{y} - \overline{x \cdot y}}{\overline{x}^2 - \overline{x^2}} \quad \beta_0^* = \overline{y} - \beta_1^* \cdot \overline{x} \qquad (7)$$

Учитывая свойства функции $\varphi(\beta_0, \beta_1)$, нетрудно показать также, что это решение (т.е., стационарная точка функции $\varphi(\beta_0, \beta_1)$ является *точкой минимума* функции $\varphi(\beta_0, \beta_1)$. Иными словами, определяемые формулами (7) значения β_0^* и β_1^* обеспечивают получение наилучшей (в смысле задачи (2)) линейной функции, отражающей зависимость переменной отклика y от фактора x. График этой линейной зависимости называется *прямой регрессии* (y на x). На приведённом ниже рисунке эта прямая имеет красный цвет.



Найденная линейная функция позволяет *прогнозировать* значение зависимого признака (y) по заданным значениям независимого фактора (x).