Лекция 9

Ilya Yaroshevskiy

1 ноября

Содержание

1 Математическая регрессия

Определение. Регрессией X на Z называется функция $f(z) = E(X|Z = z)$.

$$x = f(z)$$

— **уравнение регрессии**, а ее график — **линия регрессии**

Пусть при n экспериментах, при значениях $z_1, \dots, z_n$ случайной величины Z наблюдаемые значения $x_1, \dots, x_n$ случайной величины X.

Обозначим ε разницу между наблюдаемыми и экспериментальными значениями X

$$\varepsilon_i = X_i - E(X|Z = z_i) = X_i - f(z_i)$$

Тогда $x_i = f(z_i) + \varepsilon_i$, $1 \le i \le n$, и ε_i — независимые случайные нормальные величины. $\varepsilon_i \in N(0, \sigma^2)$

$$a = 0, \text{ т.к. } E\varepsilon_i = E(X_i) - E(f(z_i)) = EX_i - E(X|Z = z) = 0$$

Цель: по данным значениям как можно более точно оценить функцию $f(z)$

Примечание. При этом предполагается(например из теории), что $f(z)$ — функция определенного типа, параметры которой не известны

1.1 Метод наименьшие квадратов

Метод наименьших квадратов состоит в выборе параметров функции $f(z)$, таким образом, чтобы минимизировать сумму квадратов ошибок

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (X - f(z))^2 \to \min$$

Определение. Пусть $\Theta = (\Theta_1, \dots, \Theta_n)$ — набор параметров функции $f(z)$.

$$\hat{\Theta} \quad \min \sum_{i=1}^{n} \varepsilon_i^2$$

2 Линейная парная регрессия

Пусть имеется *линейная регрессия*:

$$f(z) = a + bz$$

Тогда $X_i = a+bz_i+\varepsilon_i$, $1 \le i \le n$. Найдем оценки неизвестных параметров a и b методом наименьших квадратов (МНК).

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (X_i - a - bz_i)^2 \to min$$

$$\frac{\partial}{\partial a} \sum_{i=1}^{n} \varepsilon_i^2 = 2\sum_{i=1}^{n} (X_i - a - bz_i) \cdot (-1) = -2\sum_{i=1}^{n} X_i + 2\sum_{i=1}^{n} a + 2b\sum_{i=1}^{n} z_i = -2(n\overline{X} - na - bn\overline{z}) = 0$$

$$\frac{\partial}{\partial b} \sum_{i=1}^{n} \varepsilon_i^2 = 2\sum_{i=1}^{n} (X_i - a - bz_i) \cdot (-z_i) = -2\left(\sum_{i=1}^{n} X_i z_i - a\sum_{i=1}^{n} z_i - b\sum_{i=1]^{n} z_i^2}\right) = -2(n\overline{Xz} - an\overline{z} - bn\overline{z}) = 0$$

$$\begin{cases} n\overline{X} - na - bn\overline{z} = 0 \\ n\overline{Xz} - an\overline{z} - bn\overline{z^2} \end{cases} \Leftrightarrow \begin{cases} \overline{X} - a - b\overline{z} = 0 \\ \overline{Xz} - z\overline{z} - b\overline{z^2} \end{cases} \Leftrightarrow \begin{cases} a + b\overline{z} = \overline{X} \\ a\overline{z} + b\overline{z^2} = \overline{Xz} \end{cases}$$

$$\begin{cases} a = \overline{X} - b\overline{z} \\ (\overline{Z} - b\overline{z})\overline{z} + b\overline{z^2} = \overline{Xz} \end{cases} \Leftrightarrow \begin{cases} a = \overline{x} - b\overline{z} \\ b(\overline{z^2} - \overline{z}^2) = \overline{Xz} - \overline{X} \cdot \overline{z} \end{cases} \Leftrightarrow \begin{cases} a = \overline{X} - b\overline{z} \\ b = \frac{\overline{Xz} - \overline{X} \cdot \overline{z}}{\overline{z^2} - \overline{z}^2} \end{cases}$$

$$\begin{aligned} \overline{X_z} &= E(\overline{X}|Z = z) = f(z) \\ \overline{X_z} &= a + bz \\ \overline{X_z} &= \overline{X} - b\overline{z} + b\overline{z} \\ \overline{X_z} - \overline{X} &= b(z - \overline{z}) \\ \overline{X_z} - \overline{X} &= \frac{\overline{Xz} - \overline{X} \cdot \overline{z}}{\hat{\sigma_z}^2} \cdot (z - \overline{z}) \\ \overline{X_z} - \overline{X} &= \frac{\overline{Zx} - \overline{X} \cdot \overline{z}}{\hat{\sigma_z} \cdot \hat{\sigma_x}} \cdot \frac{\hat{\sigma}_x}{\hat{\sigma_z}} \cdot (z - \overline{z}) \\ \frac{\overline{X_z} - \overline{X}}{\hat{\sigma_x}} &= \frac{\overline{Xz} - \overline{x} \cdot \overline{z}}{\hat{\sigma_z} \cdot \hat{\sigma_x}} \cdot \frac{z - \overline{z}}{\hat{\sigma_z}} \end{aligned}$$

$$\frac{\overline{X_z} - \overline{X}}{\hat{\sigma_x}} = r_B \cdot \frac{\overline{z} - z}{\hat{\sigma_z}}$$

, где $r_B = \frac{\overline{Xz} - \overline{X} \cdot \overline{z}}{\hat{\sigma_z} \cdot \hat{\sigma_x}}$

2.1 Геометрический смысл прямой линейной регрессии

Прямая регрессии строится таким образом, чтобы сумма квадратов длин отрезков была наименьшей.

Определение. Выборочный коэффициент линейной корреляции

$$r_B = \frac{\overline{Xz} - \overline{X} \cdot \overline{z}}{\hat{\sigma_x} \cdot \hat{\sigma_y}}$$

Примечание. Отсюда видим, что выборочный коэффициент линейной корреляции является оценкой теоретического коэффициента полученного по методу моментов. Поэтому выборный коэффициент корреляции характеризует силу линейной связи между двумя случайными величинами. Знак r_B показывается является ли корреляция прямой или обратной

2.2 Проверка гипотезы по значимости коэффициента линейной корреляции

Пусть двумерная случайная величины распределена нормально. По выборке объема n вычислен r_B, а r — теоретический коэффициент линейной корреляции.

- $H_0 : r = 0$
- против $H_1 : r \neq 0$, т.е. коэффициент r_B статистически значимый

Теорема 2.1. Если H_0 верна, то статистика

$$K = \frac{r_B\sqrt{n-z}}{\sqrt{q-r_0^2}} \in T_{n-1}$$

— распределение Стьюдента с $n-2$ степенями свободы
Критерий: Пусть $t_c r$ — квантиль T_{n-2} уравнения значимости α

2.3 Выборочное корреляционное отношение

Пусть имеется k выборок случайной величины X, полученных при уровнях фактора $z_1, \ldots, z_k$. Вычислены общая дисперсия D_O, D_M — межгрупповая дисперсия, D_B — внутригрупповая дисперсия. По теореме о разложении дисперсии $D_O = D_M + D_B$ Доделать