

ПАРНАЯ КОРРЕЛЯЦИЯ И РЕГРЕССИЯ

Цель работы: исследовать взаимосвязь двух признаков методами корреляции и регрессии.

Основные теоретические сведения

Пусть предварительно установлено, что изменчивость некоторого признака y зависит от изменчивости другого признака x .

Установлено, что признаки подчиняются нормальному закону распределения. Пусть вид аналитической зависимости заранее не определен.

Тогда цель анализа — установить вид зависимости признаков y и x , определить и оценить тесноту связи этих признаков и определить наилучшую форму регрессионного уравнения.

Аналитическую модель связи будем определять в виде: прямой линейной зависимости

$$\hat{y} = a_0 + a_1 x; \quad (1)$$

обратной линейной зависимости

$$\hat{x} = b_0 + b_1 y; \quad (2)$$

параболы

$$\hat{y} = a_0 + a_1 x + a_2 x^2; \quad (3)$$

гиперболы

$$\hat{y} = a_0 + \frac{a_1}{x}. \quad (4)$$

Методы, которыми можно представить отношения в виде прямой, применимы и в случае кривых различного типа.

Типов кривых — бесконечное множество, однако в практике используются кривые нескольких основных видов, такие как параболы второго, третьего порядка, гиперболы.

Для получения количественных значений параметров этих уравнений применяется метод наименьших квадратов.

Однако вид уравнения заранее невозможно определить. Поэтому определяется несколько видов зависимостей и из них выбирается, согласно критериям, наилучший, который и принимается в качестве решения.

Для парных зависимостей исследуется несколько видов уравнений и сравниваются рассеяния результатов около линии регрессии (применяя остаточные дисперсии и критерий Фишера).

Выбранный вид уравнения регрессии служит указанием к выбору вида регрессии для множественных зависимостей. Принято определять параметры множественной регрессии, принимая порядок уравнения не ниже самого высокого из входящих парных зависимостей.

Среди самых простых видов — параболическая форма:

$$\hat{y} = a_0 + a_1x + a_2x^2.$$

Другие формы зависимости:

— показательная;

— гиперболическая;

— экспоненциальная;

— обратно-экспоненциальная;

— экспоненциально-показательная;

— обратно-экспоненциально-показательная.

Исследователь, зная основные черты явления, может выбрать те или другие виды зависимости для расчета.

Остальные рассуждения: оценка тесноты связи, определение аргумента и функции, выбор вида уравнения

— все проводится в аналогии с линейной формой и потому не требует специального рассмотрения.

Весьма часто можно усмотреть известную связь между вариациями по различным признакам. В простейшем случае такая связь однозначна. Но это бывает редко. Если составить так называемую **корреляционную решетку**, то видна некоторая размазанность корреляции.

Собственно корреляционная решетка (табл. 6.1) включает частоты одного признака в конкретном интервале, распределенные по всем интервалам другого признака. Заполнение корреляционной решетки рекомендуется сделать следующим образом: выписать из табл. 6.2 все значения признака x , входящие в первый интервал, и соответствующие значения y . Затем распределить количество значений x и y по интервалам значений. Таким образом определяются условные частоты первой строки.

Другими словами, m_{11} — число значений y из первого интервала, которым соответствуют значения x из первого интервала; m_{12} — число значений y из первого интервала, которым соответствуют значения x из второго интервала и т. д.

Таблица 6.1

$\begin{matrix} \rightarrow j \\ i \end{matrix}$	$(x_H - x_b)$	$(x_H - x_b)_1$	$(x_H - x_b)_2$	...	$(x_H - x_b)_k$	m	$\tilde{x}_i$
$(y_H - y_b)$	$\tilde{y}_1$	$\tilde{x}_1$	$\tilde{x}_2$	...	$\tilde{x}_k$		
$(y_H - y_b)$	$\tilde{y}_2$	m_{11}		...			
$(y_H - y_b)$	$\tilde{y}_2$	m_{12}		...			
$\vdots$	$\vdots$	$\vdots$	$\vdots$				
$(y_H - y_b)$	$\tilde{y}_k$						
	m	$\sum m_{1k}$	$\sum m_{2k}$		$\sum m_{kk}$		
	$\tilde{y}_n$	$\tilde{y}_{11}$	$\tilde{y}_{12}$				

Аналогично следует поступить с другими интервалами. Разумеется, если просуммировать по строке значения условных частот, получим число значений y , входящих в соответствующий интервал по y , если просуммировать значения по столбцу, получим число значений x , входящих в соответствующий интервал по x , если просуммировать все значения условных частот в корреляционной решетке, получим общее число наблюдений N .

Таблица 6.2 представляет обобщенные результатов рассмотрения признаков x и y , т. е. — корреляционную решетку.

Так, первый столбец таблицы — столбец y — среднеинтервальные значения y , а верхняя строка таблицы — среднеинтервальные значения x .

На пересечениях столбцов и строк — частоты, соответствующие данным интервалам x и y .

Таблица 6.2.

$y \backslash x$	1500	1900	2300	2700	3100	3500
44						1
38			1	1	1	
32						
26		1	1			
20	4	5	10	3	1	
14	2	4		2	1	

Установление корреляции между признаками не дает оснований считать эти связи причинно-следственными. Вполне может быть, что эти признаки зависят еще от каких-то признаков.

Наглядное представление (рис. 6.1) можно получить, построив **корреляционное поле** (точечная диаграмма). Пусть, для примера имеем:

x 5,0	5,5	5,5	6,0	6,0	7,0	7,0	8,0
y 10,0	10,5	11,0	10,5	11,0	11,0	12,0	12,0

Вытянутость корреляционного поля по диагонали свидетельствует о несомненном наличии корреляции между признаками.

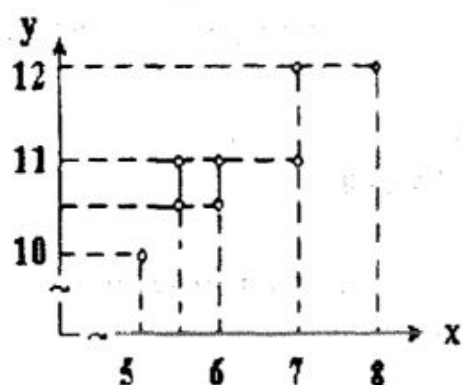


Рис. 6.1

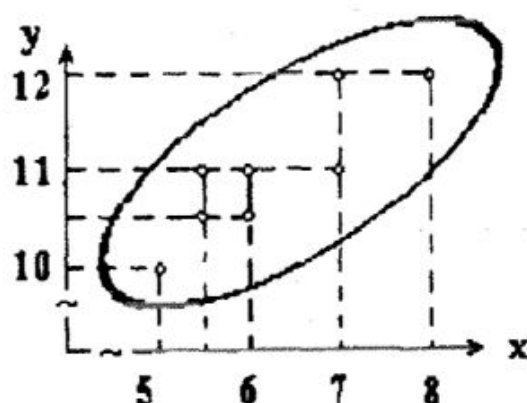


Рис. 6.2

Если число вариантов велико, то корреляционное поле имеет вид более или менее правильного эллипса со сгущением точек в центре и сравнительно редким их расположением на периферии. Такой эллипс носит название **эллипс рассеяния** (рис. 6.2). Отклонение осей эллипса от координатных направлений указывает на наличие корреляции. Вытянутость говорит только о масштабах по осям.

Обычно на одно значение признака приходится несколько значений другого признака. Так, на одно значение суточной добычи приходится несколько значений ширины выработки. При этом рассчитывается условное или групповое среднее, так, $\bar{y}_x$ — среднее значение y при условии, что x — заданная величина.

Тогда о корреляции можно говорить только в случае, когда изменения $\bar{y}_x$ при переходе от x_i к x носят систематический характер. Получим условные средние $\bar{y}_{x_0}^0$:

для

$x = 5,0$	$y = 10,0$	$\bar{y}_{x=5,0} = 10,0$
$x = 5,5$	$y = \begin{cases} 10,5 \\ 11,0 \end{cases}$	$\bar{y}_{x=5,5} = \frac{10,5 + 11,0}{2} = 10,75$
$x = 6,0$	$y = \begin{cases} 10,5 \\ 11,0 \end{cases}$	$\bar{y}_{x=6} = \frac{10,5 + 11,0}{2} = 10,75$
$x = 7,0$	$y = \begin{cases} 11,0 \\ 12,0 \end{cases}$	$\bar{y}_{x=7,0} = \frac{11,0 + 12,0}{2} = 11,5$
$x = 8,0$	$y = 12,0$	$\bar{y}_{x=8} = 12,0.$

А теперь рассчитаем условные средние $\bar{x}_y$:
для

$y = 10,0$	$x = \begin{cases} 5,0 \\ 5,5 \end{cases}$	$\bar{x}_{y=10,0} = \frac{5,0 + 5,5}{2} = 5,25$
$y = 10,5$	$x = \begin{cases} 5,5 \\ 6,0 \end{cases}$	$\bar{x}_{y=10,5} = \frac{5,5 + 6,0}{2} = 5,75$
$y = 11,0$	$x = \begin{cases} 6,0 \\ 7,0 \end{cases}$	$\bar{x}_{y=11,0} = \frac{5,5 + 6,0 + 7,0}{2} = 6,17$
$y = 12,0$	$x = \begin{cases} 7,0 \\ 8,0 \end{cases}$	$\bar{x}_{y=12,0} = \frac{7,0 + 8,0}{2} = 7,5.$

Построим зависимости $(\bar{y}_x, x)$ и $(y, \bar{x}_y)$, обозначим их соответственно через (*) и (0) и назовем их эмпирическими линиями регрессии (рис. 6.3).

Линии не совпадают между собой, что является результатом размазанности корреляции. В общем случае

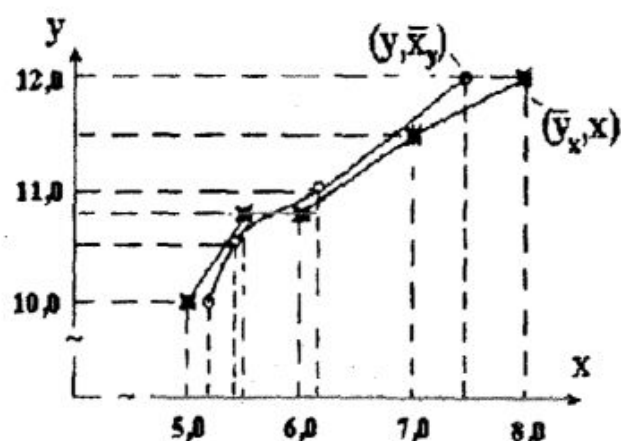


Рис 6.3

линии — ломаные, но в простейшем случае — средние значения одного признака зависят от значений другого признака линейно. Такой случай называют линейной регрессией. Для простоты рассмотрим линейный случай.

Точки, изображающие зависимость $\bar{y}_x$ от x и $\bar{x}_y$ от y , никогда не ложатся на одну прямую. Поэтому речь может идти только о том, чтобы найти такую прямую, которая проходила бы наиболее близко ко всем точкам. Смысл «близости» может быть разным:

а) наилучшая прямая та, при которой max эмпирического отклонения от расчетного — наименьший. Но тогда наиболее отклоняющаяся точка окажет большое влияние на определение прямой;

б) наилучшая прямая та, при которой площадь между ломаной эмпирических значений и прямой — наименьшая. Такую считать трудно;

в) в большинстве случаев целесообразен критерий, при котором требуется минимизировать сумму квадратов отклонений эмпирических точек от прямой (метод наименьших квадратов). При этом отклоняющаяся точка не имеет решающего значения.

Для практических целей могут представлять интерес оба уравнения регрессии. При размазанности корреляции коэффициенты уравнений не являются обратными.

$r_{xy} = r_{yx}$, если не размазаны данные.

Важной задачей теории корреляции является построение численного параметра, который давал бы количественное выражение степени или силы корреляции между признаками.

Корреляция тем более сильна, чем теснее точки корреляционного поля группируются около линии регрессии. И если корреляция полная (т. е. неучитываемых влияний нет), то имеем функциональную зависимость $r_{xy} = r_{yx} = 1$.

Если корреляция отсутствует (т. е. y в общем не зависит от x), то $r_{xy} = r_{yx} = 0$.

Корреляционную связь можно считать реальной, если полученный коэффициент корреляции значительно отличается от нуля. С этой целью используют таблицы r -распределения.

Если вычисленный коэффициент корреляции превосходит табличное значение для выбранного уровня значимости при числе степеней свободы $f = N - 2$, где N — число испытаний, то его можно считать значимо отличающимся от нуля.

Для выполнения анализа необходимо, чтобы коэффициенты корреляции были безусловно значимы.

Если $r_{yx} = 0$, то это только означает, что не может существовать линейная корреляционная связь, а криволинейная вполне может.