Задание

1. Ознакомиться с теоретический материалом темы, законспектировать.

2. Фотоотчет и сообщение присылать на электронную почту

С уважением, Хвастова Светлана Ивановна

!!! Если возникнут вопросы обращаться по телефону 0721389311 (ватсап). Электронная почта: xvsviv@rambler.ru

Функциональная, статистическая и корреляционная зависимости. Выборочное уравнение регрессии

§ 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 Y от одной случайной (или неслучайной) величины X, а затем от нескольких величин.

Две случайные величины могут быть связаны либо функциональной зависимостью, либо зависимостью другого рода, называемой статистической, либо быть независимыми.

Строгая функциональная зависимость реализуется редко, так как обе величины или одна из них подвержены еще действию случайных факторов, причем среди них могут быть и общие для обеих величин (под «общими» здесь подразумеваются такие факторы, которые воздействуют и на Y и на X). В этом случае возникает статистическая зависимость.

Например, если Y зависит от случайных факторов Z_1, Z_2, V_1, V_2, а X зависит от случайных факторов Z_1, Z_2, U_1, то между Y и X имеется статистическая зависимость, так как среди случайных факторов есть общие, а именно: Z_1, Z_2.

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 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.

§ 2. Условные средние

В качестве оценок условных математических ожиданий принимают условные средние, которые находят по данным наблюдений (по выборке).

Условным средним $\bar{y}_x$ называют среднее арифметическое наблюдавшихся значений Y, соответствующих X = x. Например, если при x_1=2 величина Y приняла значения y_1=5, y_2=6. y_3=10, то условное среднее $\bar{y}_{x_1}$ =(5+6+10)/3=7.

Аналогично определяется условное среднее $\bar{x}_y$.

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

§ 3. Выборочные уравнения регрессии

В гл. XIV, § 15 были введены уравнения регрессии Y на X и X на Y:

M(У|x)=f(x), M(X|y)=φ(y).

Условное математическое ожидание M (Y | x) является функцией от x, следовательно, его оценка, т. е. условное среднее $\bar{y}_x$, также функция от x; обозначив эту функцию через f* (x), получим уравнение

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

Это уравнение называют выборочным уравнением регрессии Y на X; функцию f*(x) называют выборочной регрессией Y на X, а ее график— выборочной линией регрессии Y на X. Аналогично уравнение

$$\bar{x}_y = \varphi^*(y)$$

называют выборочным уравнением регрессии X на Y; функцию φ*(y) называют выборочной регрессией X на Y, а ее график—выборочной линией регрессии X на Y.

Как найти по данным наблюдений параметры функций f*(x) и φ*(y), если вид их известен? Как оценить силу (тесноту) связи между величинами X и Y и установить, коррелированы ли эти величины? Ответы на эти вопросы изложены ниже.

§ 4. Отыскание параметров выборочного уравнения прямой линии среднеквадратичной регрессии по не сгруппированным данным

Пусть изучается система количественных признаков (X, Y). В результате n независимых опытов получены и пар чисел (x_1, y_1), (x_2, y_2), . . ., (x_n, y_n).

Найдем по данным наблюдений выборочное уравнение прямой линии среднеквадратичной регрессии (см. гл. XIV, § 20). Для определенности будем искать уравнение

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

регрессии Y на X.

Поскольку различные значения x признака X и соответствующие им значения y признака Y наблюдались по одному разу, то группировать данные нет необходимости. Также нет надобности использовать понятие условной средней, поэтому искомое уравнение можно записать так:

$$У = kx + b.$$

Угловой коэффициент прямой линии регрессии Y на X называют выборочным коэффициентом регрессии Y на X и обозначают через ρ_{yx}; он является оценкой коэффициента регрессии β.

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

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

Подберем параметры ρ_{yx} и b так, чтобы точки (x_1, y_1), (x_2, y_2), . . ., (x_n, y_n), построенные по данным наблюдений, на плоскости xOy лежали как можно ближе к прямой (*). Уточним смысл этого требования. Назовем отклонением разность

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

где Y_i – вычисленная по уравнению (*) ордината, соответствующая наблюдаемому значению x_i; y_i – наблюдаемая ордината, соответствующая 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):

$$F(\rho,b) = \sum_{i=1}^{n}(Y_i - y_i)^2,$$

или

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

Для отыскания минимума приравняем нулю соответствующие частные производные:

$$\frac{dF}{d\rho} = 2\sum_{i=1}^{n}(\rho x_i + b - y_i)x_i = 0;$$

$$\frac{dF}{db} = 2\sum_{i=1}^{n}(\rho x_i + b - y_i) = 0.$$

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

$$\left(\sum x^2\right)\rho + \left(\sum x\right)b = \sum xy; \quad \left(\sum x\right)\rho + nb = \sum y. \qquad (**)$$

Решив эту систему, найдем искомые параметры:

$$\rho_{yx} = \left(n \cdot \sum xy - \sum x \cdot \sum y\right) / \left(n \sum x^2 - \left(\sum x\right)^2\right);$$

$$b = \left(\sum x^2 \cdot \sum y - \sum x \cdot \sum xy\right) / \left(n \sum x^2 - \left(\sum x\right)^2\right). \qquad (***)$$

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

$$\bar{x}_y = \rho_{xy}\, x + C,$$

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 1,00 1,50 3,00 4,50 5,00

y 1,25 1,40 1,50 1,75 2,25

Решение. Составим расчетную табл. 11.

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

ρ_{xy} = (5·26,975— 15·8,15)/(5·57,5—15^2) ==0,202;

Ь= (57,5·8,15—15·26,975)/62,5= 1,024.

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

Таблица 11

x_i	y_i	x^2_i	$x_i y_i$
1,00	1,25	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	7,875
5,00	2.25	25,00	11,250
$\sum x_i = 15$	$\sum y_i = 8{,}15$	$\sum x_i^2 = 57{,}50$	$\sum x_i y_i = 26{,}975$

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 y_i найдем отклонения Y_i—y_i. Результаты вычислений приведены в табл. 12.

Таблица 12

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	1,933	1,75	0,183
5,00	2,034	2,25	—0,216

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