МИНИСТЕРСТВО СЕЛЬСКОГО ХОЗЯЙСТВА РОССИЙСКОЙ ФЕДЕРАЦИИ
ФАКУЛЬТЕТ ПЕРЕРАБАТЫВАЮЩИХ ТЕХНОЛОГИЙ
Кафедра технологии хранения и переработки растениеводческой продукции

Лекция №10
«Регрессионный анализ. Метод наименьших квадратов»

Регрессионный анализ

Регрессионный анализ неразрывно связан с корреляционным анализом. Если корреляция позволяет измерить связь между признаками Х и У, то регрессионный анализ позволяет найти форму этой связи с помощью нахождения уравнения регрессии.

Регрессионный анализ

Если величина X и У связаны точно линейной функцией $y=в_0+в_1x_1$, то $r=\pm1$, а знак будет соответствовать коэффициенту $в_1$, если величины X и У связаны произвольной зависимостью, коэффициент имеет значение $-1< r <1$.

Регрессионный анализ

Найти уравнение регрессии – это значит по эмпирическим (фактическим) данным математически описать изменения взаимно коррелируемых величин.

Рассчитанные по уравнению регрессии значения результативного признака называют теоретическим и обычно обозначают y_x (y выровненный по x) и рассматривается как функция : $y=f(x)$.

Оценка значимости

Выбор теоретической линии регрессии обусловлен формой эмпирической линии регрессии, а также с учетом природы изучаемых показателей и специфики их взаимосвязи.

Могут использоваться уравнения:

- 1 $y=a_0+a_1x$ (прямая)
- 2 $y= a_0+a_1x+a_2x^2$ (парабола 2-го порядка)
- 3 $y= a_0+a_1 1/x$ (гипербола)
- 4 $y= a_0a_1^x$ (показательная функция)
- 5 $y= a_0+b\lg x$ (логарифмическая)

Обычно зависимость, выражаемую уравнением прямой, называют прямолинейной, а все остальное – криволинейными.

Выбрав тип функции, по эмпирическим данным определяют параметры уравнения.

Регрессионная прямая

Существует несколько методов нахождения параметров уравнения регрессии. Наиболее часто используется метод наименьших квадратов (МНК).

Суть метода заключается в следующем искомые теоретические значения результативного признака y_x должны быть такими, при которых бы обеспечивалась минимальная сумма квадратов их отклонений от эмпирических значений, т.е.

$$S = \sum_{i=1}^{n} (y_i - \overline{y_x}) = \min$$

Предполагается что разброс точек относительно кривой подчиняется закону нормального распределения.

$$y = f(x)$$

Зависимость переменной у от х может выражаться формулой:

$$y = b_0 + b_1 x$$

У – зависимая, х – независимая переменная.

Если же x представляет зависимую, а y независимую, то речь идет о регрессии x по y

$$x = b_0 + b_2 y$$

Величины b_0, b_1, b_2 – коэффициенты регрессии, постоянные величины.

Они вычисляются по формулам:

$$b_1 = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2}$$

$$b_0 = \overline{y} - b_1 \overline{x}$$

Где x_i и y_i – частные эмпирические значения изучаемых величин.

Оценка значимости коэффициентов уравнения регрессии проводится по критерию Стьюдента:

$$t_p = \frac{(b_i)}{Sb_i}$$

Где b_i – коэффициенты уравнения регрессии;

Sb_i – среднее квадратичное отклонение для коэффициентов уравнения регрессии b_0 и b_1 находят:

$$S_{a1} = \sqrt{\frac{S^2{}_{воспх} \sum x_i^2}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}}$$

$$S_{b1} = \sqrt{\frac{nS^2{}_{воспх}}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}}$$

Дисперсию воспроизводимости определяют:

$$S^2_{восп} = \frac{\sum_{i=1}^{n} (y_i - \overline{y})^2}{n-1}$$

где n – количество экспериментов.

Если $t_{расч} > t_{табл}$, то коэффициент значим.

Адекватность уравнения проверяют по критерию Фишера:

$$F_{расч} = \frac{S^2_{ад}}{S^2_{воспр}}$$

Дисперсия адекватности определяется уравнением:

$$S^2_{ад} = \frac{\sum_{i=1}^{n} (\overline{y}_i - \hat{y}_i)^2}{n - l}$$

Где l – число значимых коэффициентов в уравнении регрессии

(n-l) – число степеней свободы адекватности.

При постановке ряда экспериментальных задач необходимо не только нахождение уравнения регрессии, описывающего зависимость тех или иных факторов, но и поиск их оптимальных значений.

Существует ряд методов оптимизации – метод золотого сечения, метод координатного спуска, метод спирального координатного спуска.

Задачи оптимизации, независимо от метода ставят таким образом: поиск уравнения регрессии и дальнейший анализ его с поиском наилучших результатов.