**Лекция**

**Регрессионный анализ**

**2.1. Понятие регрессии**

В практических исследованиях возникает необходимость аппроксимировать (описать приблизительно) диаграмму рассеяния математическим уравнением. То есть зависимость между переменными величинами Y и Х можно выразить аналитически с помощью формул и уравнений и графически в виде геометрического места точек в системе прямоугольных координат. График корреляционной зависимости строится по уравнениям функции  которые называются **регрессией** (термин “регрессия” происходит от лат. regressio — движение назад). Здесь  и — средние арифметические из числовых значений зависимых переменных Y и X.

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, учитывая изменение средней величины признака Y при изменении значений xi признака X, и, наоборот, показывают изменение средней величины признака Х по измененным значениям yi 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. 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.

**2.2. Уравнение линейной регрессии**

Обычно признак Y рассматривается как функция многих аргументов — x1, x2, x3, ...— и может быть записана в виде:

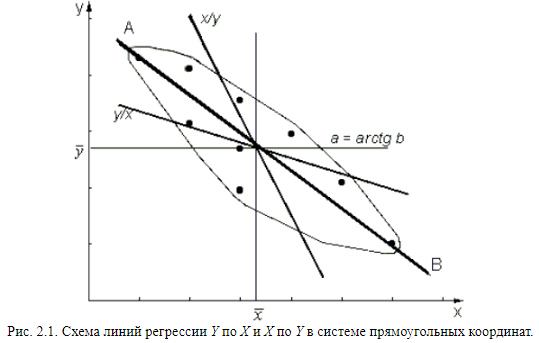
y = a + bx1 + cx2 + dx3 + ... ,

где: а, b, с 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:

y = a + bx (2.1)

В этом уравнении параметр а — свободный член; графически он представляет отрезок ординаты (у) 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 Х и X по Y в системе прямоугольных координат дает рисунок 2.1.



Линии регрессии, как показано, пересекаются в точке 0 (), соответствующей средним арифметическим значениям корреляционно связанных друг с другом признаков Y и X. Линия АВ, 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 АВ, и, наоборот, чем слабее связь между варьирующими признаками, тем более удаленными оказываются линии регрессии от АВ. 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.

**2.3. Коэффициенты уравнения парной линейной регрессии**

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

Y = a1 + by/x X — прямое

и X = a2 + bx/y Y — обратное, (2.2)

где: a и b – коэффициенты, или параметры, которые надлежит определить.

Значение коэффициентов регрессии вычисляется по формуле:

;

.

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

Коэффициенты а определяются по формуле:

;

.

Чтобы вычислить этот 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:

;

.

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:

;

.

Значение этой оценки, если r = ±1, равно нулю и, если r = 0, максимально. Остаточное среднее квадратическое отклонение характеризует колеблемость Y относительно линии регрессии по Х в прямом уравнении регрессии и, наоборот, в обратном случае. А, следовательно, чем меньше величина относительной погрешности уравнения регрессии, тем точнее будет оно осуществлять прогноз значений одного показателя по заранее известным значениям другого.

2.4. 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:

 (2.7)

где: и - средние квадратические ошибки.

Приведенное выражение позволяет оценить параметр регрессии без решения системы нормальных уравнений при условии, что коэффициент корреляции уже определен. На основе формулы (2.7) легко показать, что *выборочный коэффициент корреляции* равен среднему геометрическому выборочных коэффициентов регрессии. Действительно, Сравнив формулы (2.3) с основной формулой (1.1) коэффициента корреляции, видим, что их числители равны . Это свидетельствует об определенной связи между этими характеристиками. Выборочный коэффициент корреляции выражается тогда равенством , откуда

 (2.8)

Эта формула ценна тем, что, во-первых, может быть использована для нахождения неизвестной величины коэффициента корреляции по известным значениям коэффициента регрессии *by/x* и *bx/y*, а во-вторых, позволяет контролировать правильность расчета коэффициента корреляции, если известны величины *by/x* и *bx/y*. Знак выборочного коэффициента корреляции совпадает со знаком выборочных коэффициентов регрессии, что следует из формулы (2.3). Если зависимость между признаками функциональная, то *by/x = 1 / bx/y* и, следовательно, *r=1*. И, наоборот, при полном отсутствии взаимосвязи между признаками *by/x=0, bx/y=0*, и *r = 0*.

**9.5. 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:



.

Ряды регрессии — это ряды усредненных значений (yx и xy) варьирующих признаков Y и X, соответствующих значениям аргументов xi и yi. Поэтому эмпирические уравнения регрессии следует записывать так:

yx = ay/x + by/x\*x

и xy = ax/y + bx/y\*y (2.9)

Формулы для определения параметров а и b принимают следующие выражения:



 (2.10)

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



 (2.11)

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



Поскольку  и , то параметр *b* выразится в виде приведенной формулы (2.3); параметр *а* легко найти по формуле (2.4). Если средние https://masters.donntu.ru/2005/kita/tokarev/diss/images/1(10).gif и https://masters.donntu.ru/2005/kita/tokarev/diss/images/1(11).gif перенести в правую часть уравнения (2.11), то при https://masters.donntu.ru/2005/kita/tokarev/diss/images/1(12).gif система нормальных уравнений принимает следующий вид:

https://masters.donntu.ru/2005/kita/tokarev/diss/images/1(13).gif

И https://masters.donntu.ru/2005/kita/tokarev/diss/images/1(14).gif. (2.12)

Заменив в формуле (2.11) параметры *by/x* и *bx/y* на их значения из формулы (2.3), получим систему уравнений парной линейной регрессии:



 (2.13)

Эти уравнения удобны для определения параметров при отыскивании эмпирических уравнений регрессии в практической работе для точности прогнозирования результатов.

**2.6. Графическое представление уравнения парной линейной регрессии**

Эмпирические ряды регрессии Y по Х и Х 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 (xi) от средней арифметической (https://masters.donntu.ru/2005/kita/tokarev/diss/images/2.gif) есть величина наименьшая, т. е. https://masters.donntu.ru/2005/kita/tokarev/diss/images/2(1).gif. Отсюда и название метода, который нашел широкое применение не только в биологии, но и в технике. Мы уже говорили об этом методе и применяли его, когда находили параметры а и b линейной регрессии, отыскивая эмпирическое уравнение.

При графическом изображении эмпирического уравнения регрессии (используется следующая последовательность:

1. Определив форму и направление взаимосвязи между эмпирическими данными на основе данных расчета нормированного коэффициента корреляции, производят расчет уравнений регресиии (прямого и обратного) по формуле (2.13).
2. Подставляя в конечный вид уравнений, выражающих зависимость между переменными величинами Y и X, эмпирические данные xi и yi находят координаты точек линий регрессии для усредненных значений yx и xy.
3. На графике, выполненном в прямоугольной системе координат, на оси x откладывают значения переменных xi, на оси у – значения yi и отмечают точками рассчитанные координаты линий регрессии для усредненных значений yx и xy (рис.2.2).
4. Две линии регрессии на графике пересекаются в точке М с координатами средних значений показателей xi и yi.

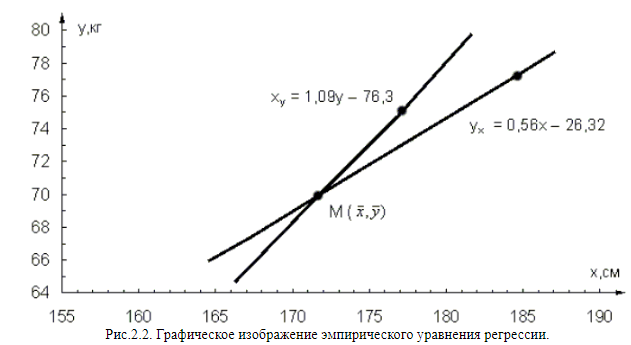


График линий регрессии отражает ряды теоретически ожидаемых значений функции по известным значениям аргумента. При этом, чем сильнее взаимосвязь между величинами xi и yi, тем меньше угол между линиями регрессии. При r =±1 линии уравнения регресии либо совпадают, либо расположены параллельно, так как 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.