

Методы научных исследований

9 Лекция. Регрессионный анализ. Метод наименьших квадратов

Исполнитель: Байболов Асан Ерболатович

Электронный адрес: asan.baibolov@kaznaru.edu.kz

ПЛАН ЛЕКЦИИ

- 1) Регрессионный анализ;
- 2) Метод наименьших квадратов

СПИСОК ЛИТЕРАТУРЫ:

1. Дрейпер Н., Смит Г. Прикладной регрессионный анализ. Множественная регрессия. Applied Regression Analysis. - 3-е изд. - М.: «Диалектика», 2007. - 912 с.
2. Радченко С. Г. Методология регрессионного анализа. — К.: «Корнийчук», 2011.
3. Сборник задач по математике: Учеб. пособие для втузов : В 4 ч. Ч. 4: Теория вероятностей. Математическая статистика / Под общ. ред. А. В. Ефимова, А. С. Пospelova. - 3-е изд., перераб. и доп. - М.: Физматлит, 2004. - 432 с.

Регрессионный анализ



Регрессионный анализ - набор статистических методов исследования влияния одной или нескольких независимых переменных $X_1, X_2, \dots, X_p$ на зависимую переменную Y . Независимые переменные иначе называют регрессорами, а зависимые переменные - критериальными. Терминология зависимых и независимых переменных отражает лишь математическую зависимость переменных, а не причинно-следственные отношения. Наиболее распространённый вид регрессионного анализа - линейная регрессия, когда находят линейную функцию, которая, согласно определённым математическим критериям, наиболее соответствует данным. Например, в методе наименьших квадратов вычисляется прямая (или гиперплоскость), сумма квадратов между которой и данными минимальна.

Метод наименьших квадратов



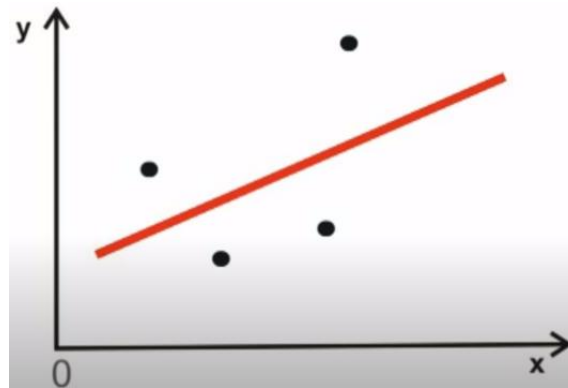
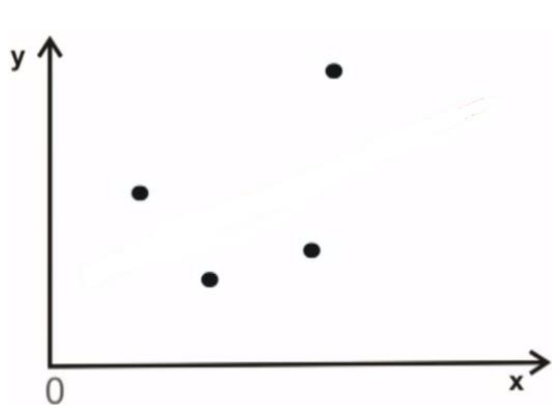
Метод наименьших квадратов (МНК) - один из методов регрессионного анализа для оценки неизвестных величин по результатам измерений, содержащих случайные ошибки. Применяется также для приближённого представления заданной функции другими (более простыми) функциями и часто оказывается полезным при обработке наблюдений.

В настоящее время широко применяется при обработке количественных результатов естественнонаучных опытов, технических данных, астрономических и геодезических наблюдений и измерений.

Можно выделить следующие достоинства метода:

- а) расчеты сводятся к механической процедуре нахождения коэффициентов;
- б) доступность полученных математических выводов.

Основным недостатком МНК является чувствительность оценок к резким выбросам, которые встречаются в исходных данных.



$$y_i = a + bx_i$$

Необходимо найти такие значения параметров a и b , которые бы доставляли минимум функции, т. е. минимизировали бы сумму квадратов отклонений наблюдаемых значений результативного признака y от теоретических значений $\tilde{y}$ (значений, рассчитанных на основании уравнения регрессии):

$$F = \sum_{i=1}^n (y_i - \tilde{y}_i)^2 = \sum_{i=1}^n (y_i - \tilde{a} + \tilde{b}x_i)^2 \rightarrow \min$$

При минимизации функции неизвестными являются значения коэффициентов регрессии a и b . Значения зависимой и независимой переменных известны из наблюдений.

Для того чтобы найти минимум функции двух переменных, нужно вычислить частные производные этой функции по каждой из оцениваемых параметров и приравнять их к нулю. В результате получаем стационарную систему уравнений для функции:

$$\begin{cases} \frac{\partial F}{\partial a} = -2 \sum_{i=1}^n (y_i - \tilde{a} - \tilde{b}) = 0 \\ \frac{\partial F}{\partial b} = -2 \sum_{i=1}^n (y_i - \tilde{a} + \tilde{b}) \times x_i = 0 \end{cases}$$

Если разделить обе части каждого уравнения системы на (-2), раскрыть скобки и привести подобные члены, то получим систему:

$$\begin{cases} \tilde{b} \sum_{i=1}^n x_i^2 + \tilde{a} \sum_{i=1}^n x_i = \sum_{i=1}^n x_i \times y_i \\ \tilde{b} \sum_{i=1}^n x_i + \tilde{a} \times n = \sum_{i=1}^n y_i \end{cases}$$

Эта система нормальных уравнений относительно коэффициентов a и b для зависимости

$$y_i = a + bx_i$$

Решением системы нормальных уравнений являются оценки неизвестных параметров уравнения регрессии a и b

$$b = \frac{n \sum_{i=1}^n x_i \times y_i - \sum_{i=1}^n x_i \times \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2} = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2} = Cov(x, y)$$
$$a = \bar{y} - b \times \bar{x},$$

где $\bar{y}$ - среднее значение зависимого признака;
 $\bar{x}$ - среднее значение независимого признака;
 $\overline{xy}$ - среднее арифметическое значение произведения зависимого и независимого признаков;
 $G^2(x)$ - дисперсия независимого признака;
 $Cov(x, y)$ - ковариация между зависимым и независимым признаками.

ВОПРОСЫ ДЛЯ САМОКОНТРОЛЯ

1. Основные понятия и определения математической статистики
2. Что называется генеральной совокупностью и выборкой;
3. Выборочная функция распределения

Спасибо за внимание !