点到直线距离与垂线距离的平方和最小法在直线回归中的研究

1　垂线距离最小二乘法原理

最小二乘法（又称最小平方法）是一种数学优化技术，于19世纪初由勒让德和高斯 分别独立提出.

现以一元线性回归为例. 设有一组试验数据 $(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)$，欲寻求线性函数(理论分析证明二者存在线性相关性) ：$y=ax+b$. 式中a,b均为待常数. 现行的最小二乘法理论认为，只要RSS(residual sum of squares)最小，则a,b即为最佳拟合. 简化为

$$Q=\sum[y_i-ax_i-b]^2 \tag{1}$$

最小。

对式（1）求偏导并另偏导为0，则

$$\frac{\partial Q}{\partial}=-2\sum(y_i-ax_i-b)x_i=0$$
$$\frac{\partial Q}{\partial}=-2\sum y_i-ax_i-b=0$$

即

$$a\sum x_i^2+b\sum x_i=\sum x_iy_i$$
$$a\sum x_i+nb=\sum y_i$$

该一元二次方程解得$a=l_{xy}/l_{xx}$,b=$\bar{y}-a\bar{x}$.

式中$l_{xy}=\sum(x_i-\bar{x})\,(y_i-\bar{y})=\sum x_iy_i-\frac{1}{n}\sum x_i\sum y_i$,$l_{xx}=\sum(x_i-\bar{x})^2=$

$\sum {x_i}^2-\frac{1}{n}(\sum x_i)^2$.

2. 点到直线最小二乘法原理

由点到直线距离公式$d_i=|y_i-ax_i-b|/\sqrt{1+a^2}$,转化为对 Q=$\sum d_i^2$求最小.

则

$$\frac{\partial Q}{\partial a}=-2\sum x_i(y_i-ax_i-b)/(1+a^2)-2a(ax_i-y_i+b)^2/(1+a^2)^2\ =0 \tag{2}$$
$$\frac{\partial Q}{\partial b}=-2\sum\frac{y_i-ax_i-b}{1+a^2}=0 \tag{3}$$

由 3 式可计算得

$$a\sum x_i+nb=\sum y_i,\quad \text{b}=\bar{y}-a\bar{x}. \tag{4}$$

平均点$(\bar{x}\ ,\bar{y})$ 落在拟合曲线上.

把 4 式代入 2 式，最后得

$$a^2 l_{xy} + a\left(l_{xx} - l_{xy}\right) - l_{xy} = 0 \tag{5}$$

上式为关于 a 的一元二次方程，可解其根形式为：

$$a_2 = {(l_{yy} - l_{xx}) \pm \sqrt{(l_{yy} - l_{xx})^2 + 4{l_{xy}}^2}} \Big/ {2l_{xy}} \tag{6}$$

分析当倾角为锐角，$a_2 > 0$，式 6 中取“+”有效；

当倾角为钝角，$a_2 < 0$，式 6 中取“+”有效；

因此

$$a_2 = {(l_{yy} - l_{xx}) + \sqrt{(l_{yy} - l_{xx})^2 + 4{l_{xy}}^2}} \Big/ {2l_{xy}} \tag{7}$$

代入式 4 可得b_2.

3. 实例对比

使用程序软件：R

数据集S1: 100个数据点，符合曲线$y = 2x + 1$变化趋势，考虑数据的正常误差是影响因素(包括操作人员、机械等)的微小变化造成的,这类误差即随机误差,往往呈现正态分布 ，因此加一个正态随机偏差，符合N（0，5）分布。

数据集合S2: 100个数据点，符合曲线$y = 100x + 50$变化趋势，采用S1类似偏差，符合N（0，500）分布。

数据集合S3：在S1基础上加5个噪点，噪点在可行域内均匀分布.

用两种方法对以上数据集进行拟合，拟合结果如下表1所示：

表1 拟合结果参数表

数据集	相关系数	a_1	a_2	b_1	b_2
S1	0.9968362	1.998968	2.009143	0.7784571	0.7784571
S2	0.9732198	100.5679	103.3350	69.89993	-69.83708
S3	0.9044248	1.897437	0.4766561	7.381926	-0.6093772

用r生成图像如下所示：

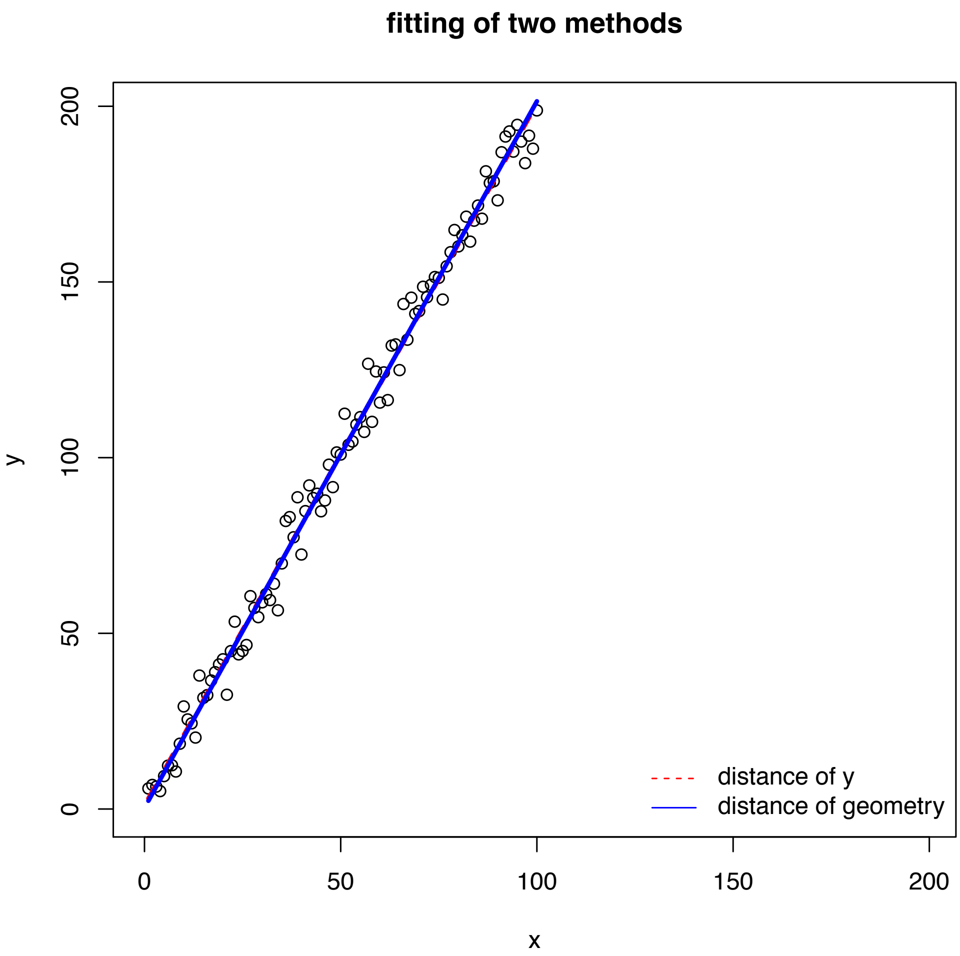

图1. S1图像

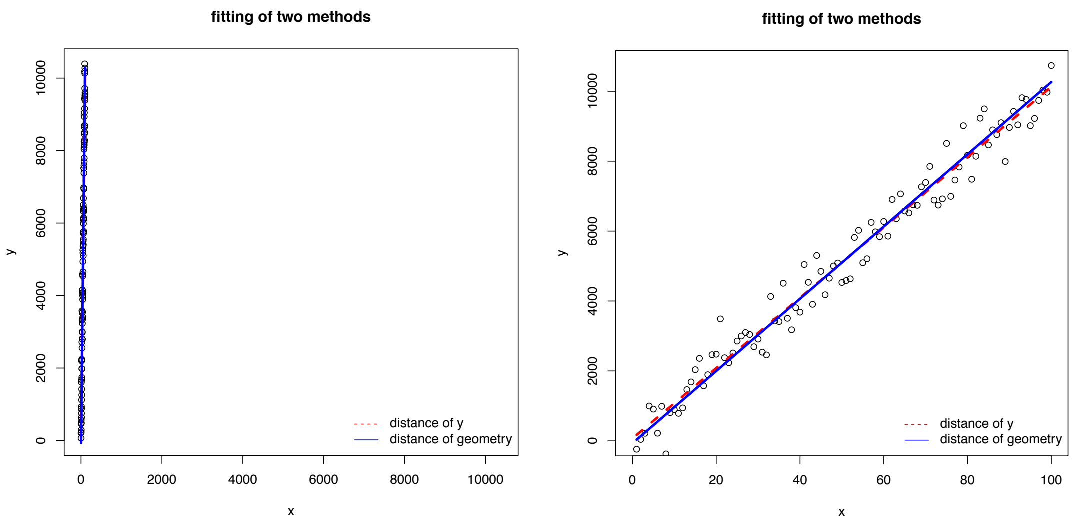

图2. S2图像取不同坐标

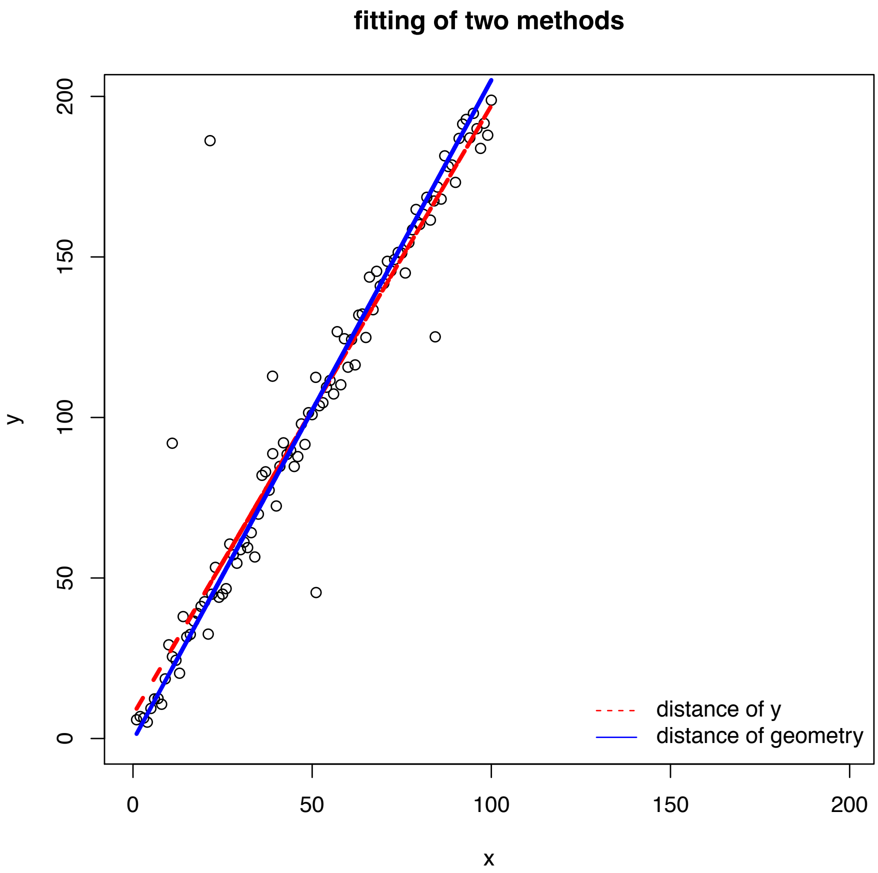

图3.　S3图像

由上可知，当2个变量试验数据间线性相关系数靠近1时 ，差别不大，当相关系数较小，偏离度变大。