ЛЕКЦИЯ 20

1. Множественная регрессия

Для того чтобы построить модель нелинейной связи выходного показателя с несколькими переменными, необходимо применить методы множественной регрессии. В общем случае условию равенства (1) всегда соответствует относительно большая система уравнений. Например, для отыскания коэффициентов уравнения вида

$$y = a_0 + a_1 x + a_{11} x^2 + a_2 z + a_{22} z^2 \qquad (1)$$

необходимо решить систему уравнений:

$$\begin{cases} \sum_i y_i = a_0 N + a_1 \sum_i x_i + a_{11} \sum_i x_i^2 + a_2 \sum_i z_i + a_{22} \sum_i z_i^2; \\ \sum_i x_i y_i = a_0 \sum_i x_i + a_1 \sum_i x_i^2 + a_{11} \sum_i x_i^3 + a_2 \sum_i x_i z_i + a_{22} \sum_i x_i z_i^2; \\ \sum_i x_i^2 y_i = a_0 \sum_i x_i^2 + a_1 \sum_i x_i^3 + a_{11} \sum_i x_i^4 + a_2 \sum_i x_i^2 z_i + a_{22} \sum_i x_i^2 z_i^2; \\ \sum_i z_i y_i = a_0 \sum_i z_i + a_1 \sum_i z_i x_i + a_{11} \sum_i z_i x_i^2 + a_2 \sum_i z_i^2 + a_{22} \sum_i z_i^3; \\ \sum_i z_i^2 y_i = a_0 \sum_i z_i^2 + a_1 \sum_i x_i z_i^2 + a_{11} \sum_i x_i^2 z_i^2 + a_2 \sum_i z_i^3 + a_{22} \sum_i z_i^4; \end{cases} \qquad (2)$$

Подобным образом можно получить систему уравнений для уравнения с любым числом членов, но при добавлении каждого нового члена трудоемкость решения резко возрастает. Точность модели оценивается по остаточной дисперсии:

$$s^2_{y,\text{ост}} = \frac{\sum_{i=1} (y_i - y_{i,\text{мод}})^2}{N - k - 1}, \qquad (3)$$

где k — число коэффициентов уравнения.

Строго математически исходными предпосылками регрессионного и корреляционного анализов являются условия:

— случайные величины нормально распределены;

— дисперсия зависимой переменной одинакова при всех значениях аргумента;

— отдельные наблюдения переменных не связаны друг с другом, т. е. являются независимыми.

Практически первые два условия невыполнимы. Поэтому возникает вопрос, можно ли пользоваться изложенными методами при каких-либо отклонениях указанных условий от требуемых? В настоящее время считается, что методами корреляционного и регрессионного анализов можно пользоваться и тогда, когда зависимая переменная не распределена нормально, а наблюдения зависимы, лишь бы распределения были одновершинными и в некоторой степени симметричными.

Коэффициенты уравнений теоретической линии регрессии в общем случае оцениваются с ошибкой

$$s_{aj} = \frac{s_{y.ост}}{\sqrt{\sum_{i=1}^{N} x_{ij}^2}}, \qquad (4)$$

где x — общее обозначение члена функции $y = \sum_{j=0}^{k} a_j x_j$.

Следует помнить, что при небольшом числе степеней свободы $s_{y.ост}^2$ становится ненадежной оценкой. В этом случае следует пользоваться ошибкой воспроизводимости s_y. Тогда оценкой среднего квадратического отклонения кривой регрессии будет величина

$$s_{\bar{y}}(x) = s_y \sqrt{\sum_{j=0}^{k} \frac{x_j^2}{\sum_{i=1}^{N} x_{ij}^2}}. \qquad (5)$$

Доверительные интервалы для теоретической линии регрессии

$$\bar{y}(x) - t s_{\bar{y}}(x) \leq \bar{y}_{И} \leq \bar{y}(x) + t s_{\bar{y}}(x). \qquad (6)$$

Пример.

Определить коэффициенты a_0, a_1, a_2, a_3. в уравнении регрессии: $y = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_3$. Исходные данные приведены в таблице 1. В каждой точке эксперимента было проведено по три опыта.

Таблица 1

Результаты эксперимента

N	x_1	x_2	x_3	y_1	y_2	y_3	$\bar{y}$	$x_1\bar{y}$	$x_2\bar{y}$	$x_3\bar{y}$	x_1x_2	x_1x_3	x_2x_3	x_1^2	x_2^2	x_3^2	$\hat{y}$	S_y^2	S_A^2
1	1	0	1	46	45	53	48	48	0	48	1	0	0	1	0	1	47	19	3
2	0	1	2	40	47	45	44	0	44	88	0	0	2	0	1	4	45	13	3
3	2	2	0	36	38	49	41	82	82	0	4	0	0	4	4	0	40	49	3
4	3	3	3	27	29	37	31	93	93	93	9	9	9	9	9	9	32	28	3
Σ	6	6	6				164	223	219	229	13	10	11	14	14	14		109	12

Система уравнений для определения коэффициентов регрессии имеет вид:

$$\begin{aligned}
\sum y &= a_0 N + a_1 \sum x_1 + a_2 \sum x_2 + a_3 \sum x_3 \\
\sum yx_1 &= a_0 \sum x_1 + a_1 \sum x_1^2 + a_2 \sum x_1 x_2 + a_3 \sum x_1 x_3 \\
\sum yx_2 &= a_0 \sum x_2 + a_1 \sum x_1 x_2 + a_2 \sum x_2^2 + a_3 \sum x_2 x_3 \\
\sum yx_3 &= a_0 \sum x_3 + a_1 \sum x_1 x_3 + a_2 \sum x_2 x_3 + a_3 \sum x_3^2
\end{aligned} \qquad (1)$$

Подставим соответствующие суммы из таблицы 1 в систему (1) и решим её.

$$\begin{cases} 164=4a_0+6a_1+6a_2+6a_3 & a_0=50 \\ 223=6a_0+14a_1+13a_2+10a_3 & a_1=-2 \\ 219=6a_0+13a_1+14a_2+11a_3 & a_2=-3 \\ 229=6a_0+10a_1+11a_2+14a_3 & a_3=-1 \end{cases}$$

Уравнение регрессии имеет вид:

$$y = 50 - 2x_1 - 3x_2 - x_3$$

Используя данные таблицы 1 определим дисперсию адекватности и эксперимента:

$$S_A^2 = \sum_1^n \cdot k(\bar{y}_i - \hat{y})^2 = 12 \qquad S_y^2 = \sum_1^n \sum_1^k \frac{(y_i - \bar{y}_i)^2}{k-1} = 109$$

Число степеней свободы этих дисперсий равно:

$$\varphi_1 = n - 2 = 2 \qquad \varphi_2 = N - n = 12 - 4 = 8$$

Критерий Фишера:

$$F = \frac{S_{ад}^2 / f_2}{S_{ад}^2 / f_1} = \frac{109/8}{12/2} = 2{,}2$$

Табличное значение критерия Фишера $F_{кр}$=19,2. Условие адекватности модели :

$F_{\text{êð}} = 19{,}2 > F = 2{,}2$. Вывод: модель адекватна.

Произведем проверку эксперимента на воспроизводимость по критерию Кохрена. В таблице 1 находим максимальную дисперсию и рассчитываем отношение:

$$G = \frac{S_{y3}^2}{S_y^2} = \frac{49}{109} \approx 0{,}45$$

Для числа степеней свободы $f_2=k+1=4$, $f_1=N-1=3$ табличное значение критерия Кохена при уровне значимости **P=95%** $G_{табл}$=0,6. Так как $G_{табл}$=0,6>0,45, то отличие дисперсий незначимо и, следовательно, дисперсии однородны.

2. Множественная корреляция

Для получения уравнения множественной корреляции обычно используют систему, составленную из найденных коэффициентов парной корреляции.

Пусть необходимо получить связь

$$y = a_0 + a_1 x_1 + a_2 x_2. \tag{7}$$

Система нормальных уравнений для нее

$$\begin{cases} \sum_i y_i = a_0 N + a_1 \sum_i x_{1i} + a_2 \sum_i x_{2i}; \\ \sum_i y_i x_{1i} = a_0 \sum_i x_{1i} + a_1 \sum_i x_{1i}^2 + a_2 \sum_i x_{1i} x_{2i}; \\ \sum_i y_i x_{2i} = a_0 \sum_i x_{2i} + a_1 \sum_i x_{1i} x_{2i} + a_2 \sum_i x_{2i}^2. \end{cases} \tag{8}$$

Представим связь (7) в стандартизованной форме

$$\frac{y_i - \bar{y}}{s_y} = \ell_1 \frac{x_{1i} - \bar{x}_1}{s_{x_1}} + \ell_2 \frac{x_{2i} - \bar{x}_2}{s_{x_2}} \tag{9}$$

и запишем систему уравнений для нее, разделив все члены на $N - 1$

$$\frac{1}{N-1}\sum_i \frac{y_i - \bar{y}}{s_y} \cdot \frac{x_{1i} - \bar{x}_1}{s_{x_1}} = \ell_1 \frac{1}{N-1}\sum_i \left(\frac{x_{1i} - \bar{x}_1}{s_{x_1}}\right)^2 + \quad + \ell_2 \frac{1}{N-1}\sum_i \frac{x_{1i} - \bar{x}_1}{s_{x_1}} \cdot \frac{x_{2i} - \bar{x}_2}{s_{x_2}}; \tag{10}$$

$$\frac{1}{N-1}\sum_i \frac{y_i - \bar{y}}{s_y} \cdot \frac{x_{2i} - \bar{x}_2}{s_{x_2}} = \ell_1 \frac{1}{N-1}\sum_i \left(\frac{x_{1i} - \bar{x}_1}{s_{x_1}}\right)\left(\frac{x_{2i} - \bar{x}_2}{s_{x_2}}\right) + \quad + \ell_2 \frac{1}{N-1}\sum_i \left(\frac{x_{2i} - \bar{x}_2}{s_{x_2}}\right)^2. \tag{11}$$

В первом уравнении системы (7) все суммы обращаются в ноль, откуд$\ell_0 = 0$

Отсюда

$$\begin{cases} r_{yx_1} = \ell_1 + \ell_2 r_{x_1x_2}; \\ r_{yx_2} = \ell_1 r_{x_1x_2} + \ell_2. \end{cases} \qquad (12)$$

Теснота связи уравнения множественной корреляции с экспериментальными данными оценивается коэффициентом
множественной корреляцииR

Коэффициент множественной корреляции не может быть по абсолютной величине больше единицы.

Доверительные интервалыR оцениваются так же, как и для коэффициента парной корреляции .

$$R = \sqrt{r_{yx_1}\ell_1 + r_y}. \qquad (13)$$