Методические указания для выполнения лабораторной работы 2

Найти выборочное уравнение линейной регрессии Y на X на основании корреляционной таблицы.

Методические указания

Регрессией Y на X или условным математическим ожиданием случайной величины Y относительно случайной величины X называется функция вида

$$M(Y/x)=f(x).$$

Регрессией X на Y называется функция вида $M(X/y)=\varphi(y)$.

Оценками этих функций являются выборочные уравнения регрессии, или условные средние

$$\bar{y}_x = f(x),\ \bar{x}_y = \varphi(y)$$

В том случае, когда варианты парной выборки встречаются по нескольку раз, причем с одним значением варианты x_i может встретиться несколько вариант y_j, их обычно представляют в виде корреляционной таблицы.

Выборочное уравнение прямой линии регрессии Y на X имеет вид

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

где $r_в = \dfrac{\sum n_{xy} xy - n\bar{x}\cdot\bar{y}}{n\sigma_x\sigma_y}$ - выборочный коэффициент корреляции.

Для упрощения расчетов используются условные варианты, которые рассчитываются по формулам

$$u_i = \frac{x_i - C_1}{h_1}, \qquad v_j = \frac{y_j - C_2}{h_2},$$

где C_1 , C_2 – ложные нули (в качестве ложного нуля будем принимать варианту, расположенную в середине вариационного ряда),
h_1 , h_2 – шаги, т.е. разности между двумя соседними вариантами.
В этом случае выборочный коэффициент корреляции

$$r_в = \frac{\sum n_{uv} uv - n\bar{u}\cdot\bar{v}}{n\sigma_u\sigma_v},$$

причем слагаемое $\sum n_{uv} uv$ удобно вычислять, используя расчетную таблицу 1.

Величины $\bar{u}, \bar{v}, \sigma_u, \sigma_v$ могут быть найдены по формулам

$$\bar{u} = \frac{\sum n_u u}{n}, \bar{v} = \frac{\sum n_v v}{n}, \sigma_u = \sqrt{\overline{u^2} - (\bar{u})^2}, \sigma_v = \sqrt{\overline{v^2} - (\bar{v})^2}$$

Для обратного перехода применяются выражения

$$\bar{x} = \bar{u}h_1 + C_1,\ \bar{y} = \bar{v}h_2 + C_2,\ \sigma_x = \sigma_u h_1, \sigma_y = \sigma_v h_2$$

Пример Найти выборочное уравнение линейной регрессии Y на X на основании корреляционной таблицы.

У/Х	15	20	25	30	35	40
100	2	1		7		
120	4		2			3
140		5		10	5	2
160			3	1	2	3

Решение. Для упрощения расчетов перейдем к условным вариантам, которые рассчитываются по формулам

$$u_i = \frac{x_i - 30}{5}, \quad v_j = \frac{y_j - 120}{20}$$

и составим преобразованную корреляционную таблицу с условными вариантами

u/v	**-3**	**-2**	**-1**	**0**	**1**	**2**	$\mathbf{n_v}$
-1	2	1		7			10
0	4		2			3	9
1		5		10	5	2	22
2			3	1	2	3	9
n_u	6	6	5	18	7	8	N=50

Затем составим новую таблицу, в которую внесем посчитанные значения $n_{ij}U_i$ в правый верхний угол заполненной клетки и $n_{ij}V_j$ в левый нижний угол, после чего суммируем верхние значения по строкам для получения значений V_j и нижние значения по столбцам для U_i и подсчитаем величины u_iU_i и v_jV_j.

u/v	-3	-2	-1	0	1	2	V_j	v_jV_j
-1	-6 2 -2	-2 1 -1		0 7 -7			-8	8
0	-12 4 0		-2 2 0			6 3 0	-8	0
1		-10 5 5		0 10 10	5 5 5	4 2 2	-1	-1
2			-3 3 6	0 1 2	2 2 4	6 3 6	5	10
U_i	-2	4	6	5	9	8	-	$\Sigma = 17$
u_i Ui	6	-8	-6	0	9	16	$\Sigma = 17$	-

Подсчитываем суммы $\sum_{i=1}^{k_1} u_i U_i$ и $\sum_{j=1}^{k_2} v_j V_j$. Параллельный подсчет этих сумм осуществляется для контроля правильности расчетов. В данном случае

$$\sum_{i=1}^{k_1} u_i U_i = \sum_{j=1}^{k_2} v_j V_j = 17$$

Находим $\bar{u}, \bar{v}$

$$\bar{u} = (-3 \cdot 6 - 2 \cdot 6 - 1 \cdot 5 + 1 \cdot 7 + 2 \cdot 8)/50 = -0{,}24,$$

$$\bar{v} = (-1 \cdot 10 + 1 \cdot 22 + 2 \cdot 9)/50 = 0{,}6$$

Находим $\overline{u^2}, \overline{v^2}$

$$\overline{u^2} = (9 \cdot 6 + 4 \cdot 6 + 1 \cdot 5 + 1 \cdot 7 + 4 \cdot 8)/50 = 2{,}44$$

$$\overline{v^2} = (1 \cdot 10 + 1 \cdot 22 + 4 \cdot 9)/50 = 1{,}36$$

Определяем σ_u, σ_v

$$\sigma_u = \sqrt{\overline{u^2} - (\bar{u})^2} = \sqrt{2{,}44 - (-0{,}24)^2} = 1{,}54$$

$$\sigma_v = \sqrt{\overline{v^2} - (\bar{v})^2} = \sqrt{1{,}36 - 0{,}6^2} = 1$$

Вычисляем выборочный коэффициент корреляции

$$r_в = \frac{\sum n_{uv} uv - n\bar{u} \cdot \bar{v}}{n\sigma_u \sigma_v} = \frac{17 - 50 \cdot (-0{,}24) \cdot 0{,}6}{50 \cdot 1{,}54 \cdot 1} = 0{,}314$$

Осуществим переход к исходным вариантам:

$$\bar{x} = \bar{u}h_1 + C_1 = 5 \cdot (-0{,}24) + 30 = 28{,}8,$$
$$\bar{y} = \bar{v}h_2 + C_2 = 20 \cdot 0{,}6 + 120 = 132,$$
$$\sigma_x = \sigma_u h_1 = 5 \cdot 1{,}54 = 7{,}7,$$
$$\sigma_y = \sigma_v h_2 = 20 \cdot 1 = 20.$$

Находим уравнение регрессии Y на X $\quad \bar{y}_x - \bar{y} = r_в \frac{\sigma_y}{\sigma_x}(x - \bar{x})$

$\bar{y}_x - 132 = 0{,}314 \frac{20}{7{,}7}(x - 28{,}8)$ или $\bar{y}_x = 0{,}81x + 108{,}51$

Методические указания для выполнения лабораторной работы 3

При уровне значимости $\alpha = 0{,}05$ методом дисперсионного анализа проверить нулевую гипотезу о влиянии фактора на качество объекта на основании пяти измерений для трех уровней фактора $\Phi_1 - \Phi_3$.

Методические указания

Дисперсионным анализом называется статистический метод анализа результатов испытаний, цель которого – оценить влияние одного или нескольких качественных факторов на рассматриваемую величину X.

Рассмотрим схему однофакторного дисперсионного анализа на примере исследования влияния различных видов рекламы на прибыль предприятия.

Если разделить виды рекламы на несколько групп (уровней фактора) и через одинаковые интервалы времени измерять прибыль, то результаты можно представить в виде таблицы

Номер измерения	**Уровни фактора**			
	Φ_1	**Φ_2**	**…**	**Φ_p**
1	X_{11}	X_{12}	…	X_{1p}
2	X_{21}	X_{22}	…	X_{2p}
.	.	.	…	.
q	X_{q1}	X_{q2}	…	X_{qp}
Групповая средняя	$\bar{x}_{г1}$	$\bar{x}_{г2}$	…	$\bar{x}_{гр}$

Общую среднюю можно получить как среднее арифметическое групповых средних

$$\bar{x} = \sum_{j=1}^{p} \bar{x}_{гj} / p$$

На разброс прибыли относительно общей средней влияют как измерения уровня рассматриваемого фактора, так и случайные факторы. Для того чтобы учесть влияние данного фактора, общая выборочная дисперсия разбивается на две части, первая из которых называется факторной $(S_{ф}^2)$, а вторая - остаточной $(S_{ост}^2)$.

С целью учета этих составляющих вначале рассчитываются общая сумма квадратов отклонений вариант от общей средней

$$R_{общ} = \sum_{j=1}^{p} \sum_{i=1}^{q} x_{ij}^{\,2} - pq(\bar{x})^2$$

и факторная сумма квадратов отклонений групповых средних от общей средней, которая и характеризует влияние данного фактора,

$$R_{ф} = q \sum_{j=1}^{p} \left((\bar{x}_{гj})^2 - p(\bar{x})^2 \right)$$

Остаточная сумма квадратов отклонений получается как разность

$$R_{ост} = R_{общ} - R_{ф}$$

Факторная и остаточная дисперсии определяются по формулам:

$$S_{ф}^{2} = \frac{R_{ф}}{p - 1}, \qquad S_{ост}^{2} = \frac{R_{ост}}{p(q - 1)}$$

С целью оценки влияния фактора на изменения рассматриваемого параметра рассчитывается величина

$$f_{набл} = \frac{S_{ф}^{2}}{S_{ост}^{2}}$$

Так как отношение двух выборочных дисперсий распределено по закону Фишера – Снедекора, то полученное значение $f_{набл}$ сравнивают со значением функции распределения в критической точке $f_{кр}$, соответствующей выбранному уровню значимости α. Если $f_{набл} > f_{кр}$, то фактор оказывает существенное воздействие и его следует учитывать, в противном случае он оказывает незначительное влияние, которым можно пренебречь.

Пример

Для проверки влияния внутрицехового оформления на качество продукции рассмотрены три участка по производству однотипной продукции и проведена выборочная проверка процента брака за пять месяцев. Методом дисперсионного анализа при уровне значимости $\alpha = 0{,}05$ проверить нулевую гипотезу о существенном влиянии оформления участка на качество продукции.

Номер измерения	**Ф₁**	**Ф₂**	**Ф₃**
1	2	3	1
2	4	5	4
3	3	4	5
4	2	3	10
5	1	6	3

Решение Находим общую среднюю

Номер измерения	**Ф₁**	**Ф₂**	**Ф₃**
1	2	3	1
2	4	5	4
3	3	4	5
4	2	3	10
5	1	6	3
Групповая средняя	2,4	4,2	4,6

$$\bar{x} = \frac{2{,}4 + 4{,}2 + 4{,}6}{3} = 3{,}73$$

Для расчета $R_{общ}$ составляем таблицу квадратов вариант

Номер измерения	Φ_1	Φ_2	Φ_3
1	4	9	1
2	16	25	16
3	9	16	25
4	4	9	100
5	1	36	9
$\sum$	34	95	151

Вычисляем $R_{общ}$

$$R_{общ} = 34 + 95 + 151 - 3 \cdot 5 \cdot 3{,}73^2 = 71{,}3$$

Находим $R_ф$ по формуле

$$R_ф = 5(2{,}4^2 + 4{,}2^2 + 4{,}6^2 - 3 \cdot 3{,}73^2) = 14{,}1$$

Получаем $R_{ост}$

$$R_{ост} = R_{общ} - R_ф = 71{,}3 - 14{,}1 = 57{,}2$$

Определяем факторную и остаточную дисперсии:

$$S_ф^2 = \frac{R_ф}{p - 1} = \frac{14{,}1}{2} = 7{,}05,$$

$$S_{ост}^2 = \frac{R_{ост}}{p(q - 1)} = \frac{57{,}2}{12} = 4{,}77$$

Находим $f_{набл} = 7{,}05 / 4{,}77 = 1{,}48$

Для уровня значимости $\alpha = 0{,}05$, чисел степеней свободы 2 и 12 находим $f_{кр}$ из таблицы распределения Фишера – Снедекора $f_{кр}(0{,}05; 2; 12) = 3{,}89$

В связи с тем, что $f_{набл} < f_{кр}$, нулевую гипотезу о существенном влиянии внутрицехового оформления на процент брака отвергаем.

Методические указания для выполнения лабораторной работы 4

Дана таблица недельного дохода (X) и недельного потребления (У) для 10 домохозяйств. Необходимо:

а) оценить коэффициенты линейной регрессии по МНК;

б) вычислить стандартную ошибку регрессии;

в) проверить статистическую значимость коэффициентов при уровне значимости $\alpha = 0{,}05$;

г) рассчитать 95% -е доверительные интервалы для теоретических коэффициентов регрессии;

д) рассчитать коэффициент детерминации для построенного уравнения регрессии

X	100	120	140	160	180	200	220	240	260	280
У	60	70	90	100	110	120	120	150	140	180

Оценим коэффициенты линейной регрессии по МНК

Эмпирическое уравнение регрессии имеет вид

$$\tilde{y} = b_0 + b_1 x$$

Рассчитаем коэффициенты уравнения регрессии по формулам

$$\begin{cases} b_1 = \dfrac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \bar{x}^2} \\ b_0 = \bar{y} - b_1 \bar{x} \end{cases}$$

Для определения сумм составим расчетную таблицу

i	x	y	x^2	xy	y^2	$\tilde{y}$	e	e^2
1	100	60	10000	6000	3600	60,9	-0,9	0,81
2	120	70	14400	8400	4900	72,7	-2,7	7,29
3	140	90	19600	12600	8100	84,5	5,5	30,25
4	160	100	25600	16000	10000	96,3	3,7	13,69
5	180	110	32400	19800	12100	108,1	1,9	3,61
6	200	120	40000	24000	14400	119,9	0,1	0,01
7	220	120	48400	26400	14400	131,7	-11,7	136,89
8	240	150	57600	36000	22500	143,5	6,5	42,25
9	260	140	67600	36400	19600	155,3	-15,3	234,09
10	280	180	78400	50400	32400	167,1	12,9	166,41
Сумма	1900	1140	394000	236000	142000	1140	0	635,3
Сред.	190	114	39400	23600	14200	114	0	63,53

$$\begin{cases} b_1 = \dfrac{23600 - 190 \cdot 114}{39400 - 190^2} = \dfrac{1940}{3300} = 0{,}59 \\ b_0 = 114 - 0{,}59 \cdot 190 = 1{,}9 \end{cases}$$

Коэффициент регрессии b_1 показывает абсолютную силу связи между вариацией X и вариацией У.

$$\tilde{y} = 1{,}9 + 0{,}59x$$

Вычислим стандартную ошибку регрессии S

$$S = \sqrt{\frac{\sum e_i^2}{n-2}}, \qquad S = \sqrt{\frac{635{,}3}{8}} = 8{,}91$$

Проверим статистическую значимость коэффициентов при уровне значимости $\alpha = 0{,}05$. Она может быть решена по схеме:

$$H_0 : b_1 = 0$$
$$H_1 : b_1 \neq 0$$

Вычислим значения

$$t_{b_1} = \frac{b_1}{S_{b_1}}, \qquad \text{где } S_{b_1} = \sqrt{\frac{S^2}{n(\overline{x^2} - \bar{x}^2)}}$$

и $t_{кр} = t_{\frac{\alpha}{2}, n-2}$ (по таблице распределения Стьюдента)

$$S_{b_1} = \frac{8{,}91}{\sqrt{10 \cdot (39400 - 36100)}} = 0{,}05$$

$$t_{b_1} = \frac{0{,}59}{0{,}05} = 11{,}8 \quad t_{кр} = 2{,}3$$

Сравним модуль наблюдаемого значения $|t_{b_1}|$ с критическим значением. Если $|t_{b_1}| > t_{кр}$, то нулевая гипотеза отвергается, следовательно, коэффициент b_1 статистически значим. Если нулевая гипотеза принимается, то коэффициент b_1 статистически незначим. Следовательно, в нашем случае коэффициент b_1 статистически значим.

Аналогично проверяется статистическая значимость коэффициента b_0.

$$S^2_{b_0} = S^2_{b_1} \overline{x^2} = \frac{S^2 \sum x_i^2}{n(\overline{x^2} - \bar{x}^2)} = 0{,}05^2 \cdot 39400 = 98{,}5$$

$$t_{b_0} = \frac{b_0}{S_{b0}} = \frac{1{,}9}{\sqrt{98{,}5}} = 0{,}19, \qquad t_{кр} = 2{,}3$$

Следовательно, в нашем случае коэффициент b_1 статистически незначим.

Рассчитаем 95% -е доверительные интервалы для теоретических коэффициентов регрессии. Они вычисляются по формулам

$$\left[b_0 - t_{\frac{\alpha}{2},n-2} S_{b0}; b_0 + t_{\frac{\alpha}{2},n-2} S_{b0} \right]$$
$$\left[b_1 - t_{\frac{\alpha}{2},n-2} S_{b1}; b_1 + t_{\frac{\alpha}{2},n-2} S_{b1} \right]$$

$$\alpha = 0.05$$

$$(1{,}9 - 2{,}3 \cdot 9{,}92; \quad 1{,}9 + 2{,}3 \cdot 9{,}92)$$
$$(0{,}59 - 2{,}3 \cdot 0{,}05; \quad 0{,}59 + 2{,}3 \cdot 0{,}05)$$

$$(-20{,}916; \quad 24{,}716)$$
$$(-0{,}475; \quad 0{,}705)$$

Рассчитаем коэффициент детерминации для построенного уравнения регрессии

$$R^2 = 1 - \frac{\sum e_i^2}{\sum (y_i - \bar{y})^2} = r_{xy}^2$$

$$r_{xy} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sqrt{\overline{x^2} - \bar{x}^2}\sqrt{\overline{y^2} - \bar{y}^2}}$$

$$r_{xy} = \frac{23600 - 190 \cdot 114}{\sqrt{3300}\sqrt{14200 - 114^2}} = \frac{1940}{1993{,}3} = 0{,}97$$

Такое значение линейного коэффициента корреляции говорит о высокой тесноте связи между X и У.

$$R^2 = 0{,}97^2 = 0{,}947$$

Квадрат коэффициента корреляции называется коэффициентом детерминации. $R^2 = 0{,}947$. Это означает, что доля вариации У, объясненная вариацией фактора X, включенного в уравнение регрессии, равна 94,7%, а остальные 5,3% вариации приходятся на долю других факторов, не учтенных в уравнении регрессии.