（二）回归方程的稳定性

回归方程的稳定性是指回归值 $\hat{y}$ 的波动大小，波动越小，回归方程的稳定性越好。和对待一般的估计值一样，$\hat{y}$ 的波动大小用 $\hat{y}$ 的标准差 $\sigma_{\hat{y}}$ 来表示。根据随机误差传递公式及回归方程式（6-2）有

$$\sigma_{\hat{y}}^2 = \sigma_{b_0}^2 + x^2\sigma_b^2 + 2x\sigma_{b_0b} \tag{6-16}$$

式中，σ_{b_0}、σ_b 为 b_0、b 的标准差；σ_{b_0b} 为 b_0 和 b 的协方差。

设 σ 为测量数据 y 的残余标准差［有关 σ 的进一步说明见本节二、（三）及式（6-33）］，由相关矩阵式（6-6）可得

$$\sigma_{b_0}^2 = \frac{\sum_{t=1}^{N} x_t^2}{N\sum_{t=1}^{N} x_t^2 - \left(\sum_{t=1}^{N} x_t\right)^2}\sigma^2 = \left(\frac{1}{N} + \frac{\bar{x}^2}{l_{xx}}\right)\sigma^2 \tag{6-17}$$

$$\sigma_b^2 = \frac{N}{N\sum_{t=1}^{N} x_t^2 - \left(\sum_{t=1}^{N} x_t\right)^2}\sigma^2 = \frac{\sigma^2}{l_{xx}} \tag{6-18}$$

$$\sigma_{b_0b} = \frac{-\sum_{t=1}^{N} x_t}{N\sum_{t=1}^{N} x_t^2 - \left(\sum_{t=1}^{N} x_t\right)^2} = -\frac{\bar{x}}{l_{xx}}\sigma^2 \tag{6-19}$$

将式（6-17）、式（6-18）、式（6-19）代入式（6-16）得

$$\sigma_{\hat{y}}^2 = \left(\frac{1}{N} + \frac{\bar{x}^2}{l_{xx}}\right)\sigma^2 + x^2\frac{\sigma^2}{l_{xx}} - 2x\frac{\bar{x}}{l_{xx}}\sigma^2$$

$$= \left(\frac{1}{N} + \frac{(x-\bar{x})^2}{l_{xx}}\right)\sigma^2 \tag{6-20}$$

或

$$\sigma_{\hat{y}} = \sigma\sqrt{\frac{1}{N} + \frac{(x-\bar{x})^2}{l_{xx}}} \tag{6-21}$$

由式（6-21）可见，回归值的波动大小不仅与残余标准差 σ 有关，而且还取决于实验次数 N 及自变量 x 的取值范围。N 越大，x 的取值范围越小，回归值 $\hat{y}$ 的精度越高。

二、回归方程的方差分析及显著性检验

回归方程式（6-15）求出来了，但它是否有实际意义呢？这里有两个问题需要解决：其一，就这种求回归直线的方法本身而言，对任何两个变量 x 和 y 的一组数据 (x_t, y_t)，$t=1, 2, \cdots, N$，都可以用最小二乘法给它们拟合一条直线。要知道这条直线是否基本上符合 y 与 x 之间的客观规律，这就是回归方程的显著性检验要解决的问题。其二，由于 x 与 y 之间是相关关系，知道了 x 值，并不能精确地知道 y 值。那么，用回归方程，根据自变量 x 值预报（或控制）因变量 y 值，其效果如何？这就是回归方程的预报精度问题。为此，必须对回归问题作进一步分析。现介绍一种常用的方差分析法，其实质是对 N 个观测值与其算术

平均值之差的平方和进行分解，将对 N 个观测值的影响因素从数量上区别开，然后用 F 检验法对所求回归方程进行显著性检验。

（一）回归问题的方差分析

观测值 y_1，y_2，…，y_N 之间的差异（称变差），是由两个方面原因引起的：①自变量 x 取值的不同；②其他因素（包括实验误差）的影响。为了对回归方程进行检验，首先必须把它们所引起的变差从 y 的总变差中分解出来（见图 6-2）。

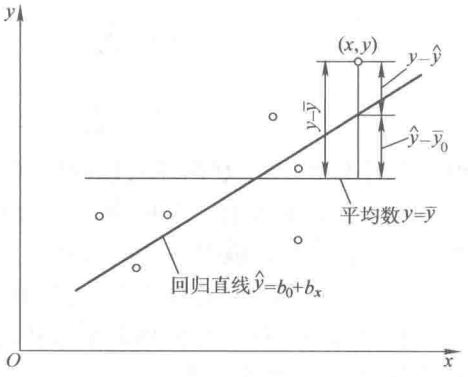

图 6-2

N 个观测值之间的变差，可用观测值 y 与其算术平均值 $\bar{y}$ 的离差平方和来表示，称为总的离差平方和，记作

$$S = \sum_{t=1}^{N} (y_t - \bar{y})^2 = l_{yy} \qquad (6\text{-}22)$$

因为

$$S = \sum_{t=1}^{N} (y_t - \bar{y})^2 = \sum_{t=1}^{N} [(y_t - \hat{y}_t) + (\hat{y}_t - \bar{y})]^2$$

$$= \sum_{t=1}^{N} (\hat{y}_t - \bar{y})^2 + \sum_{t=1}^{N} (y_t - \hat{y}_t)^2 + 2\sum_{t=1}^{N} (y_t - \hat{y}_t)(\hat{y}_t - \bar{y})$$

可以证明，交叉项

$$\sum_{t=1}^{N} (y_t - \hat{y})(\hat{y}_t - \bar{y}) = 0$$

因此总的离差平方和可以分解为两个部分，即

$$\sum_{t=1}^{N} (y_t - \bar{y})^2 = \sum_{t=1}^{N} (\hat{y}_t - \bar{y})^2 + \sum_{t=1}^{N} (y_t - \hat{y}_t)^2 \qquad (6\text{-}23)$$

或者写成

$$S = U + Q \qquad (6\text{-}24)$$

式（6-23）、式（6-24）中右边第一项，即

$$U = \sum_{t=1}^{N} (\hat{y}_t - \bar{y})^2 \qquad (6\text{-}25)$$

称为回归平方和，它反映了在 y 总的变差中由于 x 和 y 的线性关系而引起 y 变化的部分。因此回归平方和也就是考虑了 x 与 y 的线性关系部分在总的离差平方和 S 中所占的成分，以便从数量上与 Q 值相区分。

式（6-24）中右边第二项，即

$$Q = \sum_{t=1}^{N} (y_t - \hat{y}_t)^2 \qquad (6\text{-}26)$$

称为残余平方和，即所有观测点到回归直线的残余误差 $y_t - \hat{y}_t$ 的平方和。它是除了 x 对 y 的线性影响之外的一切因素（包括实验误差、x 对 y 的非线性影响以及其他未加控制的因素）对 y 的变差作用，这部分的变差是仅考虑 x 与 y 的线性关系所不能减少的部分。

这样，通过平方和分解式（6-23）就把对 N 个观测值的两种影响从数量上区分开来。U

和 Q 的具体计算通常并不是按它们的定义式（6-25）和式（6-26）进行，而是按下式计算：

$$U = \sum_{i=1}^{N} (\hat{y}_i - \bar{y})^2 = \sum_{i=1}^{N} (b_0 + bx_i - b_0 - b\bar{x})^2$$

$$= b^2 \sum_{i=1}^{N} (x_i - \bar{x})^2 = b \sum_{i=1}^{N} (x_i - \bar{x})(\hat{y}_i - \bar{y}) = bl_{xy} \tag{6-27}$$

$$Q = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 = S - U = l_{yy} - bl_{xy} \tag{6-28}$$

因此，在计算 S、U、Q 时就可以利用回归系数计算过程中的一些结果。

对每个平方和都有一个称为“自由度”的数据跟它相联系。如果总的离差平方和是由 N 项组成，其自由度就是 $N-1$。如果一个平方和是由几部分相互独立的平方和组成，则总的自由度等于各部分自由度之和。正如总的离差平方和在数值上可以分解成回归平方和与残余平方和两部分一样，总的离差平方和的自由度 ν_S 也等于回归平方和的自由度 ν_U 与残余平方和的自由度 ν_Q 之和，即

$$\nu_S = \nu_U + \nu_Q \tag{6-29}$$

在回归问题中，$\nu_S = N-1$，而 ν_U 对应于自变量的个数，因此在一元线性回归问题中 $\nu_U = 1$，故根据式（6-29），Q 的自由度 $\nu_Q = N-2$。

（二）回归方程显著性检验

由回归平方和与残余平方和的意义可知，一个回归方程是否显著，也就是 y 与 x 的线性关系是否密切，取决于 U 及 Q 的大小，U 越大、Q 越小说明 y 与 x 的线性关系越密切。回归方程显著性检验通常采用 F 检验法，因此要计算统计量 F：

$$F = \frac{U/\nu_U}{Q/\nu_Q} \tag{6-30}$$

对一元线性回归

$$F = \frac{U/1}{Q/(N-2)} \tag{6-31}$$

再查附表 4 F 分布表。F 分布表中的两个自由度 ν_1 和 ν_2 分别对应于式（6-30）中的 ν_U 和 ν_Q，即式（6-31）中的 1 和 $N-2$。检验时，一般需查出 F 分布表中对三种不同显著性水平 a 的数值，设记为 $F_a(1, N-2)$，将这三个数与由式（6-31）计算的 F 值进行比较，若 $F \geqslant F_{0.01}(1, N-2)$，则认为回归是高度显著的（或称在 0.01 水平上显著）；若 $F_{0.05}(1, N-2) \leqslant F < F_{0.01}(1, N-2)$，则称回归是显著的（或称在 0.05 水平上显著）；若 $F_{0.10}(1, N-2) \leqslant F < F_{0.05}(1, N-2)$，则称回归在 0.1 水平上显著；若 $F < F_{0.10}(1, N-2)$，一般认为回归不显著，此时，y 对 x 的线性关系就不密切。

（三）残余方差与残余标准差

残余平方和 Q 除以它的自由度 ν_Q 所得商

$$\sigma^2 = \frac{Q}{N-2} \tag{6-32}$$

称为残余方差，它可以看作在排除了 x 对 y 的线性影响后（或者当 x 固定时），衡量 y 随机波动大小的一个估计量。残余方差的正平方根

$$\sigma = \sqrt{\frac{Q}{N-2}} \tag{6-33}$$

称为残余标准差，与 σ^2 的意义相似，它可以用来衡量所有随机因素对 y 的一次性观测的平均变差的大小，σ 越小，回归直线的精度越高。当回归方程的稳定性较好时，σ 可作为应用回归方程时的精度参数。

（四）方差分析表

上述把平方和及自由度进行分解的方差分析所有结果可归纳在一个简单的表格中，这种表称为方差分析表，见表6-3。

表 6-3

来　源	平方和	自由度	方　差	F	显著性
回　归	$U=bl_{xy}$	1		$F=\dfrac{U/1}{Q/(N-2)}$	—
残　余	$Q=l_{yy}-bl_{xy}$	$N-2$	$\sigma^2=Q/(N-2)$		
总　计	$S=l_{yy}$	$N-1$	—	—	—

例 6-2　在例6-1电阻对温度的回归中，由表6-2及表6-3可得表6-4的方差分析结果。

表 6-4

来　源	平方和/Ω^2	自由度	方差/Ω^2	F	显著性
回　归	60.574	1	—	1.18×10^3	$\alpha=0.01$
残　余	0.257	5	0.0514	—	—
总　计	60.831	6	—	—	—

显著性一栏中的 $\alpha=0.01$，表明前面所得的回归方程式（6-15）在 $\alpha=0.01$ 水平上显著，即可信赖程度为99%以上，这是高度显著的。

利用回归方程，可以在一定显著性水平 α 上，确定与 x 相对应的 y 的取值范围。反之，若要求观测值 y 在一定的范围内取值，利用回归方程可以确定自变量 x 的控制范围。

三、重复实验情况

应该指出，用残余平方和检验回归平方和所作出的“回归方程显著”这一判断，只表明相对于其他因素及试验误差来说，因素 x 的一次项对指标 y 的影响是主要的，但它并没有告诉我们：影响 y 的除 x 外，是否还有一个或几个不可忽略的其他因素，以及 x 和 y 的关系是否确实为线性。换言之，在上述意义下的回归方程显著，并不一定表明这个回归方程是拟合得很好的。其原因是由于残余平方和中除包括实验误差外，还包括了 x 和 y 线性关系以外的其他未加控制的因素的影响。为了检验一个回归方程拟合得好坏，可以做些重复实验，从而获得误差平方和 Q_E 和失拟平方和 Q_L（它反映了非线性及其他未加控制的因素的影响），用误差平方和对失拟平方和进行 F 检验，就可以确定回归方程拟合得好坏。

设取 N 个实验点，每个实验点都重复 m 次实验，此时各种平方和及其相应的自由度可按下列各式计算：

$$S = U + Q_L + Q_E, \quad \nu_S = \nu_U + \nu_L + \nu_E \tag{6-34}$$