

ОСНОВЫ НАУЧНЫХ ИССЛЕДОВАНИЙ

Задание для ГП-17

22.12.2020

Ссылка для подключения в сервисе BigBlueButton:

<http://disrm1.zabgu.ru/b/dcd-g6q-a7r>

Начало подключения – 11 час.50 мин.

1) Ознакомиться с практическим занятием №9 «Метод наименьших квадратов. Корреляционный, регрессионный, дисперсионный анализы».

Сделать письменный конспект практического занятия.

Практическое занятие №9

Корреляционный, регрессионный, дисперсионный анализы. Метод наименьших квадратов.

Регрессионный анализ

Под регрессионным анализом понимают исследование закономерностей связи между явлениями (процессами), которые зависят от многих, иногда неизвестных, факторов. Часто между переменными x и y существует связь, но не вполне определенная, при которой одному значению x соответствует несколько значений (совокупность) y . В таких случаях связь называют регрессионной. Таким образом, функция $y = f(x)$ является регрессионной (корреляционной), если каждому значению аргумента соответствует статистический ряд распределения y . Следовательно, регрессионные зависимости характеризуются вероятностными или стохастическими связями. Поэтому установление регрессионных зависимостей между величинами y и x возможно лишь тогда, когда выполнимы статистические измерения.

Статистические зависимости описываются математическими моделями процесса, т.е. регрессионными выражениями, связывающими независимые значения x (факторы) с зависимой переменной y (результативный признак, функция цели, отклик). Модель по возможности должна быть простой и адекватной. Например, модуль упругости материала E зависит от его плотности ρ так, что с возрастанием плотности модуль упругости материала увеличивается. Но выявить эту закономерность можно только при наличии большого количества измерений, так как при исследованиях каждой отдельной парной связи в зависимости $E=f(\rho)$ наблюдаются большие отклонения.

Суть регрессионного анализа сводится к установлению уравнения регрессии, т.е. вида кривой между случайными величинами (аргументами x и функцией y), оценке тесноты связей между ними, достоверности и адекватности результатов измерений.

Чтобы предварительно определить наличие такой связи между x и y , наносят точки на график и строят так называемое корреляционное поле (рис. 1).

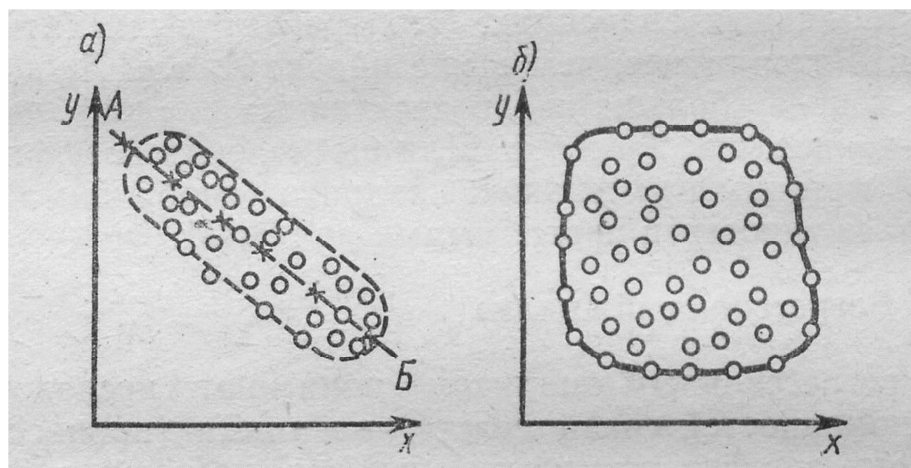


Рис. 1. Корреляционное поле

По тесноте группирования точек вокруг прямой или кривой линии, по наклону линии можно визуально судить о наличии корреляционной связи. Так, из рис. 1, *a* видно, что экспериментальные данные имеют определенную связь между x и y , а измерения, приведенные на рис. 1, *б*, такой связи не показывают.

Корреляционное поле характеризует вид связи между x и y . По форме поля можно ориентировочно судить о форме графика, характеризующего прямолинейную или криволинейную зависимости. Даже для вполне выраженной формы корреляционного поля вследствие статистического характера связи исследуемого явления одно значение x может иметь несколько значений y . Если на корреляционном поле осреднить точки, т. е. для каждого значения x_i определить $\bar{y}_i$ и соединить точки $\bar{y}_i$, то можно будет получить ломаную линию, называемую *экспериментальной регрессионной зависимостью (линией)*. Наличие ломаной линии объясняется погрешностями измерений, недостаточным количеством измерений, физической сущностью исследуемого явления и др. Если на корреляционном поле провести плавную линию между $\bar{y}_i$, которая равноудалена от них, то получится новая теоретическая регрессионная зависимость — линия *АБ* (см. рис. 1, *a*).

Различают *однофакторные (парные) и многофакторные регрессионные зависимости*. Парная регрессия при парной зависимости может быть аппроксимирована прямой линией, параболой, гиперболой, логарифмической, степенной или показательной функцией, полиномом и др. Двухфакторное поле можно аппроксимировать плоскостью, параболоидом второго порядка, гиперболоидом. Для переменных факторов связь может быть установлена с помощью n -мерного пространства уравнениями второго порядка:

$$y = b_0 + \sum_1^n b_i x_i + \sum_j^n b_{ij} x_i x_j + \sum_1^n b_{ij} x_i^2 \quad (1)$$

где y — функция цели (отклика) многофакторных переменных; x_i — независимые факторы; b_i — коэффициенты регрессии, характеризующие влияние фактора x_i на функцию цели; b_{ij} — коэффициенты, характеризующие двойное влияние факторов x_i и x_j на функцию цели.

При построении теоретической регрессионной зависимости оптимальной является такая функция, в которой соблюдаются условия наименьших квадратов $\sum (y_i - y)^2 = \min$, где y_i — фактические ординаты поля; y — среднее значение ординаты с абсциссой x . Суть метода заключается в том, что сумма квадратов отклонений фактических значений результативного признака (y_i) от значений, найденных по уравнению регрессии (y), должна быть наименьшей. Поле корреляции аппроксимируется уравнением прямой $y = a + bx$. (Аппроксимация, или приближение — научный метод, состоящий в замене одних объектов другими, в каком-то смысле близкими к исходным, но более простыми). Нахождение уравнения регрессии сводится к определению его параметров. Определение параметров производится, как правило, по методу наименьших квадратов. При этом кривая AB (см. рис. 1, а) наилучшим образом выравнивает значения постоянных коэффициентов a и b , т. е. коэффициентов уравнения регрессии.

Их вычисляют по выражениям

$$b = (n \sum xy - \sum x \sum y) \div n \sum x^2 - (\sum x)^2 \quad (2)$$

$$a = y - bx = \frac{\sum y}{n} - (b \sum x)/n \quad (3)$$

Критерием близости корреляционной зависимости между x и y к линейной функциональной зависимости является коэффициент парной или просто *коэффициент корреляции* r , показывающий степень тесноты связи x и y определяемый отношением

$$r = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{[n \sum x_i^2 - (\sum x_i)^2] [n \sum y_i^2 - (\sum y_i)^2]}} \quad (4)$$

где n — число измерений. Значение коэффициента корреляции всегда меньше единицы. При $r=1,0$ x и y связаны функциональной связью (в данном случае линейной), т. е. каждому значению x соответствует только одно значение y . Если $r < 1$, то линейной связи не существует. При $r=0$ линейная корреляционная связь между x и y отсутствует, но может существовать нелинейная регрессия. Обычно считают тесноту связи удовлетворительной при $r \geq 0,5$; хорошей при $r=0,8 \dots 0,85$. Для определения процента разброса (изменчивости) искомой функции y относительно ее среднего значения, определяемого изменчивостью фактора x , вычисляют коэффициент детерминации

$$r_d = r^2 \quad (5)$$

Уравнение регрессии прямой можно представить выражением

$$y = \bar{y} + r \frac{\sigma_y}{\sigma_x} (x - \bar{x}) \quad (6)$$

Пусть, например, имеется статистический ряд парных измерений:

1 2 3 4 5 6 7 8 9 10

8 11 14 16 21 26 27 32 34 41

по которому можно найти уравнение прямолинейной регрессии, оценить тесноту связей и оценить степень достоверности. Расчет целесообразно вести в табличной форме (табл. 10.9).

В табл. 10.10 приведена сходимость экспериментальной и теоретической регрессии.

$$\begin{aligned} \bar{x} &= \frac{55}{10} = 5,5; & \bar{y} &= \frac{230}{10} = 23; & \sigma_x &= \\ &= \frac{82,50}{10} = 8,25; & \sigma_y &= \frac{1054}{10} = 105,4. \end{aligned}$$

Таблица 10.9

Расчет уравнения регрессии

x	y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})^2$	$(y - \bar{y})^2$	x^2	y^2	xy	$(x - \bar{x})(y - \bar{y})$
1	8	-4,5	-15	20,25	225	1	64	8	67,5
2	11	-3,5	-12	12,25	144	4	121	22	42,0
3	14	-2,5	-9	6,25	81	9	196	42	22,5
4	16	-1,5	-7	2,25	49	16	256	64	10,5
5	21	-0,5	-2	0,25	4	25	441	105	1,0
6	26	0,5	+3	0,25	9	36	676	156	1,5
7	27	1,5	+4	2,25	16	49	729	189	6,0
8	32	2,5	+9	6,25	81	64	1024	256	22,5
9	34	3,5	+11	12,25	121	81	1156	306	31,5
10	41	4,5	+18	20,25	324	100	1681	410	81,0
55	230	—	—	82,50	1054	385	6344	1558	286,0

Таблица 10.10

Сходимость экспериментальной и теоретической регрессии

y	8	11	14	16	21	26	27	32	34	41
y_0	7,1	10,6	14,2	17,7	21,8	24,8	28,3	31,9	35,4	39,0

Коэффициент корреляции согласно (4)

$$r = \frac{10 * 1558 - 55 * 230}{(10 * 385 - 55^2) * (10 * 6344 - 230^2)} = -0.99$$

Из (2) и (3)

$$b = \frac{10 * 1558 - 55 * 230}{10 * 385 - 55^2} = 3,55$$

$$a = \frac{230}{10} - 3,55 \frac{55}{10} = 3,48$$

Уравнение регрессии имеет вид

$$y = 3,48 + 3,55x$$

Как видно из расчетов, сходимость оказалась хорошей.

Коэффициент детерминации, найденный по формуле (5), составляет $r_d^2 = 0,99^2 = 0,98$, что означает, что 98% разброса определяется изменчивостью x , а 2% — другими причинами, т. е. изменчивость функции y почти полностью характеризуется разбросом (природой) фактора x .

На практике часто возникает потребность в установлении связи между y и многими параметрами $x_1, \dots, x_n$ на основе многофакторной регрессии.

Многофакторные теоретические регрессии аппроксимируются полиномами первого или второго (1) порядка.

Полином первого порядка

$$y = A_0 + A_1x + A_2x^2 + A_3x^3 + \dots + A_nx^n$$

Математические модели характеризуют стохастический процесс изучаемого явления, уравнение регрессии определяет систематическую, а ошибки разброса — случайную составляющие.

Теоретическую модель множественной регрессии можно получить методами математического планирования, т. е. активным экспериментом, а также пассивным, когда точки факторного пространства выбираются в процессе эксперимента произвольно.

Дисперсия - это среднее арифметическое квадратов отклонений значений переменной от её среднего значения.

Дисперсия как статистическая величина характеризует, насколько частные значения отклоняются от средней величины в данной выборке. Чем больше дисперсия, тем больше отклонения или разброс данных.

$$D = \frac{1}{n} \cdot \sum_{i=1}^n (X_i - \bar{X})^2$$

В статистическом понимании дисперсия

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

есть среднее арифметическое из квадратов отклонений величин x_i от их среднего арифметического

$$\bar{x} = (x_1 + x_2 + \dots + x_n) / n.$$

где s^2 - выборочная дисперсия, или просто дисперсия;

$(X_i - \bar{X})$ - выражение, означающее, что для всех x , от первого до последнего в данной выборке необходимо вычислить разности между частными и средними значениями, возвести эти разности в квадрат и просуммировать;

n - количество значений, по которым вычисляется дисперсия.