

«Линейная регрессия и корреляция: смысл и оценка параметров»

- ***F* критерий Фишера** - оценивает качество уравнения регрессии - состоит в проверке гипотезы H_0 (о том, что коэффициент регрессии равен нулю, т.е. $b=0$, т.е. фактор x не оказывает влияния на результат y).

- Расчету F-критерия предшествует анализ дисперсии.
- Центральное место в нем занимает разложение общей суммы квадратов отклонений на две части «объясненную» и «необъясненную».

$$\sum (y - \bar{y})^2 = \sum (\hat{y}_x - \bar{y})^2 + \sum (y - \hat{y}_x)^2$$

- Общая объясненная остаточная
- (необъясненная)

- Любая сумма квадратов отклонений связана с числом степеней свободы – df , т.е. с числом свободы независимого варьирования признака.
- Для общей суммы квадратов требуется $(n-1)$ число отклонений.

- Для расчета df объясненной суммы квадратов имеем:

$$\hat{y}_x = a + bx$$

$$a = \bar{y} - b\bar{x}$$

$$\hat{y}_x = \bar{y} - b(x - \bar{x})$$

- Число степеней свободы равно 1.

- Число степеней свободы остаточной суммы квадратов = число степ. свободы для общей суммы квадратов – число степ. свободы для объясненной регрессии.

- дисперсии на одну степень свободы

$$D_{\text{общ}} = \frac{\sum (y - \bar{y})^2}{n - 1}$$

$$D_{\text{факт}} = \frac{\sum (\hat{y}_x - \bar{y})^2}{1}$$

$$D_{\text{ост}} = \frac{\sum (y - \hat{y}_x)^2}{n - 2}$$

$$F = \frac{D_{\text{факт}}}{D_{\text{ост}}}$$

- Значение F-критерия признается достоверным, если оно больше табличного. В этом случае гипотеза H_0 отклоняется.

- Если $F_{\text{табл}} < F_{\text{факт}}$, то H_0 - гипотеза о случайной природе оцениваемых характеристик отклоняется и признается их статистическая значимость и надежность.
- Если $F_{\text{табл}} > F_{\text{факт}}$, то гипотеза H_0 не отклоняется и признается статистическая незначимость, ненадежность уравнения регрессии.

$$\sum (\hat{\sigma}_x - n\bar{y})^2 = r^2 \cdot \frac{2}{y} \cdot$$

$$\sum (\sigma - \hat{\psi}_x)^2 = (1 - r^2) \cdot \frac{2}{y} \cdot$$

$$F_{\text{факт}} = \frac{r^2}{1 - r^2} (n - 2)$$

- **Таблица значений F-критерия Фишера при уровне значимости $\alpha = 0,05$**

k1	1	2	3	4	5	6	8	12	24	∞
k2										
1	161,45	199,50	215,72	224,57	230,17	233,97	238,89	243,91	249,04	254,32
2	18,51	19,00	19,16	19,25	19,30	19,33	19,37	19,41	19,45	19,50
3	10,13	9,55	9,28	9,12	9,01	8,94	8,84	8,74	8,64	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,04	5,91	5,77	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,82	4,68	4,53	4,36
6	5,99	5,14	4,76	4,53	4,39	4,28	4,15	4,00	3,84	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,73	3,57	3,41	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,44	3,28	3,12	2,93

ПРИМЕР

(количество факторов – 1)

- Дисперсионный анализ результатов регрессии

Вариация результата y	Число степеней свободы	Сумма квадратов отклонений	Дисперсия на одну степень свободы, D	$F_{факт}$	$F_{табл}$
Общая	$df = n - 1 = 8$	6,316	-	-	-
Факторная	$k_1 = 1$	5,116	?	?	?
Остаточная	$k_2 = ?$	1,200	?	-	-

- Для оценки статистической значимости коэффициентов регрессии и корреляции рассчитываются ***t-критерий Стьюдента*** каждого из показателей и **доверительные интервалы**.
- Оценка значимости коэффициентов регрессии и корреляции с помощью *t*-критерия Стьюдента проводится путем сопоставления их значений с величиной случайной ошибки:

$$t_a = \frac{a}{m_a}$$

$$t_b = \frac{b}{m_b}$$

$$t_r = \frac{r}{m_r}$$

$$m_a = S_{ocm} \frac{\sqrt{\sum x^2}}{n\sigma_x}$$

$$m_b = \frac{S_{ocm}}{\sigma_x \sqrt{n}}$$

$$m_r = \sqrt{\frac{1-r^2}{n-2}}$$

$$S_{ocm}^2 = \frac{\sum (y - \hat{y}_x)^2}{n-2}$$

- Сравнимая фактическое и критическое (табличное) значения t -статистики - $t_{табл}$ и $t_{факт}$ - принимаем или отвергаем гипотезу H_0 .
- Если $t_{табл} < t_{факт}$ то гипотеза H_0 - о незначимости параметра отклоняется, т.е. a , b и не случайно отличаются от нуля и сформировались под влиянием систематически действующего фактора x .
- Если $t_{табл} > t_{факт}$ то гипотеза H_0 не отклоняется и признается случайная природа формирования a , b или r_{xy} .

- ***доверительный интервал***
- для расчета доверительного интервала определяем ***предельную ошибку*** Δ для каждого показателя

$$\Delta_b = t_{\text{табл}} m_b \quad \Delta_a = t_{\text{табл}} m_a$$

- для коэффициентов регрессии границы доверительного интервала составят:

$$(a - \Delta_a, a + \Delta_a) \quad (b - \Delta_b, b + \Delta_b)$$

- Если в границы доверительного интервала попадает ноль, т.е. нижняя граница отрицательна, а верхняя положительна, то оцениваемый параметр принимается нулевым, так как он не может одновременно принимать и положительное, и отрицательное значения.

- **Средняя ошибка аппроксимации** - среднее отклонение расчетных значений от фактических (дает оценку качества построенной модели):

$$\bar{A} = \frac{1}{n} \sum \left| \frac{y - \hat{y}_x}{y} \right| \cdot 100$$

- Допустимый предел значений - не более 8-10%.