数学建模中的回归分析方法

华北航天工业学院 翟秀娜

摘要 本文论述了利用回归分析建立数学模型的一般方法，并附有实例。

关键词 数学建模 回归分析 一般方法

在数学建模中，经常会遇到两类变量，一类带有“原因”的性质，称为回归变量；另一类带有“结果”的性质，称为响应变量。人们关心的问题是通过一组试验（或观测记录）数据来研究两类变量之间的关系，从而建立起一个数学模型，应用该模型去分析因果关系，或者用于预测、优化和控制等多种目的。这就是回归分析研究的主要内容。

1 用回归分析建立数学模型的一般步骤

1.1 线性回归的数学模型

设响应变量为 y，回归变量为 $x_1,x_2,\cdots\cdots x_k$，通过试验（或观测）得到 n 组数据：

$$(y_i,x_{i1},x_{i2},\cdots\cdots x_{ik}) \qquad i=1,2,\cdots\cdots n$$

假设 y 与 $x_1,x_2\cdots\cdots x_k$ 之间有如下的线性关系：

$$y=\beta_0+\beta_1x_1+\cdots\cdots+\beta_kx_k+\varepsilon \tag{1}$$

则有

$$\begin{cases} y_1=\beta_0+\beta_1x_{11}+\cdots\cdots+\beta_kx_{1k}+\varepsilon_1 \\ y_2=\beta_0+\beta_1x_{21}+\cdots\cdots+\beta_kx_{2k}+\varepsilon_2 \\ \cdots\cdots\cdots\cdots \\ y_n=\beta_0+\beta_1x_{n1}+\cdots\cdots+\beta_kx_{nk}+\varepsilon_n \end{cases} \tag{2}$$

其中 $\beta_0,\beta_1,\cdots\cdots\beta_k$ 是 $k+1$ 个待估计的参数，$\varepsilon_1,\varepsilon_2\cdots\cdots\varepsilon_n$ 是 n 个相互独立的服从 $N(0,\sigma^2)$ 的随机变量，（2）式就是线性回归的数学模型。用矩阵表示：

$$y=\begin{pmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{pmatrix} \qquad Z=\begin{pmatrix} 1\ x_{11},x_{12},\cdots\cdots x_{1k} \\ 1\ x_{21},x_{22},\cdots\cdots x_{2k} \\ \cdots\cdots\cdots\cdots \\ 1\ x_{n1},x_{n2},\cdots\cdots x_{nk} \end{pmatrix}$$

$$\beta=\begin{pmatrix} \beta_0 \\ \beta_1 \\ \cdots \\ \beta_k \end{pmatrix} \qquad \varepsilon=\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdots \\ \varepsilon_n \end{pmatrix}$$

模型（2）可写成：$Y=Z\beta+\varepsilon$ （3）

1.2　参数 β_j 的最小二乘估计

通常采用最小二乘法估计参数 $\beta_j, j = 0,1,2\cdots\cdots k$。

设 $b_0, b_1\cdots\cdots b_k$ 是 $\beta_0, \beta_1\cdots\cdots\beta_k$ 的点估计，则有

$$\hat{y} = b_0 + b_1 x_1 + \cdots\cdots + b_k x_k \tag{4}$$

对n组观测值 $(x_{i1}, x_{i2}\cdots\cdots x_{ik})$ $i = 1,2,\cdots\cdots n$ 有 y_i 的估计值 $\tilde{y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots\cdots + b_k x_{ik}$ $i = 1,2,\cdots\cdots n$ 考虑使偏差的平方和 $Q = \sum\limits_{i=1}^{n}(y_i - \tilde{y}_i)^2$ 达到最小值的 $b_0, b_1\cdots\cdots b_k$ 作为 $\beta_0, \beta_1\cdots\cdots\beta_k$ 的点估计，称 $b_j, j = 0,\cdots\cdots k$ 为 $\beta_j, j = 0,\cdots\cdots k$ 的最小二乘估计量，称(4)为线性回归方程。

1.3　回归方程的显著性检验

由上面的讨论可以看出，对任意的 n 组试验数据 $(y_i, x_{i1}, x_{i2}\cdots\cdots x_{ik})$ $i = 1,2,\cdots\cdots n$，都可以按上述方法求出一个线性回归方程。变量 y 与 $x_1, x_2\cdots\cdots x_k$ 之间是否有真的近似的线性关系呢?这就需要在求出线性回归方程后进行统计检验，即检验回归方程是否有意义。在回归分析中，这个问题称为回归的显著性检验。

我们知道，观测值 $y_1, y_2\cdots\cdots y_n$ 之间的差异是由(1)自变量 x_j $j = 1,2,\cdots\cdots k$ 取值不同引起的。(2)试验误差和模型不当引起的(自变量选取不当，如重要因素没考虑)。

为了检验这两个方面的影响哪一个是主要的。首先把它们所引起的误差从总误差中分离出来。

$$S_{总} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\tilde{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \tilde{y}_i)^2$$

其中，$\bar{y} = \frac{1}{n}(y_1 + y_2 + \cdots\cdots + y_n)$　$y_i, i = 1,2,\cdots\cdots n$ 是试验数据。

$$\tilde{y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots\cdots b_k x_{ik} \qquad i = 1,2,\cdots\cdots n$$

令 $u = \sum\limits_{i=1}^{n}(\tilde{y}_i - \bar{y})^2, Q = \sum\limits_{i=1}^{n}(y_i - \tilde{y}_i)^2$

称 u 为回归平方和，其自由度 $f_u = k$，它反映了 y 与 $x_1, x_2\cdots\cdots x_k$ 的线性关系引起的观测值的波动，因此，它的大小反映回归变量的重要程度。称 Q 为残差(剩余)平方和，其自由度 $f_Q = n - k - 1$ 它反映了试验误差以及其它未加控制的因素所引起的观测值的波动。

要检验 y 与 $x_1, x_2\cdots\cdots x_k$ 之间是否存在线性关系，需检验假设。

$$H_0: \beta_1 = \beta_2 = \cdots\cdots = \beta_k = 0$$

选统计量 $F = \dfrac{\frac{u}{k}}{\frac{Q}{n-k-1}} \sim F(k, n-k-1)$

若对于给定的 n 组试验数据 $(y_i, x_{i1}, x_{i2}\cdots\cdots x_{ik})$ 算得 $F = \dfrac{\frac{u}{k}}{\frac{Q}{n-k-1}} > F_\alpha(k, n-k-1)$

否定 H_0，这时可认为 y 对 $x_1, x_2, \cdots\cdots x_k$ 存在线性关系。若 $F < F_\alpha(k, n-k-1)$，H_0 相容，说明 y 对 $x_1, x_2\cdots\cdots x_k$ 不完全是线性关系，或还有重要的因素没考虑到。

1.4　模型的选择

(1) 当模型 $y = b_0 + b_1 x_1 + \cdots\cdots + b_k x_k$ 有显著意义 $(F > F_\alpha)$ 时，说明进入模型中的回

归变量足够多了，但不能说进入模型中的变量都是必要的，可能有多余的，为了使利用模型预报、控制时更方便，需要把多余的变量剔除。

(2) 当模型没有显著意义时($F < F_a$)，说明需要引进新的变量，或选取非线性回归方程。

2 用回归分析方法建立数学模型的实例

例 某种水泥在凝固时放出的热量 y(卡/克) 可能与下列四种化学成分有关：

x_1: $3CaO \cdot Al_2O_3$ 的成分(%) x_2: $3CaO \cdot SiO_2$ 的成分(%)

x_3: $4CaO \cdot Al_2O_3 \cdot Fe_2O_3$ 的成分(%) x_4: $2CaO \cdot SiO_3$ 的成分(%) 需要求出 y 与 x_1, x_2, x_3, x_4 之间的相依关系，并说明哪种成分是影响 y 的主要因素。

今实际测得 13 组数据如下：

	x_1	x_2	x_3	x_4	y
1	7	26	6	60	78.5
2	1	29	15	52	74.3
3	11	56	8	20	104.3
4	11	31	8	47	87.6
5	7	52	6	33	95.9
6	11	55	9	22	109.2
7	3	71	17	6	102.7
8	1	31	22	44	72.5
9	2	54	18	22	93.1
10	21	47	4	26	115.9
11	1	40	23	34	83.8
12	11	66	9	12	113.3
13	10	68	8	12	109.4

解 (1) 建立数学模型

设 $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$ 将 13 组试验数据代入上式，得到：

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \varepsilon_i \qquad i = 1, 2, \cdots\cdots 13$$

用矩阵表示：

$$Y = Z\beta + \varepsilon$$

其中 $Y=\begin{pmatrix} y_1 \\ y_2 \\ \cdots \\ y_{13} \end{pmatrix}=\begin{pmatrix} 78.5 \\ 74.3 \\ \cdots \\ 109.4 \end{pmatrix}$ $\quad Z=\begin{pmatrix} 1,x_{11},x_{12},x_{13},x_{14} \\ 1,x_{21},x_{22},x_{23},x_{24} \\ \cdots\cdots\cdots\cdots \\ 1,x_{131},x_{132},x_{133},x_{134} \end{pmatrix}=\begin{pmatrix} 1 & 7 & 26 & 6 & 60 \\ 1 & 1 & 29 & 15 & 52 \\ \cdots\cdots & & & & \\ 1 & 10 & 68 & 8 & 12 \end{pmatrix}$

$$\beta=\begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} \qquad \varepsilon=\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \vdots \\ \varepsilon_{13} \end{pmatrix}$$

$\beta_j, j=0,1,2,3,4$，是待估计的参数

$\varepsilon_i \sim N(0,\sigma^2)\ i=1,2,\cdots\cdots 13$ 且 $\varepsilon_i, i=1,2,\cdots\cdots 13$ 相互独立。

(2) 用最小二剩法估计 $\beta_0,\beta_1,\beta_2,\beta_3,\beta_4$

设 b_0,b_1,b_2,b_3,b_4 分别是 $\beta_0,\beta_1,\beta_2,\beta_3,\beta_4$ 的估计值则由

$$\begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \\ b_4 \end{pmatrix}=(Z'Z)^{-1}Z'Y \qquad \text{得} \begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \\ b_4 \end{pmatrix}=\begin{pmatrix} 62.4052 \\ 1.5511 \\ 0.5101 \\ 0.1019 \\ -0.1441 \end{pmatrix}$$

∴线性回归方程为

$$y=62.4052+1.5511x_1+0.5101x_2+0.1019x_3-0.1441x_4$$

(3) 对方程进行显著性检验 (α=0.05)

H_0:　$\beta_1=\beta_2=\beta_3=\beta_4=0$

查表 $F_0.05(k,n-k-1)=F_0.05(4,8)=3.84$

计算统计量：

$$F=\frac{u/k}{Q/n-k-1}=\frac{2667.9/4}{47.86/8}=111.5$$

∵$F>F_\alpha$　∴ 否定 H_0，认为回归方程有意义。

(4) 考虑有些变量是否可以从回归方程中去掉

在回归方程中，b_3(β_3 的估计值) 的绝对值较小，这是否表示相对来说 x_3 和 y 的线性关系不明显因而可以去掉呢？

为此，进行检验统计假设。

$H_1:\beta_3=0$

选取统计量：

$$F=\frac{b_3^2}{\frac{C_{33}Q}{n-k-1}} \sim F(1,n-k-1)=F(1,8)$$

其中 C_{33} 是矩阵$(Z'Z)^{-1}$ 中位于第 3 行第 3 列的元素，查表 $F_{0.05}$ (1.8) =5.32

计算统计量 $$f=\frac{8\times(0.1019)^2}{0.095255\times 47.86}=0.018$$

$\because F<F_{\alpha}$。$\therefore$ 接受 H_1，可以认为 $\beta_3=0$

现在考虑回归模型 $y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_4x_4+\varepsilon$ 对 β_1,β_2,β_4 是否为零，也需检验。用类似的方法可以得到 $\beta_4=0$ 的判断。最后得到的回归方程是：

$$y=52.5773+1.4683x_1+0.6623x_2$$

上述结果表明：$3CaO\cdot Al_2O_3$ 和 $3CaO\cdot SiO_2$ 是影响 y 的主要因素。

参 考 文 献

1 姜启源．数学模型．第一版，北京：高等教育出版社，1987.

（上接 36 页）

因此结深 X_j 由扩散系数 D，扩散时间 t 和比值 N_0/N_S 确定，但 D 又依赖于 t，N_0 和 N_S。为了提供工艺上的估算，先假定 D 不赖于浓度 N_0 和 N_S，当 T 确定后 X_j 仅是扩散时间 t 和比值 N_0/N_S 的函数，假如 $N_0/N_S=10^8$ 则就可由 t 和 T 值得知上述二种函数分布情况下 X_j 值。通常制作大功率平面管情况下，扩散温度范围在 900～1300℃，时间在 10′(分)～16 小时已完全足够了，在硅中硼的扩散结深 X_j 与扩散时间 t 和温度 T 的关系。下面以实例说明，求硅扩散再分布后的结深 X_j，如果 T=1200℃，t=90（分），$N_S/N_0\approx 10^2$ 得出 $X_j=4.6\mu m$ 当硅的再分布（主扩散）温度 $T=1100$℃ 时 1 次得 $X_j=30\mu m$，求扩散时间 $t=300'$（分）=5 小时。

总之，研究杂质在晶体中扩散的实验近年来取得非常大的突破，尤其是在固体的分立元器件，大规模集成电路的电子器件平面工艺中，其关键是为了获得均匀平坦和结深及浓度能得到精确控制的大面积的目的。这对当前飞跃发展的电子工业有其重要的意义。

参 考 文 献

1 L. E. Rei. 统计物理现代教程．科技出版社，1989.
2 方俊鑫、陆栋编．固体物理学．科技出版社，1992.
3 吴大猷著．理论物理（第五册）．科技出版社，1987.