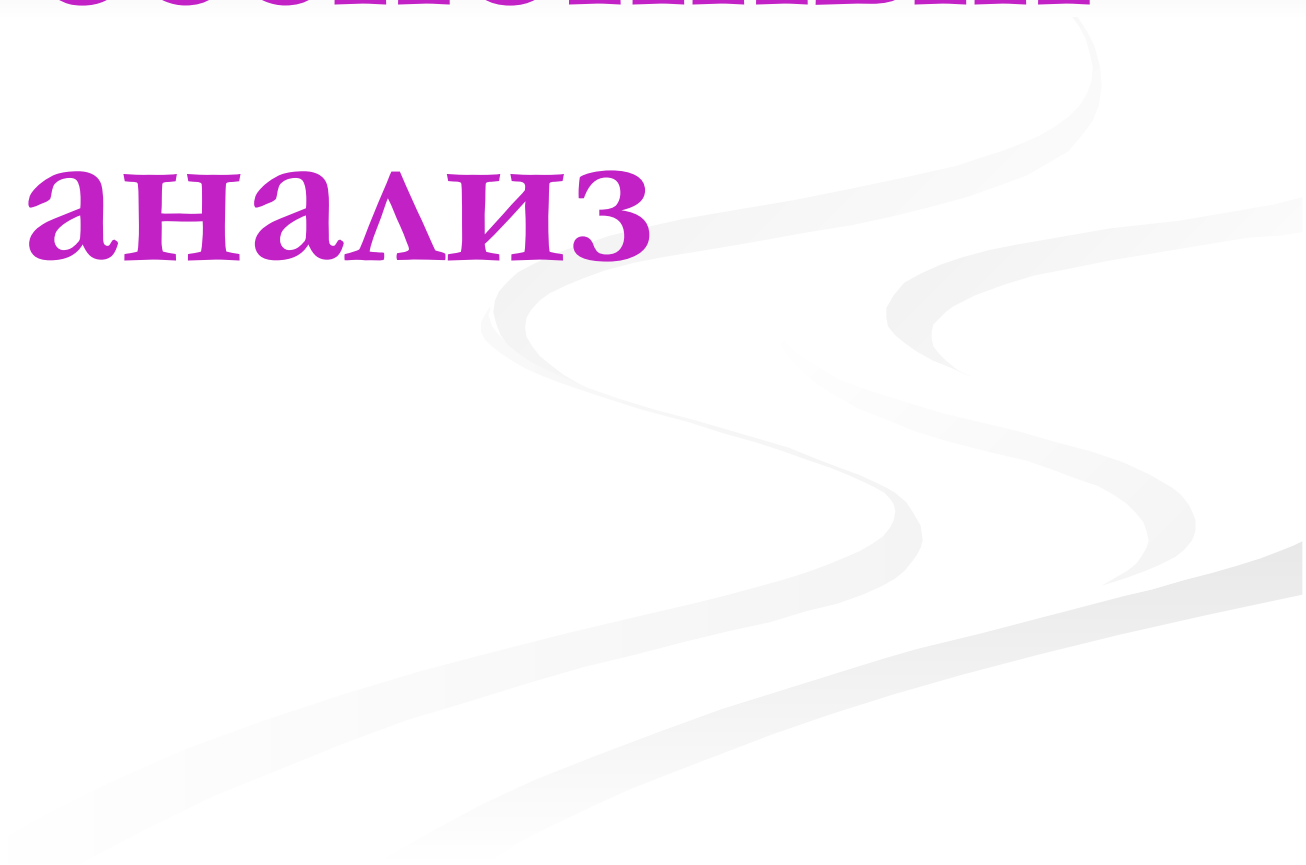


Регрессионный анализ



Понятие регрессии

- Для выражения регрессии служат эмпирические и теоретические ряды, их графики — *линии регрессии*, а также *корреляционные уравнения (уравнения регрессии)* и *коэффициент линейной регрессии*.
- Показатели регрессии выражают корреляционную связь двусторонне, учитывают изменение средней величины признака Y при изменении значений x_i признака X , и, наоборот, показывают изменение средней величины признака X по измененным значениям y_i признака Y . Исключение составляют временные ряды, или ряды динамики, показывающие изменение признаков во времени. Регрессия таких рядов является односторонней.

Ряды регрессии, особенно их графики, дают наглядное представление о форме и тесноте корреляционной связи между признаками, в чем и заключается их ценность. Форма связи между показателями, влияющими на уровень спортивного результата и общей физической подготовки занимающихся физической культурой и спортом, может быть разнообразной. И поэтому задача состоит в том, чтобы любую форму корреляционной связи выразить уравнением определенной функции (линейной, параболической и т.д.), что позволяет получать нужную информацию о корреляции между переменными величинами Y и X , предвидеть возможные изменения признака Y на основе известных изменений X , связанного с Y корреляционно.

Уравнение линейной регрессии

Обычно признак Y рассматривается как функция многих аргументов — x_1, x_2, x_3, \dots — и может быть записана в виде:

$$\underline{y = a + bx_1 + cx_2 + dx_3 + \dots}$$

a, b, c и d — параметры уравнения, определяющие соотношение между аргументами и функцией. В практике учитываются не все, а лишь некоторые аргументы, в простейшем случае, как при описании линейной регрессии, — всего один:

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

В этом уравнении параметр a — свободный член; графически он представляет отрезок ординаты (y) в системе прямоугольных координат. Параметр b называется коэффициентом регрессии. С точки зрения аналитической геометрии b — угловой коэффициент, определяющий наклон линии регрессии по отношению к осям, координат. В области регрессионного анализа этот параметр показывает, насколько в среднем величина одного признака (Y) изменяется при изменении на единицу меры другого корреляционно связанного с Y признака X . Наглядное представление об этом параметре и о положении линий регрессии Y по X и X по Y в системе прямоугольных координат дает следующий рисунок -

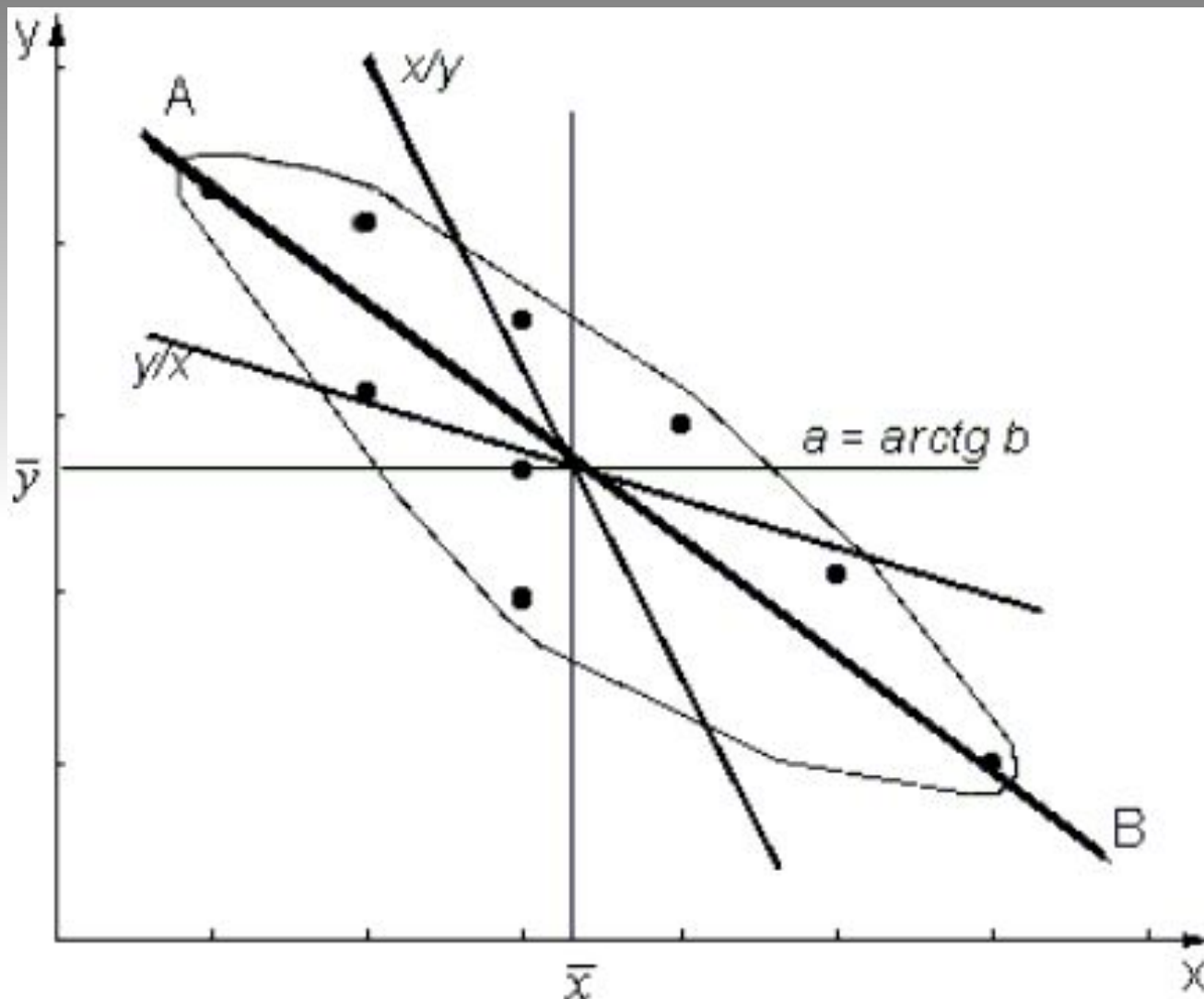


Схема линий регрессии Y по X и X по Y в системе прямоугольных координат.

Линии регрессии, как показано, пересекаются в точке O , соответствующей средним арифметическим значениям корреляционно связанных друг с другом признаков Y и X . Линия AB , проходящая через эту точку, изображает полную (функциональную) зависимость между переменными величинами Y и X , когда коэффициент корреляции $r = 1$. Чем сильнее связь между Y и X , тем ближе линии регрессии к AB , и, наоборот, чем слабее связь между варьирующими признаками, тем более удаленными оказываются линии регрессии от AB . При отсутствии связи между признаками, когда $r = 0$, линии регрессии оказываются под прямым углом (90°) по отношению друг к другу.

Уравнение регрессии тем лучше описывает зависимость, чем меньше рассеяние диаграммы, чем больше теснота взаимосвязи. Уравнение прямой линии пригодно для описания только линейных зависимостей. В случае нелинейных зависимостей математическая запись может отображаться уравнениями параболы, гиперболы и др. Необходимо также сделать одно важное замечание о значении показателей, характеризующих взаимосвязь признаков (коэффициентов корреляции, регрессии и т. п.). Все они дают лишь количественную меру связи, но ничего не говорят о причинах зависимости. Определить эти причины — дело самого исследователя.

Коэффициенты уравнения парной линейной регрессии

Как уже было определено выше, в случае линейной зависимости уравнение регрессии является уравнением прямой линии. Таких уравнений два:

$$Y = a_1 + b_{y/x}X \text{ — прямое}$$
$$\text{и } X = a_2 + b_{x/y}Y \text{ — обратное,}$$

где: a и b — коэффициенты, или параметры, которые надлежит определить. Значение коэффициентов регрессии вычисляется по формуле:

$$b_{x/y} = r \cdot \frac{\sigma_x}{\sigma_y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (y_i - \bar{y})^2}$$

$$b_{y/x} = r \cdot \frac{\sigma_y}{\sigma_x} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

Коэффициенты регрессии b имеют размерность, равную отношению размерностей изучаемых показателей X и Y , и тот же знак, что и коэффициент корреляции.

Коэффициенты a определяются по формуле:

$$a_1 = \bar{y} - b_{y/x} \cdot \bar{x}$$

$$a_2 = \bar{x} - b_{x/y} \cdot \bar{y}$$

Чтобы вычислить эти коэффициенты, надо просто в уравнения регрессии подставить средние значения коррелируемых переменных. Для оценки качества уравнений регрессии вычисляются остаточные средние квадратические отклонения (или абсолютные погрешности уравнений) по формуле:

$$\sigma_{y/x} = \sigma_y \cdot \sqrt{1 - r^2}$$

и

$$\sigma_{x/y} = \sigma_x \cdot \sqrt{1 - r^2}$$

Эти оценки абсолютны и, следовательно, не могут быть сравнимы друг с другом. Поэтому вводят оценки **относительной погрешности уравнений**, которые выражаются в процентах и **служат для точности предсказания (прогнозирования) результатов одного показателя по заранее известным значениям другого**. Относительные погрешности уравнений регрессии определяются по формуле:

$$\sigma'_{y/x} = \frac{\sigma_{y/x}}{\bar{y}} \cdot 100\%$$

и

$$\sigma'_{x/y} = \frac{\sigma_{x/y}}{\bar{x}} \cdot 100\%$$

Связь между коэффициентами регрессии и корреляции

Между коэффициентом корреляции и параметром парной линейной регрессии существует зависимость, которая применительно к выборочным оценкам может быть представлена следующим образом:

$$b = r_{xy} \cdot \frac{S_y}{S_x}$$

где: S_y и S_x — средние квадратические ошибки.

Приведенное выражение позволяет оценить параметр регрессии без решения системы нормальных уравнений при условии, что коэффициент корреляции уже определен. На основе формулы легко показать, что **выборочный коэффициент корреляции** равен среднему геометрическому выборочных коэффициентов регрессии. Действительно, сравнив формулы с основной формулой коэффициента корреляции, видим, что их числители равны $\sum (x_i - \bar{x})(y_i - \bar{y})$

Это свидетельствует об определенной связи между этими характеристиками. Выборочный коэффициент корреляции выражается тогда равенством $r^2 = b_{y/x} \cdot b_{x/y}$, откуда

$$r = \pm \sqrt{b_{y/x} \cdot b_{x/y}}$$

Эта формула ценна тем, что, во-первых, может быть использована для нахождения неизвестной величины коэффициента корреляции по известным значениям коэффициента регрессии $b_{y/x}$ и $b_{x/y}$, а во-вторых, позволяет контролировать правильность расчета коэффициента корреляции, если известны величины $b_{y/x}$ и $b_{x/y}$.

Знак выборочного коэффициента корреляции совпадает со знаком выборочных коэффициентов регрессии, что следует из формулы . Если зависимость между признаками функциональная, то $b_{y/x} = 1 / b_{x/y}$ и, следовательно, $r = 1$. И, наоборот, при полном отсутствии взаимосвязи между признаками $b_{y/x} = 0$, $b_{x/y} = 0$, и $r = 0$.

Определение параметров парной линейной регрессии

Определение параметров линейной регрессии – одна из задач регрессионного анализа. Она решается способом наименьших квадратов, основанным на требовании, чтобы сумма квадратов отклонений вариант от линии регрессии была наименьшей. *Этому требованию удовлетворяет следующая система нормальных уравнений:*

$$an + b \sum x = \sum y,$$

$$a \sum x + b \sum x^2 = \sum xy.$$

Ряды регрессии — это ряды усредненных значений (y_x и x_y) варьирующих признаков Y и X , соответствующих значениям аргументов x_i и y_i . Поэтому эмпирические уравнения регрессии следует записывать так:

$$y_x = ay/x + by/x * x$$

$$x_y = ax/y + bx/y * y$$

Формулы для определения параметров **a** и **b** принимают следующие выражения:

$$a_{y/x} = \bar{y} - b_{y/x} \cdot \bar{x}$$

$$a_{x/y} = \bar{x} - b_{x/y} \cdot \bar{y}$$

Уравнение линейной регрессии можно выразить в виде отклонений вариант от их средних арифметических:

$$y_x - \bar{y} = b_{y/x} \cdot (x_i - \bar{x})$$

$$x_y - \bar{x} = b_{x/y} \cdot (y_i - \bar{y})$$

В таком случае система нормальных уравнений для определения параметров a и b будет следующая:

$$an + b \sum (x_i - \bar{x}) = \sum (y_i - \bar{y});$$

$$a \sum (x_i - \bar{x}) + b \sum (x_i - \bar{x})^2 = \sum (y_i - \bar{y})(x_i - \bar{x}).$$

Заменяя параметры $b_{y/x}$ и $b_{x/y}$ получим систему уравнений парной линейной регрессии:

$$y_x = \bar{y} + \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \cdot (x - \bar{x})$$

$$x_y = \bar{x} + \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (y_i - \bar{y})^2} \cdot (y - \bar{y})$$

Формулы расчета свободных коэффициентов

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

$$b_0 = \frac{\sum x_i \cdot \sum y_i^2 - \sum y_i \cdot \sum x_i y_i}{\sum y_i^2 - \sum (y_i)^2}$$

Упрощенные формулы для расчета коэффициентов регрессии получаются из систем уравнений

$$\begin{cases} a_0 \cdot N + a_1 \sum x_i = \sum y_i \\ a_0 \cdot \sum x_i + a_1 \cdot \sum (x_i \cdot x_i) = \sum y_i \cdot x_i \end{cases}$$

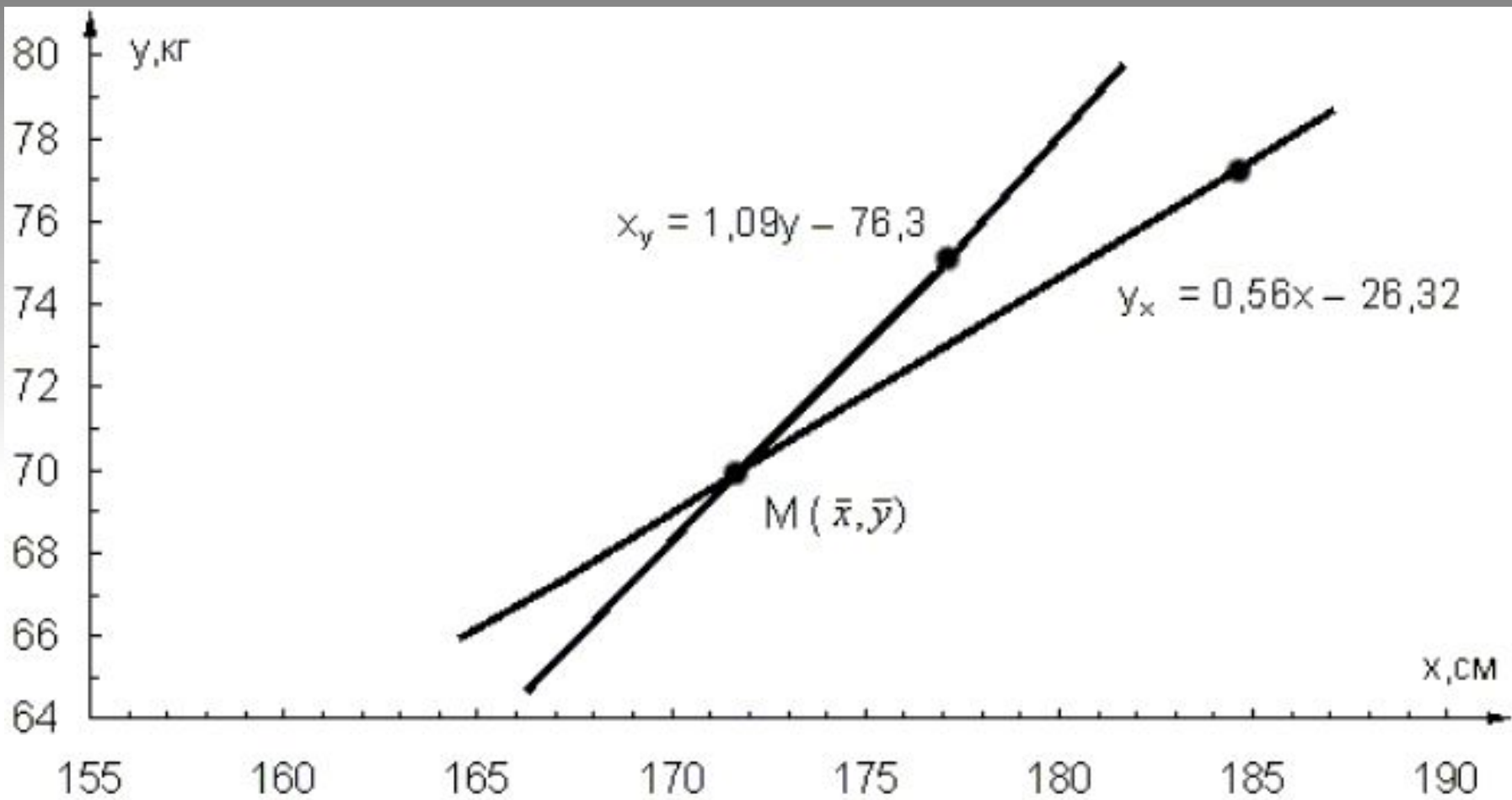
$$\begin{cases} b_0 \cdot N + b_1 \sum y_i = \sum x_i \\ b_0 \cdot \sum y_i + b_1 \cdot \sum (y_i \cdot y_i) = \sum y_i \cdot x_i \end{cases}$$

Графическое представление уравнения парной линейной регрессии

Эмпирические ряды регрессии Y по X и X по Y изображаются в виде линейного графика, при построении которого наиболее точным является использование способа наименьших квадратов, предложенного в 1806 г. К. Гауссом и независимо от него А. Лежандром. В основу этого способа положена теорема, согласно которой сумма квадратов отклонений вариант (x_i) от средней арифметической \bar{x} - есть величина наименьшая, т. е.
$$\sum (x_i - \bar{x})^2 = \min$$

При графическом изображении эмпирического уравнения регрессии (например, показатели роста и веса 10 исследуемых), представленного на рисунке используется следующая последовательность:

1. Определив форму и направление взаимосвязи между эмпирическими данными на основе данных расчета нормированного коэффициента корреляции, производят расчет уравнений регрессии (прямого и обратного) по формуле.
2. Подставляя в конечный вид уравнений, выражающих зависимость между переменными величинами Y и X , эмпирические данные x_i и y_i находят координаты точек линий регрессии для усредненных значений u_x и u_y .
3. На графике, выполненном в прямоугольной системе координат, на оси x откладывают значения переменных x_i , на оси y – значения y_i и отмечают точками рассчитанные координаты линий регрессии для усредненных значений u_x и u_y .
4. Две линии регрессии на графике пересекаются в точке M с координатами средних значений показателей x_i и y_i .



Графическое изображение эмпирического уравнения регрессии.

График линий регрессии отражает ряды теоретически ожидаемых значений функции по известным значениям аргумента. При этом, чем сильнее взаимосвязь между величинами x_i и y_i , тем ^{± 1} меньше угол между линиями регрессии. При $r = \pm 1$ линии уравнения регрессии либо совпадают, либо расположены параллельно, так как корреляционная зависимость между признаками в этом случае переходит в функциональную. И, наоборот, чем слабее зависимость между признаками, тем больше угол между линиями на графике. При $r = 0$ линии регрессии расположены перпендикулярно.

Коэффициент τ - Кендалла

$$\tau_{\text{ýïï}} = \frac{P - Q}{N \cdot (N - 1)} \cdot \frac{1}{2} = 1 - \frac{4 \cdot Q}{N \cdot (N - 1)} = \frac{4 \cdot P}{N \cdot (N - 1)} - 1$$

P – число совпадений

Q – число инверсий

N – число ранжируемых признаков

Расчет уровня значимости коэффициента корреляции

$$t_{\hat{\rho}} = |r_{y\ddot{y}}| \cdot \sqrt{\frac{n-2}{1-r_{y\ddot{y}}^2}}$$

Величина T_{ϕ} проверяется на уровень значимости по таблице для t-критерия Стьюдента