

第二章 统计学

2.4 成对数据的统计分析

2.4.1 最小二乘法

十八世纪末,天文学和大地测量经常遇到一种困境,比如一条行星轨道方程只由几个参数决定,但天文学家手上有海量的观测记录,由于观测总有误差,无论怎么选取参数,都会与一部分观测数据产生矛盾。显然观测误差是无法避免的,那么如何确定轨道方程的参数呢?

一种思路是,确定一条轨道曲线,使得它与所有观测数据的偏差尽可能小,也就是让误差的平方和最小(因误差有正有负,但二者不可抵消)。这便是最小二乘法(Least Squares Method)。

1805 年,法国数学家勒让德在《计算彗星轨道的新方法》中正式提出并命名了最小二乘法,这是该方法第一次公开发表。而高斯则声称自己从 1795 年起就已使用这一方法¹,并据此算出了谷神星的轨道。他在 1809 年出版的《天体运动论》中系统阐述了最小二乘法,并给出了理论支撑:当观测误差服从正态分布时,最小二乘估计在统计意义下是最优的。后来的高斯-马尔科夫定理则从理论上进一步巩固了最小二乘的地位:在误差零均值、互不相关、同方差的线性模型中,最小二乘估计是所有线性无偏估计中方差最小的,即最佳线性无偏估计。今天,最小二乘法已经成为回归分析、曲线拟合、信号处理以及机器学习中最基础、最常用的工具之一。

已知成对观测数据 $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, 可将它们视为平面直角坐标系中的 n 个点。希望找到一条直线 $Y = a + bX$, 使得对每一个 $x_i (i = 1, 2, \dots, n)$, 由这个直线方程计算得到的预测值 $a + bx_i$ 与实际观测值 y_i 的差距(这称为第 i 个数据的残差)整体最小。

记第 i 个数据的残差为 $e_i = y_i - (a + bx_i)$, 我们希望所有残差的绝对值之和最小, 即最小化 $\sum_{i=1}^n |e_i|$, 该优化目标在数学上是合理的, 但在实际计算中却不方便。我们考虑使用残差平方和作为衡量标准, 即通过确定 a, b 的取值以最小化

$$\sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - (a + bx_i))^2$$

这一优化目标可导, 是凸函数, 且对离群点敏感度较高(平方级增长)。

先研究简单的情形, 考虑 3 对数据 $(x_1, y_1), (x_2, y_2), (x_3, y_3)$, 求 a, b 的值以最小化

$$\sum_{i=1}^3 (y_i - (a + bx_i))^2$$

¹今天一般认为, 他们各自独立发现了这一方法, 高斯使用更早, 勒让德发表更早。

用向量解决这一问题. 设偏差向量

$$\begin{aligned}\mathbf{e} &= (y_1 - (a + bx_1), y_2 - (a + bx_2), y_3 - (a + bx_3)) \\ &= (y_1, y_2, y_3) - a(1, 1, 1) - b(x_1, x_2, x_3) \\ &= \mathbf{y} - a\mathbf{1} - b\mathbf{x}\end{aligned}$$

问题简化为求 a, b 的值以最小化 $|\mathbf{e}|$, 设 $\overrightarrow{OY} = \mathbf{y}, \overrightarrow{OI} = \mathbf{1}, \overrightarrow{OX} = \mathbf{x}, \overrightarrow{OY}, \overrightarrow{OI}, \overrightarrow{OX}$ 均为已知向量, 记 α 为过点 O, I, X 的平面, 当 a, b 任意取值时, $a\overrightarrow{OI} + b\overrightarrow{OX}$ 可以表示这一平面内的任意一条向量. 于是当 $|\mathbf{e}| = |\overrightarrow{OY} - (a\overrightarrow{OI} + b\overrightarrow{OX})|$ 取到最小值时, 一定有 $\mathbf{e} \perp \alpha$, 此时 $\mathbf{e} \cdot \overrightarrow{OI} = 0, \mathbf{e} \cdot \overrightarrow{OX} = 0$, 即

$$\begin{cases} [(y_1, y_2, y_3) - a(1, 1, 1) - b(x_1, x_2, x_3)] \cdot (1, 1, 1) = 0 \\ [(y_1, y_2, y_3) - a(1, 1, 1) - b(x_1, x_2, x_3)] \cdot (x_1, x_2, x_3) = 0 \end{cases}$$

解得

$$\hat{b} = \frac{\sum_{i=1}^3 x_i y_i - 3\bar{x}\bar{y}}{\sum_{i=1}^3 x_i^2 - 3\bar{x}^2}, \hat{a} = \bar{y} - \hat{b}\bar{x}$$

推广为 n 对数据 $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ 时, 直觉地猜测, 使得 $\sum_{k=1}^n [y_k - (a + bx_k)]^2$ 最小化的 a, b 取值应当为

$$\hat{b} = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}, \hat{a} = \bar{y} - \hat{b}\bar{x}$$

类比先前的步骤, 设 $\mathbf{y} = (y_1, y_2, \dots, y_n)^T, \mathbf{1} = (1, \dots, 1)^T, \mathbf{x} = (x_1, \dots, x_n)^T, \mathbf{z} = a\mathbf{1} + b\mathbf{x}$, 于是 $\mathbf{e} = \mathbf{y} - \mathbf{z}, \mathbf{x}$ 张成一个二维子平面 W , 我们断言, $\|\mathbf{e}\|$ 取到最小值当且仅当 $\mathbf{e}$ 正交于平面 W 中的任意向量, 即 $\mathbf{e} \cdot \mathbf{1} = 0, \mathbf{e} \cdot \mathbf{x} = 0(*)$.

证明. 设 a_0, b_0 使得 $\mathbf{z}_0 = a_0\mathbf{1} + b_0\mathbf{x}$ 且 $\mathbf{y} - \mathbf{z}_0$ 正交于平面 W 中的任意向量, 于是

$$\|\mathbf{e}\|^2 = \|\mathbf{y} - \mathbf{z}\|^2 = \|\mathbf{y} - \mathbf{z}_0 + \mathbf{z}_0 - \mathbf{z}\|^2 = \|\mathbf{y} - \mathbf{z}_0\|^2 + \|\mathbf{z}_0 - \mathbf{z}\|^2 \geq \|\mathbf{y} - \mathbf{z}_0\|^2$$

□

将 $(*)$ 式变形即可得到结论.

2.4.2 样本相关系数

给定随机变量 X 和 Y , 可知, 若 X 和 Y 的线性相关性好, 则观测值 y_i 和预测值 $\hat{a} + \hat{b}x_i$ 的差距应该很小, 最理想的情况是

$$y_i = \hat{a} + \hat{b}x_i, i = 1, 2, \dots, n$$

又有 $\bar{y} = \hat{a} + \hat{b}\bar{x}$, 将其代入上式, 消去 $\hat{a}$, 得

$$y_i - \bar{y} = \hat{b}(x_i - \bar{x}), i = 1, 2, \dots, n$$

记向量 $\mathbf{u} = (x_1 - \bar{x}, x_2 - \bar{x}, \dots, x_n - \bar{x}), \mathbf{v} = (y_1 - \bar{y}, y_2 - \bar{y}, \dots, y_n - \bar{y})$, 则上式可记为 $\mathbf{v} = \hat{b}\mathbf{u}$. 这表明线性回归方程最理想的情况是向量 $\mathbf{u}, \mathbf{v}$ 共线. 因此可以考虑用向量 $\mathbf{u}, \mathbf{v}$ 夹角的大小来刻画 X 和 Y 线性相关的程度, 记

$$r = \cos\langle \mathbf{u}, \mathbf{v} \rangle = \frac{\mathbf{u} \cdot \mathbf{v}}{|\mathbf{u}||\mathbf{v}|} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

上式也可简化为

$$r = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sqrt{\sum_{i=1}^n x_i^2 - n\bar{x}^2} \sqrt{\sum_{i=1}^n y_i^2 - n\bar{y}^2}}$$

根据柯西不等式的一般形式可以得到 $|r| \leq 1$. $|r|$ 越接近 1, 说明 X 和 Y 的线性相关性越强; $|r|$ 越接近 0, 说明 X 和 Y 的线性相关性越弱.

第二章 统计学

2.4 成对数据的统计分析

第二章 统计学

2.4 成对数据的统计分析

第二章 统计学

2.4 成对数据的统计分析

第二章 统计学

2.4 成对数据的统计分析