线性回归方向选择的分析

李 化 平

(北京钢铁学院物理系)

摘 要

本文对常见的一类简单的线性模型(直线方程 $y = a + bx$)进行最小二乘法处理中遇到的几个问题,根据普遍理论作一些具体讨论和分析,以说明在这些特殊情形下如何能实现简化处理而又不失结果的准确性. 证明了对线性函数 $y = a + bx$ 在不同方向进行线性回归分析所得出的两组不同的最小二乘解,在相关系数 $r = 1$ 时完全相同,并给出线性回归分析方向选择的一些简便判断方法。

在无系统误差的测量条件下,用线性回归分析确定线性参数的最佳估值具有唯一性、无偏性和最小方差等优点. 因此,这种方法在许多领域的精密的科学实验中进行数据处理时被广泛地采用,同时也是统计不确定度的评定的一种计算方法.

回归分析的普遍方法早在一百年前就已发展完善和成熟,即使是一般性模型(非线性、不等精度和多个未知量)的普遍情况,得到的结果也是准确的. 问题是涉及处理数据的计算量很大,因此考虑在某些特殊情形下,既能做到不失结果的准确性,同时又能大大减少处理数据的计算量就变得很重要了.

本文拟对比较常见的一类特殊情形进行最小二乘法处理时遇到的几个问题,根据普遍理论作一些具体讨论和分析,以说明在这些特殊情况下如何能实现简化处理而不失结果的准确性.

设待测物理量 y 是任意变量 x 的线性函数,

$$y = a + bx, \tag{1}$$

对 x 和 y 分别进行了 K 次等精度独立测量,要求按线性回归分析确定 a 和 b 的最佳估值. 下面分两种情况进行讨论.

一、假设仅 y_i 有误差,或 x_i 的测量误差很小,可以略去不计,则根据最小二乘法原理,

$$\sum_{i=1}^{K} v_{y_i}^2 = \sum_{i=1}^{K} [y_i - (a + bx_i)]^2 = \min. \tag{2}$$

由此可求出 a 和 b 的最佳估值为

$$\left.\begin{aligned} \hat{a} &= \frac{[y][xx] - [x][xy]}{K[xx] - [x][x]}, \\ \hat{b} &= \frac{K[xy] - [x][y]}{K[xx] - [x][x]}, \end{aligned}\right\} \tag{3}$$

式中 $[x]$, $[xy]$ 分别代表 $\sum_{i=1}^{K} x_i$ 和 $\sum_{i=1}^{K} x_i y_i$.

K 个测定方程

$$y_i - (a + bx_i) = v_{y_i}$$

中任一方程的标准差为

$$S(y) = \sqrt{\frac{\sum v_{y_i}^2}{K-2}}.$$

由误差传播定理可求得 $\hat{a}$ 和 $\hat{b}$ 的标准差,即

$$S(\hat{a}) = \sqrt{\sum_{i=1}^{K} \left(\frac{\partial \hat{a}}{\partial y_i}\right)^2 \cdot S^2(y)}.$$

因

$$\left(\frac{\partial \hat{a}}{\partial y_1}\right) = \frac{[xx] - [x]x_1}{K[xx] - [x][x]},$$

$$\left(\frac{\partial \hat{a}}{\partial y_1}\right)^2 = \frac{[xx]^2 - 2[xx][x]x_1 + [x]^2 \cdot x_1^2}{(K[xx] - [x][x])^2},$$

$$\sum_{i=1}^{K}\left(\frac{\partial \hat{a}}{\partial y_i}\right)^2 = \frac{K[xx]^2 - 2[xx][x][x] + [x]^2[xx]}{\Delta^2}$$

$$= \frac{[xx]}{\Delta} \qquad (\Delta = K[xx] - [x][x]),$$

故得 $\hat{a}$ 的标准差

$$S(\hat{a}) = \sqrt{\frac{[xx]}{\Delta}} \cdot S(y).$$

同理可求出

$$S(\hat{b}) = \sqrt{\frac{K}{\Delta}} \cdot S(y).$$

二、假设 x_i 测量有误差，y 测量误差可忽略不计，则这时应变换 $y = a + bx$ 的形式成

$$x = \frac{y}{b} - \frac{a}{b},$$

最小二乘法准则变为

$$\sum_{i=1}^{K}\left(x_i - \frac{y_i}{b'} + \frac{a'}{b'}\right)^2 = \min.$$

由此得出

$$\left.\begin{aligned} a' &= \frac{[x][yy] - [y][xy]}{[x][y] - K[xy]}, \\ b' &= \frac{[y][y] - K[yy]}{[x][y] - K[xy]}. \end{aligned}\right\} \quad (4)$$

(3)和(4)式表明，存在两个不同的最小二乘解，亦即对同一测量数组 x_i 和 y_i，可以拟合出两条最佳直线。因此，在对测量数据进行线性回归时，需要正确选择拟合方向，才能得出唯一确定的参数 a 和 b 的最佳值。

若 y 与 x 是理想的正相关，即相关系数

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \cdot \sum(y_i - \bar{y})^2}} = 1, \quad (5)$$

故有

$$\sum(x_i - \bar{x})(y_i - \bar{y}) = \sqrt{\sum(x_i - \bar{x})^2 \cdot \sum(y_i - \bar{y})^2},$$

$$\sum x_i y_i - \frac{1}{K}(\sum x_i)(\sum y_i) = \sqrt{\left\{\sum x_i^2 - \frac{1}{K}(\sum x_i)^2\right\} \cdot \left\{\sum y_i^2 - \frac{1}{K}(\sum y_i)^2\right\}}.$$

采用高斯符号，上式可写成

$$[xy] - \frac{1}{K}[x][y] = \sqrt{\left\{[xx] - \frac{1}{K}[x]^2\right\} \cdot \left\{[yy] - \frac{1}{K}[y]^2\right\}},$$

平方后乘 K^2，得

$$K^2[xy]^2 - 2K[xy][x][y] + [x]^2[y]^2 = K^2[xx][yy] - K[xx][y][y] - K[yy][x]^2 + [x]^2[y]^2,$$

或写成

$$K[xy][x][y] - K^2[xy]^2 - [x]^2[y]^2 + K[xy][x][y] = K[xy][y]^2 - K^2[xx][yy] - [x]^2[y]^2 + K[yy][x]^2,$$

故有

$$(K[xy] - [x][y])([x][y] - K[xy]) = (K[xx] - [x]^2)([y]^2 - K[yy]),$$

$$\frac{K[xy] - [x][y]}{K[xx] - [x]^2} = \frac{[y]^2 - K[yy]}{[x][y] - K[xy]}. \quad (6)$$

比较(3)，(4)，(6)式得

$$\hat{b} = \hat{b}'. \quad (7)$$

用相同方法可证明，在 $r = 1$ 时，有 $\hat{a} = \hat{a}'$。将(5)式平方后乘 K，移项整理后得

$$[y]^2[xx] - [x][y][xy] + K[xy]^2 = K[xx][yy] - [x]^2[yy] + [x][y][xy]x[x].$$

将上式两端分别减去 $K[y][xy][xx]$，得

$$[y]^2[x][xx] - K[y][xy][xx] - [x]^2[y][xy] - K[x][xy]^2 = K[x][xx][yy] - K[xy][y][xx] - [yy][x]^3 + [x]^2[y][xy],$$

故有

$$([y][xx] - [x][xy])([y][x] - K[xy]) = (K[xx] - [x]^2)([yy][x] - [y][xy])$$

或

$$\frac{[y][xx] - [x][xy]}{K[xx] - [x]^2} = \frac{[x][yy] - [y][xy]}{[x][y] - K[xy]},$$

故

$$\hat{a} = \hat{a}' \quad (8)$$

(7)和(8)式表明，当相关系数 $r = 1$ 时，两种最小二乘法处理给出的参数最佳估值是相等的。教学实验多数都不是去寻求新的规律，而是通过函数的已知关系，用最小二乘法精确地确定参数。可以说，绝大多数测量都能保证相关系数 $r \simeq 1$，因而按习惯[1,2]在 y 方向运用最小二乘法求解是完全正确的，即不存在方向选择的

问题. 如果测量精度较低，则偶然误差掩盖了它们的线性相关性，从而导致相关系数 $\gamma < 1$（即测量数据点有明显起伏），这时就需要考虑方向的选择问题.

判断线性函数 y 和自变量 x 的测量数据之间相关情况的最简便方法是，在方格纸上描出 y_i 与 x_i 的关系图. 如测量数据点 (x_i, y_i) 基本上落在一条直线上，则可认为 $\gamma \simeq 1$，故可按一般习惯在 y 方向求最小二乘解. 若 $\gamma < 1$，或测量数据点有明显的分散性（探求未知规律的科学研究实验多数是处理这样的问题），这时必须判断清楚哪个方向测量误差大，并在误差大的那个方向运用最小二乘法，才能给出参数最佳估值的正确值.

判别的方法是比较 x 和 y 两个量的测量相对误差

$$\frac{S(x)}{\bar{x}} \text{ 与 } \frac{S(y)}{\bar{y}}$$

的大小即可，考虑到

$$\frac{S(x)}{\bar{x}} : \frac{S(y)}{\bar{y}} = \frac{\sqrt{\frac{\sum v_{x_i}^2}{K-2}}}{\bar{x}} : \frac{\sqrt{\frac{\sum v_{y_i}^2}{K-2}}}{\bar{y}}$$

$$= \frac{\sqrt{\frac{1-\gamma^2}{K-2}\sum (x_i - \bar{x})^2}}{\bar{x}} : \frac{\sqrt{\frac{1-\gamma^2}{K-2}\sum (y_i - \bar{y})^2}}{\bar{y}}$$

$$= \frac{\sqrt{\sum (x_i - \bar{x})^2}}{\bar{x}} : \frac{\sqrt{\sum (y_i - \bar{y})^2}}{\bar{y}}.$$

若计算得

$$\frac{\sqrt{\sum (x_i - \bar{x})^2}}{\bar{x}} > 3\frac{\sqrt{\sum (y_i - \bar{y})^2}}{\bar{y}},$$

则按微小误差准则，y 方向的测量误差可忽略，即应在 x 方向求最小二乘解；反之，就应在 y 方向求最小二乘解.

除上述方法外，也可根据自己的实验经验来判断.

若二者的误差相近，则应该用正交回归方法来求最小二乘解. 也可分别在 x 和 y 方向求最小二乘解，亦即作出两条回归直线，则最佳直线必位于两个最小二乘解所确定的回归直线之间，若一个方向的误差比另一个方向大些，则最佳直线还应接近误差数值小的直线.

如果测量数据点 (x_i, y_i) 分散性很大，或相关系数 $\gamma < 1$ 较多，则对线性函数系为已知的测量，表明测量精度太低，偶然误差完全掩盖了 x 和 y 之间的真正相关性；对探求未知规律的科学研究实验，则表明 x 和 y 之间只存在小的相关，或者是零相关. 无论是属于前者或是后者，用线性回归分析方法来处理数据实际意义都不大，也没有必要.

以上讨论了回归分析方向的选择问题. 下面再分析等精度条件.

对上面谈的第一种情况，还要求各 y_i 测量等精度，$\hat{a}$ 和 $\hat{b}$ 才能满足

$$\sum_{i=1}^{K} v_{y_i}^2 = \sum [y_i - (\hat{a} + \hat{b}x_i)]^2 = \min. \quad (9)$$

在实际问题中，也存在测量精度不等的问题，即 y 方向的各测量值标准差不等，或 y 的测量值的标准差随着测量值的大小变化，因此需对 $\hat{a}$ 和 $\hat{b}$ 满足的(9)式加以修正.

对于不等精度，如测量值的相对误差相同（如电桥、电位差计等），则它们的标准差有如下关系：

$$S(E) = \frac{S(y)}{y}.$$

最小二乘准则变为

$$\sum \left(\frac{v_{y_i}}{y_i}\right)^2 = \sum E_{y_i}^2 = \min,$$

或

$$\sum E_{y_i}^2 = \left[\frac{y_i}{\hat{a} + \hat{b}x_i} - 1\right]^2 = \min. \quad (10)$$

由 $\frac{\partial}{\partial \hat{a}} \sum E_{y_i}^2 = 0$ 和 $\frac{\partial}{\partial \hat{b}} \sum E_{y_i}^2 = 0$，可得到

$$\sum \frac{y_i^2}{(\hat{a} + \hat{b}x_i)^3} = \sum \frac{y_i}{(\hat{a} + \hat{b}x_i)^2},$$

$$\sum \frac{y_i^2 x_i}{(\hat{a} + \hat{b}x_i)^3} = \sum \frac{x_i y_i}{(\hat{a} + \hat{b}x_i)^2}.$$

对上式求解，即可求出参数 a，b 的最佳估值 $\hat{a}$ 和 $\hat{b}$.

[1] Y. Beers, Introduction to the Theory of Errors, Addison Wesley Publ, Co. Inc.,(1957), 36.

[2] G. Parratt, Probability Experimental Errors in Science, John Wiley and Sons, Inc.,(1961),128.