Задание

1. Изучить теоретический материал темы, законспектировать пример решения задачи, подготовить 5-7 контрольных вопросов.

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

Отыскание параметров выборочного уравнения прямой линии регрессии по сгруппированным данным

Ранее для определения параметров уравнения прямой линии регрессии Y на X была получена система уравнений

$$\left.\begin{aligned}&(\sum x^2)\rho_{yx}+(\sum x)b=\sum xy,\\&(\sum x)\rho_{yx}+nb=\sum y.\end{aligned}\right\}\qquad(1)$$

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 (1) так, чтобы она отражала данные корреляционной таблицы. Воспользуемся тождествами:

$$\sum x=n\bar{x}\ \text{(следствие из } \bar{x}=\sum x/n\text{)};$$

$$\sum y=n\bar{y}\ \text{(следствие из } \bar{y}=\sum y/n\text{)};$$

$$\sum x^2=n\bar{x}^2\ \text{(следствие из } \bar{x}^2=\sum x^2/n\text{)},$$

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

Подставив правые части тождеств в систему (1) и сократив обе части второго уравнения на n, получим

$$\left.\begin{aligned}&(n\bar{x}^2)\rho_{yx}+(n\bar{x})b=\sum n_{xy}xy,\\&(\bar{x})\rho_{yx}+b=y.\end{aligned}\right\}\qquad\mathbf{(2)}$$

Решив эту систему, найдем параметры ρ_{xy} и b и, следовательно, искомое уравнение

$$\bar{y}_x=\rho_{yx}x+b.\qquad(3)$$

Однако более целесообразно, введя новую величину – выборочный коэффициент корреляции, написать уравнение регрессии в ином виде. Сделаем это. Найдем b из второго уравнения (2):

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

Подставив правую часть этого равенства в уравнение (3), получим

$$\bar{y}_x - \bar{y} = \rho_{yx}(x - \bar{x}). \qquad (4)$$

Найдем из системы (1) коэффициент регрессии, учитывая, что $\bar{x}^2 - (\bar{x})^2 = \tilde{\sigma}_x^2$

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

Умножим обе части равенства на дробь $\tilde{\sigma}_x / \tilde{\sigma}_y$

$$\rho_{yx} \cdot \frac{\tilde{\sigma}_x}{\tilde{\sigma}_y} = \frac{\sum n_{xy}xy - n\bar{x}\bar{y}}{n\tilde{\sigma}_x\tilde{\sigma}_y}. \qquad (5)$$

Обозначим правую часть равенства через r_B и назовем ее выборочным коэффициентом корреляции

$$r_B = \frac{\sum n_{xy}xy - n\bar{x}\bar{y}}{n\tilde{\sigma}_x\tilde{\sigma}_y}.$$

Подставим r_B в (5):

$$\rho_{yx}\tilde{\sigma}_x / \tilde{\sigma}_y = r_B.$$

Отсюда

$$\rho_{yx} = r_B \frac{\tilde{\sigma}_y}{\tilde{\sigma}_x}.$$

Подставив правую часть этого равенства в (4), окончательно получим выборочное уравнение прямой линии регрессии Y на X вида

$$\bar{y}_x - \bar{y} = r_B \frac{\tilde{\sigma}_y}{\tilde{\sigma}_x}(x - \bar{x}).$$

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

$$x_y - \bar{x} = r_B \frac{\tilde{\sigma}_x}{\tilde{\sigma}_y}(y - \bar{y}),$$

где $r_B\tilde{\sigma}_x / \tilde{\sigma}_y = \rho_{xy}$.

Замечание 2. Уравнения выборочной прямой регрессии можно записать в более симметричной форме:

$$\frac{\bar{y}_x - \bar{y}}{\tilde{\sigma}_y} = r_B \frac{x - \bar{x}}{\tilde{\sigma}_x}, \frac{\bar{x}_y - \bar{x}}{\tilde{\sigma}_x} = r_B \frac{y - \bar{y}}{\tilde{\sigma}_y}.$$

Замечание 3. Можно показать, используя метод моментов, что выборочный коэффициент корреляции является оценкой теоретического коэффициента корреляции

Выборочный коэффициент корреляции

Выборочный коэффициент корреляции определяется равенством

$$r_B = \frac{\sum n_{xy} xy - n\bar{x}\bar{y}}{n\tilde{\sigma}_x \tilde{\sigma}_y},$$

где x, y – варианты (наблюдавшиеся значения) признаков X и Y; n_{xy} – частота пары вариант (x, y); n – объем выборки (сумма всех частот); $\tilde{\sigma}_x$, $\tilde{\sigma}_y$ – выборочные среднеквадратические отклонения; $\bar{x}$, $\bar{y}$ – выборочные средние.

Известно, что если величины Y и X независимы, то коэффициент корреляции $r = 0$; если $.r = \pm 1$, то Y и X связаны линейной функциональной зависимостью. Итак, коэффициент корреляции r измеряет силу (тесноту) линейной связи между Y и X в соответствии с приведенной ниже таблицей.

Значение r	0 – 0,1	0,1 – 0,3	0,3 – 0,5	0,5 – 0,7	0,7 – 0,9	0,9 – 0,99	1
Теснота линейной связи	нет	слабая	умерен–ная	замет–ная	высо–кая	очень высокая	функ–ciона–льная

Выборочный коэффициент корреляции r_B является оценкой коэффициента корреляции r генеральной совокупности и поэтому также служит для измерения линейной связи между величинами – количественными признаками Y и X. Допустим, что выборочный коэффициент корреляции, найденный по выборке, оказался отличным от нуля. Так как выборка отобрана случайно, то отсюда еще нельзя заключить, что коэффициент корреляции генеральной совокупности также отличен от нуля.

Если выборка имеет достаточно большой объем и хорошо представляет генеральную совокупность (репрезентативна), то заключение о тесноте линейной зависимости между признаками, полученное по данным выборки, в известной степени может быть распространено и на генеральную совокупность. Например, для оценки коэффициента корреляции $r_г$ нормально распределенной генеральной совокупности (при $n \geq 50$) можно воспользоваться формулой

$$r_B - 3\frac{1 - r_B^2}{\sqrt{n}} \leq r_Г \leq r_B + 33\frac{1 + r_B^2}{\sqrt{n}}.$$

Методика нахождения выборочного уравнения прямой линии регрессии

Пусть требуется по данным корреляционной таблицы найти выборочное уравнение прямой линии регрессии Y на X.

Вычислим сначала выборочный коэффициент корреляции. Можно значительно упростить расчет, если к условным вариантам (при этом величина r_B не изменится)

$$u_i = (x_i - C_1)/h_1 \text{ и } v_j = (y_j - C_2)/h_2$$

В этом случае выборочный коэффициент корреляции вычисляют по формуле

$$r_B = (\sum n_{uv} uv - n\bar{u}\bar{v})/(n\bar{\sigma}_u \bar{\sigma}_v).$$

Величины $\bar{u}$, $\bar{v}$, $\bar{\sigma}_u$ и $\bar{\sigma}_v$ можно найти методом произведения, а при малом числе данных – непосредственно исходя из определенных этих величин. Остается указать способ вычисления $\sum n_{uv}uv$, где n_{uv} – частота пары условных вариант (u, v).

Можно доказать, что справедливы формулы:

$$\sum n_{uv}uv = \sum vU \text{, где } U = \sum n_{uv}u,$$

$$\sum n_{uv}uv = \sum uV \text{, где } V = \sum n_{uv}v.$$

Для контроля целесообразно вычислить расчеты по обеим формулам и сравнить результаты; их совпадения свидетельствуют о правильности вычислений.

Напишем искомое уравнения в общем виде:

$$\bar{y}_x - \bar{y} = r_B \frac{\tilde{\sigma}_y}{\tilde{\sigma}_x}(x - \bar{x}). \qquad (*)$$

Поскольку при нахождении r_B уже вычислены $\bar{u}$, $\bar{v}$, $\tilde{\sigma}_u$, $\tilde{\sigma}_v$, то целесообразно пользоваться формулами:

$\tilde{\sigma}_x = h_1\tilde{\sigma}_u$, $\tilde{\sigma}_y = h_2\tilde{\sigma}_v$, $\bar{x} = \bar{u}h_1 + c_1$, $\bar{y} = \bar{v}h_2 + c_2$.

Пример. Найти выборочное уравнение прямой линии регрессии *Y* на *X* по данным корреляционной таблицы

Y	X						n_y
	10	20	30	40	50	60	
15	5	7	–	–	–	–	12
25	–	20	23	–	–	–	43
35	–	–	30	47	2	–	79
45	–	–	10	11	20	6	47
55	–	–	–	9	7	3	19
n_x	5	27	63	67	29	9	*n*=200

Решение. Перейдем к условным вариантам: $u_i = (x_i - C_1)/h_1 = (x_i - 40)/10$ (в качестве ложного нуля C_1 взята варианта *x*=40, расположенная примерно в середине вариационного ряда; шаг h_1 равен разности между двумя соседними вариантами: 20–10=10) и $v_j = (y_j - C_2)/h_2 = (y_j - 35)/10$ (в качестве ложного нуля C_2 взята варианта *y*=35, расположенная в середине вариационного ряда; шаг h_2 равен разности между двумя соседними вариантами: 25–15=10).

Составим корреляционную таблицу в условных вариантах. Практически этоо делают так: в первом столбце вместо ложного нуля C_2 (варианты 35) пишут 0; над нулем последовательно записывают –1, –2; под нулем пишут 1, 2. В первой строке вместо ложного нуля C_1 (варианты 40) пишут 0; слева от нуля последовательно записывают –1, –2, –3; справа от нуля пишут 1, 2. все остальные данные переписываются из первоначальной корреляционной таблицы. В итоге получим корреляционную таблицу в условных вариантах.

v			*u*	

	–3	–2	–1	0	1	2	n_v
–2	5	7	–	–	–	–	12
–1	–	20	23	–	–	–	43
0	–	–	30	47	2	–	79
1	–	–	10	11	20	6	47
2	–	–	–	9	7	3	19
n_u	5	27	63	67	29	9	n=200

Теперь для вычисления искомой суммы $\sum n_{uv}uv$ составим расчетную таблицу. Пояснения к составлению таблицы.

1. В каждой клетке, в которой частота $n_{uv} \neq 0$ записывают в правом верхнем углу произведение частоты n_{uv} на варианту u. Например, в правых верхних углах клеток первой строки записаны произведения: 5*(–3)=–15; 7*(–2)=14.
2. Складывают все числа, помещенные в правых верхних углах клеток одной строки и их суму записывают в клетку этой же строки столбца U. Например, для первой строки U =–15+(–14)=–29.
3. Умножают варианту v на U и полученное произведение записывают в последнюю клетку той же строки, то естьв клетку столбца vU. Например, в первой строке таблицы v=–2, U=–29; следовательно, vU =(–2)*(–29)=58.
4. Наконец, сложив все числа столбца vU, получаем сумму $\sum_v vU$, которая равна искомой сумме $\sum n_{uv}uv$. Например, для таблицы имеем $\sum_v vU = 169$; следовательно, искомая сумма $\sum n_{uv}uv = 169$.

v	u						$U = \sum n_{uv}u$	vU
	–3	–2	–1	0	1	2		
–2	–15 5 –10	–14 7 –14	–	–	–	–	–29	58
–1	–	–40 20 –20	–23 23 –23	–	–	–	–63	63
0	–	–	–30 30 0	0 47 0	2 2 0	–	–28	0
1	–	–	–10 10 10	0 11 11	20 20 20	12 6 6	22	22
2	–	–	–	0 9 18	7 7 14	6 3 6	13	26

$V=\sum n_{uv}v$	–10	–34	–13	29	34	12		$\sum_v vU=169$
uV	30	68	13	0	34	24	$\sum_u uV=169$	$\leftarrow$ *Контроль* $\uparrow$

Для контроля аналогичные вычисления производят по столбцам: произведения $n_{uv}v$ записывают в левый нижний угол клетки, содержащей частоту $n_{uv}\neq 0$; все числа, помещенные в левых нижних углах клеток одного столбца, складывают и их сумму записывают в строку *V*; далее умножают каждую варианту u на *V* и результат записывают в клетках последней строки.

Наконец, сложив все числа последней строки, получают сумму $\sum_u uV$, которая также равна искомой сумме $\sum n_{uv}uv$. Например, для таблицы имеем $\sum_u uV=169$; следовательно, $\sum n_{uv}uv=169$.

Величины $\bar{u}$, $\bar{v}$, $\tilde{\sigma}_u$ и $\tilde{\sigma}_v$ можно вычислить методом произведений; однако, поскольку числа u_i, v_i малы, вычислим $\bar{u}$ и $\bar{v}$, исходя из определения средней, а $\tilde{\sigma}_u$ и $\tilde{\sigma}_v$ – используя формулы:

$$\tilde{\sigma}_u=\sqrt{\bar{u}^2-(\bar{u})^2},\ \tilde{\sigma}_v=\sqrt{\bar{v}^2-(\bar{v})^2}.$$

Найдем $\bar{u}$ и $\bar{v}$:

$$\bar{u}=(\sum n_u u)/n=[5\cdot(-3)+27\cdot(-2)+63\cdot(-1)+29\cdot 1+9\cdot 2]/200=-0{,}425;$$

$$\bar{v}=(\sum n_v v)/n=[12\cdot(-2)+43\cdot(-1)+47\cdot 1+19\cdot 2]/200=0{,}09;$$

Вычислим вспомогательную величину $\bar{u}^2$, а затем σ_u:

$$\bar{u}^2=(\sum n_u u^2)/n=(5\cdot 9+27\cdot 4+63\cdot 1+29\cdot 1+9\cdot 4)/200=1{,}405;$$

$$\tilde{\sigma}_u=\sqrt{\bar{u}^2-(\bar{u})^2}=\sqrt{1{,}405-(0{,}425)^2}=1{,}106.$$

Аналогично получим $\tilde{\sigma}_v=1{,}209$.

Найдем искомый выборочный коэффициент корреляции, учитывая, что ранее уже вычислена сумма $\sum n_{uv}uv=169$:

$$r_B=(\sum n_{uv}uv-n\bar{u}\bar{v})/(n\tilde{\sigma}_u\tilde{\sigma}_v)=[169-200\cdot(-0{,}425)\cdot 0{,}09]/(200\cdot 1{,}106\cdot 1{,}209)=0{,}603.$$

Итак, $r_B=0{,}603$.

Остается найти $\bar{x}$, $\bar{y}$, $\tilde{\sigma}_x$ и $\tilde{\sigma}_y$:

$$\bar{x}=\bar{u}h_1+c_1=-0{,}425\cdot 10+40=35{,}75;$$

$$\bar{y}=\bar{v}h_2+c_2=0{,}09\cdot 10+35=35{,}9;$$

$$\tilde{\sigma}_x=h_1\tilde{\sigma}_u=1{,}106\cdot 10=11{,}06;\ \tilde{\sigma}_y=h_2\tilde{\sigma}_v=1{,}209\cdot 10=12{,}09.$$

Поставим найденные величины в уравнение (*), получим искомое уравнение

$$\bar{y}_x-35{,}9=0{,}603\cdot\frac{12{,}09}{11{,}06}(x-35{,}75),$$

или окончательно

$$\bar{y}_x=0{,}659x+12{,}34.$$

Сравним условные средние, вычисленные: а) по этому уравнению; б) по данным корреляционной таблицы. Например, при x=30:

а) $\bar{y}_{30} = 0{,}659 \cdot 30 + 12{,}34 = 32{,}11$;

б) $\bar{y}_{30} = (23 \cdot 25 + 30 \cdot 5 + 10 \cdot 45)/63 = 32{,}94$.

Как видим, согласование расчетного и наблюдаемого условных средних – удовлетворительное.

Проверка гипотезы о значимости выборочного коэффициента корреляции

Пусть двумерная генеральная совокупность (X, $У$) распределена нормально. Из этой совокупности извлечена выборка объема n и по ней найдены выборочный коэффициент корреляции $r_в$, который оказался отличным от нуля. Так как выборка отобрана случайно, то нельзя заключить, что коэффициент корреляции r генеральной совокупности также отличен от нуля. Возникает необходимость при заданном уровне значимости проверить нулевую гипотезу H_0: $r = 0$ при конкурирующей гипотезе H_1: $r \neq 0$.

Если нулевая гипотеза отвергается, то X и $У$ коррелированны, то есть связаны линейной или нелинейной зависимостью. Если нулевая гипотеза будет принята, то X и $У$ некоррелированны, то есть не связаны корреляционной зависимостью.

В качестве критерия проверки нулевой гипотезы примем случайную величину

$$T_r = r_B \sqrt{\frac{n-2}{1-r_B^2}},$$

которая при справедливости нулевой гипотезы имеет распределение Стьюдента с $k = n - 2$ степенями свободы.

Критическая область – двусторонняя. По таблице критических точек распределения Стьюдента, по заданному уровню значимости α, числу степеней свободы $k = n - 2$ найдем критическую точку $t_{кр}(\alpha; k)$ для двусторонней критической области. Обозначим значение критерия, вычисленное по данным наблюдений, через $T_{набл}$.

Если $|T_{набл}| < t_{кр}$ – нет оснований отвергнуть нулевую гипотезу.

Если $|T_{набл}| > t_{кр}$ – нулевую гипотезу отвергают.