16. КОРРЕЛЯЦИОННЫЕ ЗАВИСИМОСТИ. УРАВНЕНИЯ РЕГРЕССИИ. ВЫБОРОЧНЫЙ КОЭФФИЦИЕНТ КОРРЕЛЯЦИИ И ВЫБОРОЧНОЕ КОРРЕЛЯЦИОННОЕ ОТНОШЕНИЕ

План:

1. Статистические и корреляционные зависимости

2. Выборочный коэффициент корреляции и выборочное корреляционное отношение

Ключевые слова: *функциональная зависимость, статистическая (вероятностная или стохастическая) зависимость, корреляционная зависимость, условные средние, уравнения регрессии, функция регрессии, линия регрессии, выборочный коэффициент корреляции, выборочное уравнение прямой линии регрессии, метод наименьших квадратов, выборочный коэффициент корреляции.*

Выборочный коэффициент корреляции, выборочное корреляционное отношение, теснота корреляционной зависимости, криволинейные корреляции, множественная корреляция.

1.1. Функциональная, статистическая и корреляционная зависимости

Во многих задачах требуется установить и оценить зависимость изучаемой случайной величины Y от одной или нескольких других величин. Рассмотрим сначала зависимость (связь) Y от одной случайной (или неслучайной) величины X.

В некоторых случаях эта связь является настолько тесной что, зная, какое значение приняла величина X, можно однозначно предсказать значение Y; это означает, что связь между величинами X и Y - функциональная. Возможен, однако, и другой крайний случай, когда зависимость между X и Y отсутствует вовсе, т.е. величины X и Y независимы. Точное определение независимости случайных величин было дано ранее.

В общем случае связь между величинами X и Y находит свое выражение в том, что при фиксированном значении x величины X, величина Y остается случайной, но с законом распределения, зависящим от X. Иначе

говоря, каждому значению $X = x$ отвечает свой закон, распределения величины Y. Рассмотренные выше крайние случаи – функциональная зависимость и полная независимость - вполне укладываются в эту общую схему; функциональная зависимость $Y = f(X)$ означает, что при фиксированном значении $X = x$ величина Y принимает единственное значение $f(x)$ (с вероятностью 1), а полная независимость означает, что при любом значении x величины X закон распределения величины Y - один и тот же (он не зависит от выбранного нами значения величины X).

Связь между двумя случайными величинами, проявляющаяся том, что изменение одной из них влечет за собой изменение закона распределения другой, называется статистической (или вероятностной или стохастической).

Вероятностная связь между двумя случайными величинами X и Y появляется обычно тогда, когда имеются общие случайные факторы, влияющие как на X , так и на Y (наряду с другими факторами, неодинаковыми для X и Y). Например, если X представляет собой некоторую функцию от случайных величин U и V :

$$X = f(U,V),$$

а Y есть функция от той же самой величины и другой случайной величины W :

$$X = f(U,V),$$

то величины X и Y будут связаны между собой вероятностной связью.

Определение. Статистической называют зависимость, при которой изменение одной из величин влечет изменение распределения другой. В частности, статистическая зависимость проявляется в том, что при изменении одной из величин изменяется среднее значение другой; в этом случае статистическую зависимость называют корреляционной.

Приведем пример случайной величины Y , которая не связана с величиной $\boldsymbol{X}$ функционально, а связана корреляционно. Пусть Y - урожай зерна, X - количество удобрений. С одинаковых по площади участков земли при

равных количествах внесенных удобрений снимают различный урожай, т.е. Y не является функцией от X. Это объясняется влиянием случайных факторов (осадки, температура воздуха и др.). Вместе с тем, как показывает опыт, средний урожай является функцией от количества удобрений, т.е. Y связан с X корреляционной зависимостью.

1.2. Условные средние. Корреляционная зависимость.

Уточним определение корреляционной зависимости, для чего введем понятие условной средней.

Предположим, что изучается связь между случайной величиной Y и случайной величиной X. Пусть каждому значению X соответствует несколько значений Y. Например, пусть при $x_1 = 2$ величина Y приняла значения: $y_1 = 5$, $y_2 = 6$, $y_3 = 10$. Найдем среднее арифметическое этих чисел:

$$\overline{y}_2 = \frac{5+6+10}{3} = 7.$$

Число $\overline{y}_2$ называют условным средним; черточка над буквой y служит обозначением среднего арифметического, а число 2 указывает, что рассматриваются те значения Y, которые соответствуют $x_1 = 2$.

Применительно к примеру предыдущего пункта эти данные можно истолковать так: на каждый из трех одинаковых участков земли внесли по 2 единицы удобрений и сняли соответственно 5; 6 и 10 единиц зерна; средний урожай составил 7 соответствующих единиц.

Определение. Условным средним $\overline{y}_x$ называют среднее арифметическое значений Y, соответствующих значению $X = x$.

Если каждому значению x соответствует одно значение условной средней, то, очевидно, условная средняя есть функция от x; в этом случае говорят, что случайная величина Y зависит от X корреляционно.

Определение. Корреляционной зависимостью Y от X называют функциональную зависимость условной средней $\overline{y}_x$ от x:

$$\bar{y}_x = f(x). \qquad (1)$$

Определение. Уравнение (1) называют уравнением регрессии Y на X; функцию $\boldsymbol{f(x)}$ называют регрессией Y на X, а ее график - линией регрессии Y на X.

Аналогично определяется условная средняя $\bar{x}_y$, и корреляционная зависимость X от Y.

Условным средним $\bar{x}_y$ значений X, соответствующих $Y = y$.

Определение. Корреляционной зависимостью X от Y называют функциональную зависимость условной средней $\bar{x}_y$ от y:

$$\bar{x}_y = \varphi(x). \qquad (2)$$

Определение. Уравнение (2) называют уравнением регрессии X и Y функцию $\varphi(x)$ называют регрессией X на Y, а ее график - линией регрессии X на Y.

1.3. Две основные задачи теории корреляции.

Первая задача теории корреляции - установить форму корреляционной связи, т.е. вид функции регрессии (линейная, квадратичная показательная и т. д.). Наиболее часто функции регрессии оказываются линейными. Если обе функции регрессии $f(x)$ и $\varphi(x)$ линейны, то корреляцию называют линейной; в противном случае - нелинейной. Очевидно, при линейной корреляции обе линии регрессии являются прямыми линиями.

Вторая задача теории корреляции - оценить тесноту (силу) корреляционной связи. Теснота корреляционной зависимости Y от X оценивается по величине рассеяния значений Y вокруг условного среднего $\bar{y}_x$. Большое рассеяние свидетельствует о слабой зависимости Y от X либо об отсутствии зависимости. Малое рассеяние указывает наличие достаточно сильной зависимости; возможно даже, что Y и X связаны функционально, но под воздействием второстепенных случайных факторов эта связь оказалась размы-

той, в результате чего при одном и том же значении x величина Y принимает различные значения.

Аналогично (по величине рассеяния значений X вокруг условного среднего $\overline{x}_y$) оценивается теснота корреляционной связи X от Y.

1.4. Отыскание параметров выборочного уравнения прямой линии регрессии по несгруппированным данным

Допустим, что количественные признаки X и Y связаны линейной корреляционной зависимостью. В этом случае обе линии регрессии будут прямыми.

Предположим, что для отыскания уравнений этих прямых проведено n независимых испытаний, в результате которых получены n пар чисел:

$$(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n),$$

Поскольку наблюдаемые пары чисел можно рассматривать как случайную выборку из генеральной совокупности всех возможных значений случайной величины (X, Y), то величины и уравнения, найденные по этим данным, называют выборочными.

Для определенности будем искать выборочное уравнение прямой линии регрессии Y на X.

Рассмотрим простейший случай: различные значения x признака X и соответствующие им значения y признака Y наблюдались по одному разу. Очевидно, что группировать данные нет необходимости. Также нет надобности использовать понятие условной средней, поэтому искомое уравнение

$$\overline{y}_x = kx + b$$

можно записать так:

$$Y = kx + b$$

Угловой коэффициент прямой линии регрессии Y на X принято называть выборочным коэффициентом регрессии Y на X и обозначать через ρ_{yx}.

Итак, будем искать выборочное уравнение прямой линии регрессии Y на X вида:

$$Y = \rho_{yx} x + b. \qquad (1)$$

Поставим своей задачей подобрать параметры ρ_{yx}, и b так, чтобы точки $(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)$, построенные по данным наблюдений на плоскости XOY, как можно ближе лежали вблизи прямой (1).

Уточним смысл этого требования. Назовем отклонением разность

$$Y_i - y_i \qquad (i = 1,2,\ldots,n),$$

uде Y_i - вычисленная по уравнению (1) ордината, соответствующая наблюдаемому значению x_i; y_i - наблюдаемая ордината, соответствующая $\boldsymbol{x_i}$.

Подберем параметры ρ_{yx} и b так, чтобы сумма квадратов отклонений была минимальной (в этом состоит сущность метода наименьших квадратов).

Так как каждое отклонение зависит от отыскиваемых параметров, то и сумма квадратов отклонений есть функция F этих параметров (временно вместо ρ_{yx} будем писать ρ):

$$F(\rho, b) = \sum_{i-1}^{n} (Y_i - y_i)^2, \quad \text{или} \quad F(\rho, b) = \sum_{i-1}^{n} (\rho x_i + b_i - y_i)^2.$$

Для отыскания минимума приравняем нулю соответствующие частные производные:

$$\frac{\partial F}{\partial \rho} = 2\sum_{i-1}^{n} (\rho x_i + b_i - y_i) x_i;$$

$$\frac{\partial F}{\partial b} = 2\sum_{i-1}^{n} (\rho x_i + b_i - y_i).$$

(Для простоты записи вместо $\sum_{i=1}^{n}$ будем писать $\sum$)

Выполнив элементарные преобразования, получим систему двух линейных уравнений относительно ρ и b

$$\begin{cases}\left(\sum x^2\right)\rho + \left(\sum x\right)b = \sum xy, \\ \left(\sum x\right)\rho + nb = \sum y. \end{cases} \qquad (2)$$

Решив эту систему, найдём искомые параметры:

$$\rho_{yx} = \frac{n\sum xy - \sum x \cdot \sum y}{n\sum x^2 - \left(\sum x\right)^2}, \qquad (3)$$
$$b = \frac{\sum x^2 \cdot \sum y - \sum x \cdot \sum xy}{n\sum x^2 - \left(\sum x\right)^2}.$$

Аналогично можно найти выборочное уравнение прямой линии регрессии X на Y:

$$\overline{x}_y = \rho_{xy} y + c,$$

где ρ_{xy} – выборочный коэффициент регрессии X на Y.

Пример. Найти выборочное уравнение прямой линии регрессии Y на X по данным $n = 5$ наблюдений:

x	1	1,5	3	4,5	5
y	1,25	1,4	1,5	1,75	2,25

Решение

Составим расчетную таблицу 1.

Таблица 1.

x_i	y_i	x_i^2	$x_i y_i$
1,00	125	1,00	1,250
1,50	1,40	2,25	2,100
3,00	1,50	9,00	4,500
4,50	1,75	20,25	4,875
5,00	2,25	25,00	11,250
$\sum \boldsymbol{x_i} = 15$	$\sum \boldsymbol{y_i} = 8{,}15$	$\sum \boldsymbol{x_i^2} = 57{,}50$	$\sum \boldsymbol{x_i y_i} = 26{,}975$

Найдем искомые параметры, для чего подставим вычисленные по таблице суммы в соотношения (3):

$$\rho_{yx} = \frac{5 \cdot 26{,}975 - 15 \cdot 8{,}15}{5 \cdot 57{,}5 - 15^2} = 0{,}202;$$

$$b = \frac{57{,}5 \cdot 8{,}15 - 15 \cdot 26{,}975}{5 \cdot 57{,}5 - 15^2} = 1{,}024.$$

Напишем искомое уравнение регрессии:

$$y = 0{,}202x + 1{,}024$$

Для того чтобы получить представление, насколько хорошо вычисленные по этому уравнению значения Y_i согласуются с наблюдаемыми значениями $\boldsymbol{y_i}$, найдем отклонения $Y_i - y_i$, Результаты вычислений сведены в таблицу 2.

Таблица 2.

x_i	Y_i	y_i	$Y_i - y_i$
1,00	1,226	1,25	-0,024
1,50	1,327	1,40	-0,073
3,00	1,630	1,50	0,130
4,50	1,993	1,75	0,083
5,00	2,034	2,25	-0,216

Как видно из таблицы, не все отклонения достаточно малы. Это объясняется малым числом наблюдений.

1.5. Корреляционная таблица

При большом числе наблюдений одно и то же значение x может встретиться n_x раз, одно и то же значение у может встретиться $\boldsymbol{n_y}$ раз, одна и та же пара чисел (x, y) может наблюдаться n_{xy} раз. Поэтому данные наблюдений группируют, т.е. подсчитывают частоты n_x , n_y , n_{xy} . Все сгруппированные данные записывают в виде таблицы, которую называют корреляционной.

Поясним устройство корреляционной таблицы на примере (табл. 3).

Таблица 3.

X / Y	10	20	30	40	n_y
0,4	5	-	7	14	26
0,6	-	2	6	4	12
0,8	3	19	-	-	22
n_x	8	21	13	18	$n = 60$

В первой строке таблицы указаны наблюдаемые значения (10; 20; 30; 40) признака X, а в первом столбце - наблюдаемые значения (0,4; 0,6; 0,8) признака Y. На пересечении строк и столбцов вписаны частоты n_{xy} наблюдаемых пар значений признаков, Например, частота 5 указывает, что пара чисел (10; 0,4) наблюдалась 5 раз. Все частоты помещены в прямоугольнике, клетки которого выделены. Черточка означает, что соответственная пара чисел, например (20; 0,4), не наблюдалась.

В последнем столбце записаны суммы частот строк. Например, сумма частот первой строки прямоугольника, клетки которого выделены, равна $n_y = 5 + 7 + 14 = 26$; это число указывает, что значение признака Y, равное 0,4 (в сочетании с различными значениями признака X) наблюдалось 26 раз.

В последней строке записаны суммы частот столбцов. Например, число 8 указывает, что значение признака X, равное 10 (в сочетании с различными значениями признака Y) наблюдалось 8 раз.

В клетке, расположенной в нижнем правом углу таблицы, помещена сумма всех частот (общее число всех наблюдений n). Очевидно

$$\sum n_x = \sum n_y = n.$$

В нашем примере $\sum n_x = 8 + 21 + 13 + 18 = 60$ и $\sum n_y = 26 + 12 + 22 = 60$.

1.6. Отыскание параметров выборочного уравнения прямой линии регрессии по сгруппированным данным. Выборочный коэффициент корреляции

В п. 4 для определения параметров уравнения прямой линии регрессии Y на X была получена система уравнений (2):

$$\begin{cases}\left(\sum x^2\right)\rho_{yx}+\left(\sum x\right)b=\sum xy,\\ \left(\sum x\right)\rho_{yx}+nb=\sum y.\end{cases} \tag{4}$$

Предполагалось, что значения X и соответствующие им значения Y наблюдались по одному разу. Теперь же допустим, что получено большое число данных (практически для удовлетворительной оценки искомых параметров должно быть хотя бы 50 наблюдений), среди них есть повторяющиеся, и они сгруппированы в виде корреляционной таблицы. Запишем систему (4) так, чтобы она отражала данные корреляционной таблицы. Воспользуемся тождествами:

$$\sum x = n\bar{x} \text{ следствие } \bar{x}=\frac{\sum x}{n};$$

$$\sum y = n\bar{y} \text{ следствие } \bar{y}=\frac{\sum y}{n};$$

$$\sum x^2 = n\overline{x^2} \text{ следствие } \overline{x^2}=\frac{\sum x^2}{n}.$$

$\sum n_x = \sum n_y$ (учтено, что пара чисел (x, y) наблюдалась n_{xy} раз).

Подставив правые части тождеств в систему (4) и сократив обе части второго уравнения на n, получим:

$$\begin{cases}\left(n\overline{x^2}\right)\rho_{yx}+\left(n\bar{x}\right)b=\sum n_{xy}xy,\\ \left(\bar{x}\right)\rho_{yx}+b=\bar{y}\end{cases} \tag{5}$$

Решив эту систему, найдем параметры $\boldsymbol{\rho_{yx}}$ и $\boldsymbol{b}$ и, следовательно, искомое уравнение:

$$\bar{y}_x=\rho_{yx}x+b. \tag{*}$$

Однако более целесообразно, введя новую величину - коэффициент корреляции, написать уравнение регрессии в ином виде. Сделаем это.

Найдем b из второго уравнения (5):

$$b = \bar{y} - \rho_{yx} x$$

Подставив правую часть этого равенства в уравнение (*), получим:

$$\bar{y}_x - \bar{y} = \rho_{yx}\left(x - \bar{x}\right) \qquad (6)$$

Найдем из системы (4) коэффициент регрессии, учитывая, что $\overline{x^2} - \left(\bar{x}\right)^2 = \sigma_x^2$:

$$\rho_{yx} = \frac{\sum n_{xy} xy - n\overline{xy}}{n[\overline{x^2} - \left(\bar{x}\right)^2]} = \frac{\sum n_{xy} xy - n\overline{xy}}{n\sigma_x^2}.$$

Умножим обе части равенства на дробь $\frac{\sigma_x}{\sigma_y}$:

$$\rho_{yx} \cdot \frac{\sigma_x}{\sigma_y} = \frac{\sum n_{xy} xy - n\overline{xy}}{n\sigma_x \sigma_y}.$$

Обозначим правую часть равенства через $r_в$ и назовем ее выборочным коэффициентом корреляции:

$$\rho_{yx} \cdot \frac{\sigma_x}{\sigma_y} = r_в, \quad \text{или} \quad \rho_{yx} = r_в \cdot \frac{\sigma_y}{\sigma_x}.$$

Подставив правую часть этого равенства в (6),окончательно получим выборочное уравнение прямой линии регрессии Y на X вида

$$\bar{y}_x - \bar{y} = r_в \frac{\sigma_y}{\sigma_x}\left(x - \bar{x}\right).$$

Замечание 1. Аналогично находят выборочное уравнение прямой линии регрессии X на Y вида

$$\bar{x}_y - \bar{x} = r_в \frac{\sigma_x}{\sigma_y}\left(y - \bar{y}\right),$$

где

$$r_в \frac{\sigma_x}{\sigma_y} = \rho_{xy}.$$

Замечание 2. Выборочный коэффициент корреляции имеет важное самостоятельное значение. Этот вопрос будет рассмотрено в ниже лекции.

2.1. Выборочный коэффициент корреляции

Как следует из предыдущего, выборочный коэффициент корреляции определяется равенством

$$r_{\text{в}} = \frac{\sum n_{xy} xy - n\bar{x}\,\bar{y}}{n\sigma_x \sigma_y}$$

где x, y - варианты (наблюдавшиеся значения) признаков X и Y;

n_{xy} - частота наблюдавшейся пары вариант (x, y);

n-объем выборки (сумма всех частот);

$\bar{x}$, $\bar{y}$ - выборочные средние;

σ_x, σ_y - выборочные среднеквадратические отклонения.

Ниже приведем свойства выборочного коэффициента корреляции из которых следует, что он служит для оценки тесноты линейной корреляционной зависимости.

1^0. Абсолютная величина выборочного коэффициента корреляции не превосходит единицы.

2^0. Если выборочный коэффициент корреляции равен нулю и выборочные линии регрессии прямые. Тогда X и Y не связаны линейной корреляционной зависимостью.

Замечание. Если выборочный коэффициент корреляции равен нулю, то признаки X и Y могут быть связаны нелинейной корреляционной или даже функциональной зависимостью.

3^0. Если $|r_{\text{в}}| = 1$, то наблюдаемые значения признаков связаны линейной функциональной зависимостью.

4^0. С возрастанием абсолютной величины выборочного коэффициента корреляции линейная корреляционная зависимость становится более тесной и при $|r_{\text{в}}| = 1$ переходит в функциональную зависимость.

Из приведенных свойств вытекает смысл $|r_в|$: выборочный коэффициент корреляции характеризует ***тесноту линейной связи*** между количественными признаками в выборке: чем ближе $|r_в|$ к 1, тем связь сильнее; чем ближе $|r_в|$ к 0, тем связь слабее.

2.2. Выборочное корреляционное отношение

Для оценки тесноты линейной корреляционной связи между признаками в выборке служит выборочный коэффициент корреляции. Для оценки тесноты нелинейной корреляционной связи вводят новые сводные характеристики:

η_{yx} - выборочное корреляционное отношение Y к X;

η_{xy} - выборочное корреляционное отношение X к .

Выборочным корреляционным отношением Y к X называют отношению

$$\eta_{yx} = \frac{\sigma_{\bar{y}_x}}{\sigma_y} \quad .$$

Здесь

$$\sigma_{\bar{y}_x} = \sqrt{\frac{\sum n_x \left(\bar{y}_x - \bar{y}\right)^2}{n}} \quad ; \qquad \sigma_y = \sqrt{\frac{\sum n_y \left(y - \bar{y}\right)^2}{n}} \, ,$$

где n – объем выборки (сумма всех частот);

n_x - частота значения x признака X;

n_y - частота значения y признака Y;

$\bar{y}$ - общая средняя признака Y;

$\bar{y}_x$ - условная средняя признака Y.

Аналогично определяется выборочное корреляционное отношение X к Y:

$$\eta_{xy} = \frac{\sigma_{\bar{x}_y}}{\sigma_x} \quad .$$

Пример. Найти η_{yx} по данным корреляционной таблицы.

X / Y	10	20	30	n_y
15	4	28	6	38
25	6	--	6	12
n_x	10	28	12	$n = 50$
$\overline{y_x}$	21	15	20	

Решение

Найдем общее среднюю

$$\bar{y} = \frac{\sum n_y y}{n} = \frac{38 \cdot 15 + 12 \cdot 25}{50} = 17{,}4.$$

Найдем

$$\sigma_y = \sqrt{\frac{\sum n_y \left(y - \bar{y}\right)^2}{n}} = \sqrt{\frac{38(15-17)^2 + 12(25-17{,}4)^2}{50}} = 4{,}27.$$

$$\sigma_{\bar{y}_x} = \sqrt{\frac{\sum n_x \left(\overline{y_x} - \bar{y}\right)^2}{n}} = \sqrt{\frac{10(21-17{,}4)^2 + 28(15-17{,}4)^2 + 12(20-16{,}4)^2}{50}} = 2{,}73.$$

Искомое корреляционное отношение

$$\eta_{yx} = \frac{\sigma_{\bar{y}_x}}{\sigma_y} = 0{,}64$$

Свойства выборочного корреляционного отношения.

Поскольку η_{yx} обладает теми свойствами ,что и η_{xy}, перечислим свойства только выборочного корреляционного отношения η_{yx} ,которое далее для упрощения записи будем обозначать через η и для простоты речи «корреляционным отношением».

1^0. Корреляционное отношение удовлетворяет двойному соотношению:

$$0 \le \eta \le 1.$$

2^0. Если η=0, то и признак Y с признаком X корреляционной зависимостью не связан и обратно.

3^0. Если η=1, то признак Yсвязан с признаком X функциональной зависимостью и обратно

4^0. $\eta \le |r_в|$.

5^0. Если $\eta = |r_в|$, то имеет место точная линейная корреляционная зависимость.

Во свойстве 2^0 выборочного корреляционного отношения было отмечено, при $\eta = 0$ признаки не связаны корреляционной зависимостью; при $\eta = 1$ имеет место функциональная зависимость.

В рассуждениях не делалось никаких допущений о форме корреляционной связи. Поэтому η служит мерой тесноты связи для любой, в том числе и линейной формы. В этом преимущество корреляционного отношения перед коэффициентом корреляции, который оценивает тесноту лишь линейной зависимости. Вместе с тем корреляционное отношение обладает недостатком: оно не позволяет судить, насколько близко расположены точки, найденные по данным наблюдений, к кривой определенного вида, например к параболе, гиперболе и т.д. Это объясняется тем, что при определении корреляционного отношения форма связи во внимание не принималась.

2.3. Криволинейные корреляции

Если график регрессии $\overline{y_x} = f(x)$ или $\overline{x_y} = \phi(y)$ изображается кривой линией , то корреляцию называют криволинейной.

Например, функции регрессии Y на X могут имеет вид: $\overline{y_x} = ax^2 + bx + c$ - параболическая корреляция, $\overline{y_x} = a + \frac{b}{x}$ - гиперболическая корреляция, $\overline{y_x} = ab^x$ - показательная корреляция и т. д.

Теория криволинейной корреляции решает те же задачи, что и теория линейной корреляции – установление формы и тесноты корреляционной связи.

Неизвестные параметры уравнения регрессии ищут методом наименьших квадратов. Для оценки тесноты криволинейной корреляции служат выборочные корреляционные отношения (лек.№15).

Рассмотрим параболическую корреляцию, предположив, что данные выборки позволяют считать, что имеет место именно такая корреляция . В этом случае выборочное уравнение регрессии Y на X имеет вид:

$$\bar{y}_x = Ax^2 + Bx + C\,, \qquad (1)$$

где A,B,C – неизвестные параметры.

Пользуясь МНК, получают систему линейных уравнений относительно неизвестных параметров (вывод опущен , поскольку он не содержит ничего нового сравнительно с п.1 лек.№14.)

$$\begin{cases} \left(\sum n_x x^4\right)A + \left(\sum n_x x^3\right)B + \left(\sum n_x x^2\right)C = \sum n_x \bar{y}_x x^2 \\ \left(\sum n_x x^3\right)A + \left(\sum n_x x^2\right)B + \left(\sum n_x x\right)C = \sum n_x \bar{y}_x x \\ \left(\sum n_x x^2\right)A + \left(\sum n_x x\right)B + nC = \sum n_x \bar{y}_x \end{cases} \qquad (2)$$

Найденные из этой системы параметры A,B,C подставляют в (1) в итоге получают искомое уравнение регрессии.

2.4. Понятие о множественной корреляции.

Если исследовать связь между несколькими признаками, то корреляцию называют множественной.

В простейшем случае число признаков равно трем, и связь между ними линейная:

$$z = ax + by + c$$

В этом случае возникают задачи:

1) найти по данным наблюдений выборочное уравнение связи вида

$$z = Ax + Bx + C \qquad (3)$$

т. е. требуется найти коэффициенты регрессии *A, B* и параметр *C.*;

2) оценить тесноту связи между *Z* и обоими признаками *X.Y;*

3) оценить тесноту связи между *Z* и *X* (при постоянном*Y*), между Z и.*Y* (при постоянном *X).*

Первая задача решается МНК, причем вместо уравнения (3) удобнее искать уравнение связи вида

$$z - \bar{z} = A\left(x - \bar{x}\right) + B\left(y - \bar{y}\right),$$

где

$$\boldsymbol{A} = \frac{\boldsymbol{r}_{xz} - \boldsymbol{r}_{yz}\boldsymbol{r}_{xy}}{1 - \boldsymbol{r}_{xy}^2} \cdot \frac{\sigma_z}{\sigma_x}; \qquad \boldsymbol{B} = \frac{\boldsymbol{r}_{yz} - \boldsymbol{r}_{xz}\boldsymbol{r}_{xy}}{1 - \boldsymbol{r}_{xy}^2} \cdot \frac{\sigma_z}{\sigma_y}.$$

Здесь r_{xz}, r_{yz}, r_{xy} – коэффициенты корреляции соответственно между признаками X и Z, Z и X, X и Y;

$\sigma_x, \sigma_y, \sigma_z$ -среднеквадратическое отклонения.

Теснота связи признака *Z* с признаками *X,Y* оценивается выборочным совокупным коэффициентом корреляции:

$$R = \sqrt{\frac{r_{xz}^2 - 2r_{xy}r_{xz}r_{yz} + r_{yz}^2}{1 - r_{xy}^2}};$$

причем $0 \le \boldsymbol{R} \le 1$.

Теснота связи между *Z* и *X* (при постоянном *Y*), между*Z* и *Y* (при постоянном *X*) оценивается соответственно частными выборочными коэффициентами корреляции:

$$r_{xz(y)} = \frac{r_{xz} - r_{xy}r_{yz}}{\sqrt{\left(1 - r_{xy}^2\right)\left(1 - r_{yz}^2\right)}}; \qquad r_{yz(x)} = \frac{r_{yz} - r_{xy}r_{xz}}{\sqrt{\left(1 - r_{xy}^2\right)\left(1 - r_{xz}^2\right)}}.$$

Эти коэффициенты имеют те же свойства и тот же смысл, что и обыкновенный выборочный коэффициент корреляции, т.е. служат для оценки линейной связи между признаками.

Вопросы для самопроверки

1. Дайте определение функциональной зависимости.

2. Дайте определение статистической зависимости.

3. Что называется условным средним?

4. Дайте определение корреляционной зависимости.

5. Дайте определения уравнения регрессии.

6. В чем состоит задача теории корреляции?

7. Что представляет собой метод наименьших квадратов (МНК)?

8. Что такое наблюдаемая ордината в МНК?

9. Напишите параметры выборочного уравнения прямой линии регрессии в случае, когда данные несгруппированы?

10. Поясните устройство корреляционной таблицы.

11. Напишите параметры выборочного уравнения прямой линии регрессии в случае, когда данные сгруппированы?

12. Как определяется выборочный коэффициент корреляции?

13. Как определяется выборочный коэффициент корреляции?

14. Приведите свойства выборочной коэффициента корреляции?

15. Что характеризирует коэффициент корреляции?

16. Напишите формулу для оценки коэффициента корреляции нормально распределенной генеральной совокупности при больших n?

17. Как оценивается тесноту нелинейной корреляционной связи?

18. Приведите свойства выборочного корреляционного отношения.

19. В чем проявляется недостаток корреляционного отношения?

20. Какую задачу решает теория криволинейной корреляции?

21. Какой метод используется для нахождения коэффициентов регрессионных уравнений в теории криволинейной корреляции ?

22. Что называется множественной корреляцией?

23. Как выясняется теснота связи меду признаками во множественной корреляции?