第十一章　　拟合

§11.1一元线性最小二乘法

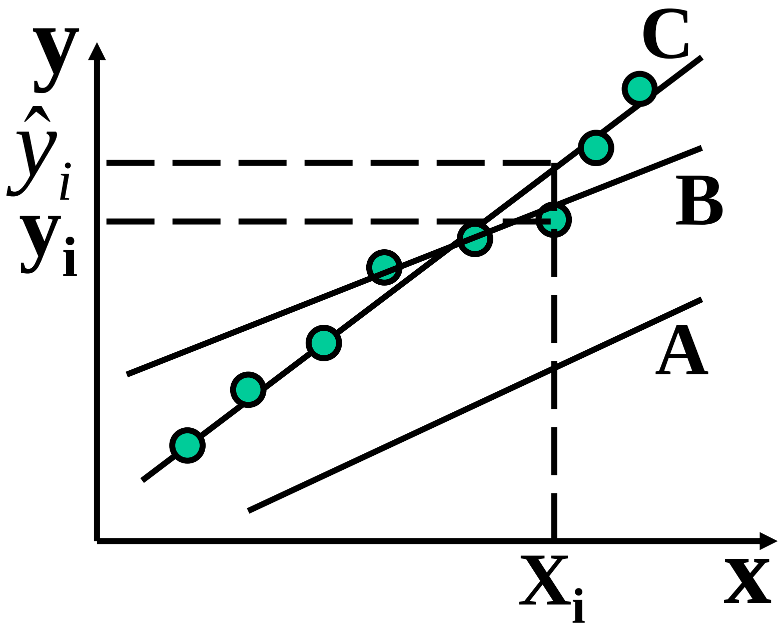
11.1.1. 一般介绍:

设已知物理量x和y间存在线性关系:

$$y=a+bx \qquad (11.1)$$

当x取值为x_1，x_2，…x_i，…x_m时，测得y值y_1，y_2，…y_i，…y_m。现需要根据这些数据（x_i，y_i）来计算式11.1中的截距a和斜率b，即建立y与x间的这种线性关系。这个问题就是一元线性拟和的问题。

如果我们把数据(x_i，y_i)画在x-y直角坐标系中，所得点一般不能准确的连成一条直线， 如图所示，这是因为实验数据肯定存在误差。



回归线示意图

为了建立y与x间的线性关系，只能找寻这样一条直线，使之尽可能靠近各个(x_i，y_i)点。这种线称为回归线，与之相应的方程称为回归方程。

目前的任务就是根据各原始点找寻一条回归线，这条回归线要满足如下两条要求:

(1)该线是直线。

(2)从总体上看，该线比任何其它 直线都更靠近各原始点。

怎样判断一条线与各原始点最为靠近呢? 目前常用的标准是：“残差平方和最小”。

所谓残差，定义为:

i点残差：(x_i，y_i)为原始点，y_i与按回归方程计算的y值 $\hat{y}$ 间的差称为残差，用δ $_i$ 表示：

$$\delta_i = y_i - \hat{y}_i = y_i - (a + bx_i)$$

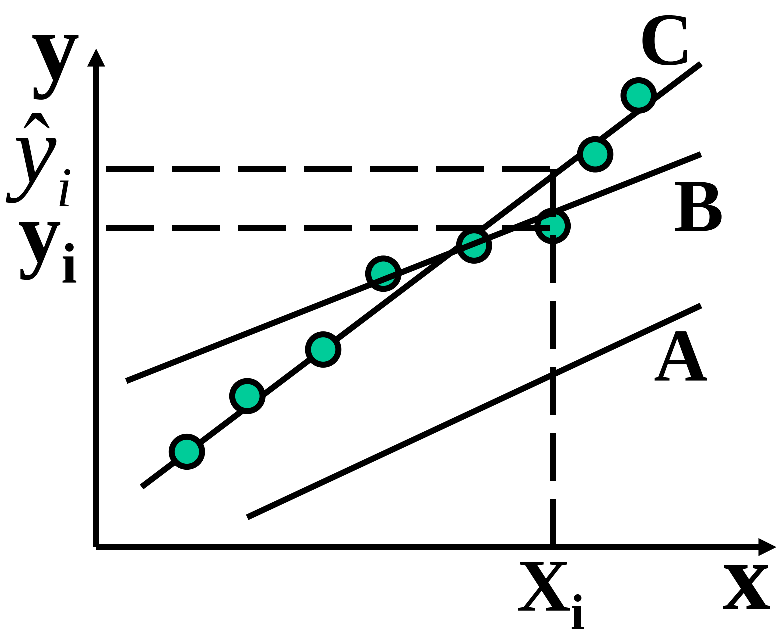
$$\hat{y}_i = a + bx_i \qquad \textbf{(11.19)}$$

残差平方和Q就是每个点的残差进行平方，再彼此相加：

$$Q = \sum_{i=1}^{m} \delta_i^2$$

看图，直线A远离各原始点，直线B比较靠进，直线B的残差平方和$Q_B < Q_A$。如果在一切直线中，直线C的残差平方和最小

$$Q=\sum_{i=1}^{m}\delta_i^2=\sum_{i=1}^{m}\left(y_i-\hat{y}_i\right)^2=\sum_{i=1}^{m}\left(y_i-a-bx_i\right)^2$$



回归线示意图

直线C就是所求的回归线，能最好地代表y与x间的线性关系。

“平方”过去曾被称为“二乘”，这就是最小二乘法名称的来历。

当回归线是只有一个自变量和一个应变量的直线时，称为一元线性最小二乘法。

11．1．2．　一元线性最小二乘法的算法：

残差平方和Q应是直线参数a，b的函数：

Q=Q（a，b）

若Q最小，则a，b应满足方程：

$$\frac{\partial Q(a,b)}{\partial a}=0 \qquad (11.5)$$

$$\frac{\partial Q(a,b)}{\partial b}=0 \qquad (11.6)$$

$$\frac{\partial^2 Q}{\partial a^2}>0 \qquad (11.7)$$

$$\frac{\partial^2 Q}{\partial b^2}>0 \qquad (11.8)$$

把Q的表示式代入上面两个等式，得：

$$\frac{\partial}{\partial a}\sum_{i=1}^{m}\left(y_i-a-bx_i\right)^2=\sum_{i=1}^{m}2\left(y_i-a-bx_i\right)(-1)$$

$$=-2\sum_{i=1}^{m}y_i+2ma+2b\sum_{i=1}^{m}x_i=0$$

$$\frac{\partial}{\partial b}\sum_{i=1}^{m}\left(y_i-a-bx_i\right)^2=\sum_{i=1}^{m}2\left(y_i-a-bx_i\right)(-x_i)$$

$$=-2\sum_{i=1}^{m}x_iy_i+2a\sum_{i=1}^{m}x_i+2b\sum_{i=1}^{m}x_i^{\ 2}=0$$

这是一个关于a和b的二元一次线性方程组，不难解出：

$$b=\frac{m\sum_{i=1}^{m}x_i y_i-\sum_{i=1}^{m}x_i\sum_{i=1}^{m}y_i}{m\sum_{i=1}^{m}x_i{}^2-\left(\sum_{i=1}^{m}x_i\right)^2}$$

$$a=\frac{1}{m}\sum_{i=1}^{m}y_i-\frac{b}{m}\sum_{i=1}^{m}x_i \qquad \text{（11.12）}$$

再看11.7和11.8式两个不等式是否满足。前面已经得到:

$$\frac{\partial Q(a,b)}{\partial a} = -2\sum_{i=1}^{m} y_i + 2ma + 2b\sum_{i=1}^{m} x_i \qquad \frac{\partial^2 Q}{\partial a^2} = 2m > 0$$

$$\frac{\partial Q(a,b)}{\partial b} = -2\sum_{i=1}^{m} x_i y_i + 2a\sum_{i=1}^{m} x_i + 2b\sum_{i=1}^{m} x_i{}^2$$

$$\frac{\partial^2 Q}{\partial b^2} = 2\sum_{i=1}^{m} x_i^2 > 0$$

所以上页式11.12中的a与b就是所求回归线的截距与斜率.

为了简化a与b的表示式， 再定义以下一些量：以 $\bar{x}$ 和 $\bar{y}$ 代表x_i和y_i的平均值,

$$\bar{x} = \frac{1}{m}\sum_{i=1}^{m} x_i \qquad \bar{y} = \frac{1}{m}\sum_{i=1}^{m} y_i \qquad (11.13)$$

将x_i与 $\bar{x}$ 之差称为x_i的离差, 全部x_i的离差的平方和, 称为x的离差平方和, 记作L_{xx}：(11.14)式

$$L_{xx} = \sum_{i=1}^{m}\left(x_i - \bar{x}\right)^2$$

$$L_{xx} = \sum_{i=1}^{m} \left(x_i - \bar{x} \right)^2$$

$$= \sum_{i=1}^{m} x_i^{\ 2} - \sum_{i=1}^{m} \left(2x_i \bar{x} - \bar{x}^{\ 2} \right)$$

$$= \sum_{i=1}^{m} x_i^{\ 2} - \bar{x} \sum_{i=1}^{m} \left(2x_i - \bar{x} \right)$$

$$= \sum_{i=1}^{m} x_i^{\ 2} - m\bar{x}^{\ 2} \ = \sum_{i=1}^{m} x_i^{\ 2} - \frac{1}{m} \left(\sum_{i=1`}^{m} x_i \right)^2$$

同样, y的离差平方和为:

$$L_{yy}=\sum_{i=1}^{m}\left(y_i-\bar{y}\right)^2=\sum_{i=1}^{m}y_i{}^2-\frac{1}{m}\left(\sum_{i=1`}^{m}y_i\right)^2 \qquad (11.15)$$

L_{xy}为全部x_i的离差与y_i离差乘积的总和:

$$L_{xy}=\sum_{i=1}^{m}\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)=\sum_{i=1}^{m}x_iy_i-m\bar{x}\bullet\bar{y}$$

(11.16)

将以上各式代入11.12式可得:

$$b=\frac{L_{xy}}{L_{xx}} \qquad a=\bar{y}-b\bar{x}$$

（11.17）

这就是我们编程时所依据的公式.

11．1．3．相关系数：

一般说来，最小二乘法本身并不要求所依据的数据符合某种特定的规律性。不管我们所测定的数据如何，按前述步骤，总能得到一条回归线。为了衡量回归方程与原始数据相符合的程度，提出了相关系数这一概念。

首先讨论y的变化规律。一般说来，对于一组测定值（x_i，y_i），x_i不同，则y_i也不同，即（$y_i - \bar{y}$）彼此不同。这称作y是有变动的。下面的问题是怎样描述y的变动。由式11.15可以看出，y的离差平方和L_{yy}能描述这种变动。y_i彼此差异越大，则L_{yy}就越大。

下面分析造成y变动的原因：

$$L_{yy} = \sum_{i=1}^{m}\left(y_i - \bar{y}\right)^2$$

$$L_{yy}=\sum_{i=1}^{m}\left(y_i-\bar{y}\right)^2=\sum_{i=1}^{m}\left(y_i-\hat{y}_i+\hat{y}_i-\bar{y}\right)^2$$

$$=\sum_{i=1}^{m}\left(y_i-\hat{y}_i\right)^2+\sum_{i=1}^{m}\left(\hat{y}_i-\bar{y}\right)^2 \qquad (11.20)$$

还有如下的项，我们证明它等于**0**。

$$2\sum_{i=1}^{m}\left(y_i-\hat{y}_i\right)\left(\hat{y}_i-\bar{y}\right)=2\sum_{i=1}^{m}\left(y_i-a-bx_i\right)\left(a+bx_i-\bar{y}\right)$$

$$=2\left(a-\bar{y}\right)\sum_{i=1}^{m}\left(y_i-a-bx_i\right)+2b\sum_{i=1}^{m}\left(y_i-a-bx_i\right)x_i=0$$

按照 $\frac{\partial Q(a,b)}{\partial a}=0$ $\frac{\partial Q(a,b)}{\partial b}=0$ 的要求:

$$\sum_{i=1}^{m}2\left(y_i-a-bx_i\right)\left(-1\right)=0 \qquad \sum_{i=1}^{m}2\left(y_i-a-bx_i\right)\left(-x_i\right)=0$$

上式确实为**0**。

这样，L_{yy}分成了两部分，表示**y**的变动由两种因素造成。一种是因为y_i与$\hat{y}_i$不同,一种是$\hat{y}_i$与 $\bar{y}$不同。

$$L_{yy}=\sum_{i=1}^{m}\left(y_i-\hat{y}_i\right)^2+\sum_{i=1}^{m}\left(\hat{y}_i-\bar{y}\right)^2$$

先看第二个求和。

若 $\hat{y}_i$ 与 $\bar{y}$ 不同,则L_{yy}≠0。而 $\hat{y}_i$与 $\bar{y}$ 不同,必定为y与x间存在回归关系11.1式所至。与不同x_i相应的 $\hat{y}_i$ 是由式11.19计算出来的，当然彼此应该不同；$\bar{y}$ 是一个常数，故 $\hat{y}_i$ 不会全等于 $\bar{y}$ 。

若 $\hat{y}_i$ 与 $\bar{y}$ 全相同，平行于x轴的直线，y跟x无关。所以第二个求和项反映了y与x间存在11.1式所示的回归关系对y变动的影响，这一个求和项称为**回归平方和**，用Sr表示:

$$S_r=\sum_{i=1}^{m}\left(\hat{y}_i-\bar{y}\right)^2 \qquad (11.22)$$

再看第一个求和项。从式11.4知，它就是残差平方和Q，它等于L_{yy}与S_r之差，反映的是除去y与x间存在回归关系11.1外，其它因素对y变动的贡献。这些因素主要是“测定误差”和“回归方程缺欠”。回归方程缺欠的意思是：y与x并不一定是严格的线性关系。比如弱酸的浓度 C 与 [H^+] 浓度关系。所以，y的离差平方和L_{yy}是由残差平方和Q和回归平方和S_r组成：

$$L_{yy}=Q+S_r$$

$$L_{yy}=\sum_{i=1}^{m}\left(y_i-\hat{y}_i\right)^2+\sum_{i=1}^{m}\left(\hat{y}_i-\bar{y}\right)^2$$

$L_{yy}=Q+S_r$

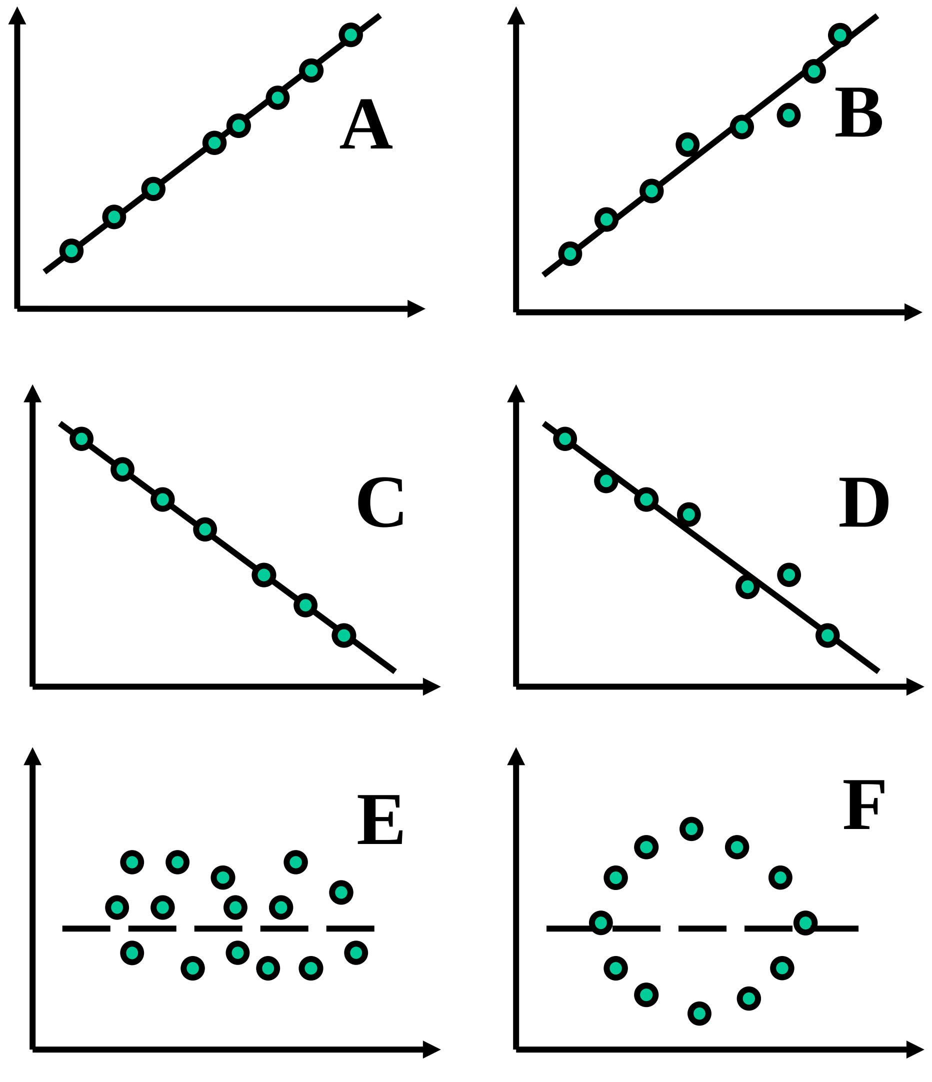
若S_r在L_{yy}中所占比例越大，则回归关系对实验点的描述就越准确；S_r所占比例愈小，测定误差及回归方程缺欠就越大。所以，S_r与L_{yy}的比值和回归方程的好坏直接相关。现定义此比值的平方根为相关系数r：

$$r=\sqrt{\frac{S_r}{L_{yy}}} \qquad (11.24)$$

用来衡量回归效果。$Q \geq 0$， $S_r \geq 0$，所以

$$0 \leq r \leq 1。$$

r=1表示数据点完全落在直线上，173页图AC线，回归效果最好。r=0表示数据点完全混乱，不存在此线性关系，如EF线。



r也可写成（**11.26**）形式。推导就不推了。

$$r = \left| \frac{L_{xy}}{\sqrt{L_{xx} L_{yy}}} \right| \qquad （11.26）$$

书上没有绝对值号。根据定义**0≤r≤1**的数，而**L**$_{\mathbf{xy}}$有正负：所以应加绝对值符号。我们编程时输出**r**要加上绝对值号。

$$\bar{x} = \frac{1}{m}\sum_{i=1}^{m} x_i \qquad \bar{y} = \frac{1}{m}\sum_{i=1}^{m} y_i$$

$$L_{xx} = \sum_{i=1}^{m} x_i^{\,2} - \frac{1}{m}\left(\sum_{i=1`}^{m} x_i\right)^2 \qquad L_{xy} = \sum_{i=1}^{m} x_i y_i - m\bar{x}\bullet\bar{y}$$

$$L_{yy} = \sum_{i=1}^{m} y_i^{\,2} - \frac{1}{m}\left(\sum_{i=1`}^{m} y_i\right)^2 \qquad b = \frac{L_{xy}}{L_{xx}}$$

$$a = \bar{y} - b\bar{x} \qquad r = \left|\frac{L_{xy}}{\sqrt{L_{xx}L_{yy}}}\right|$$

11.1.4 一元线性最小二乘法的程序：子程序

主程序中输入：

M，数据点数；

X(1)-X(M), Y(1)-Y(M)

```
11010 X1=0:X2=0:Y1=0:Y2=0:XY=0
11020 FOR I=1 TO M
11030 X1=X1+X(I):Y1=Y1+Y(I)
11040 X2=X2+X(I)*X(I):Y2=Y2+Y(I)*Y(I)
11050 XY=XY+X(I)*Y(I)
11060 NEXT I
11070 X2=X2-X1*X1/M:Y2=Y2-Y1*Y1/M
      :XY=XY-X1*Y1/M
11080 B=XY/X2:A=(Y1-B*X1)/M
11090 R=ABS(XY)/SQR(X2*Y2)
11100 RETURN
```

$$\sum_{i=1}^{m} x_i \qquad \sum_{i=1}^{m} x_i^{\ 2}$$

$$\sum_{i=1}^{m} x_i y_i$$

L_{xx}，L_{yy}，L_{xy}

子程序输出：A，B，R

11.1.5 应用示例：化学反应活化能与压力关系

定温下某反应的活化能与压力间呈直线关系：

$$E=a+bp$$

根据表中数据求回归方程系数及相关系数。

1, 40.2, 2, 40.7, 3, 40.9, 4, 41.6, 5, 41.8
6, 42.6, 7, 42.6, 8, 43.2, 9, 43.7, 10, 43.8

压力$*10^{-5}$

输出：

A=39.82

B=4.163633E-06

R=.992818

程序：

```
10 READ M
20 DIM X(M),Y(M)
30 FOR I=1 TO M:READ
X(I),Y(I):X(I)=X(I)*1E5:NEXT I
40 GOSUB 11010
50 PRINT "A=";A:PRINT "B=";B:PRINT "R=";R
60 END
200 DATA 10
210 DATA 1,40.2,2,40.7,3,40.9,4,41.6,5,41.8
220 DATA 6,42.6,7,42.6,8,43.2,9,43.7,10,43.8
11010 ----11100
```

作业：习题一，190页。

习题：一、已知铁的ΔHT与温度的数据如下：

t/°C	100	200	300	400	500	600
ΔH_T/（kJ•mol-1）	2.573	5.376	8.431	11.72	15.29	19.33

ΔH_T是TK与273.2K间的热焓变化。请用最小二乘法建立以下回归方程：

$\Delta H_T = a + bT$