“回归分析”与“时間序列分析”

統計預报短訓班

一九七三年四月

回　归　分　析

目　录

回归分析

§1. 回归模型

已知量 y 随着 m 个自变量 $x_1, x_2, \cdots, x_m$ 的变化而变化，并有如下线性关系：

$$y=\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots\cdots+\beta_m x_m+\varepsilon \qquad (1.1)$$

这一式子通常称为 y 关于 $x_1, x_2, \cdots, x_m$ 的"回归方程"，式中的 $\beta_0, \beta_1, \beta_2, \cdots, \beta_m$ 是待定参数，也称"回归系数"，ε 是随机变量，也称"误差"，这是因为 ε 可以看作是 y 中无法用 $x_1, x_2, \cdots, x_m$ 表示的各种复杂因素或随机因素组成的误差。

在实际问题中，我们常假设量 $x_1, x_2, \cdots, x_m$；y 有 N 组观测数据：

$$(x_{t1}, x_{t2}, \cdots\cdots, x_{tm}, y_t) \qquad (t=1,2,\cdots\cdots,N)$$

它们满足回归方程，即满足下面的 N 个式子：

$$y_t=\beta_0+\beta_1 x_{t1}+\beta_2 x_{t2}+\cdots\cdots+\beta_m x_{tm}+\varepsilon_t \qquad (1.2)$$

$$(t=1,2,\cdots\cdots,N)$$

这时随机变量 ε_t $(t=1,2,\cdots,N)$ 表示各次观测的误差，通常它们满足如下假设：

① 观测误差没有系统性的偏倚现象。即 ε_t 的数学期望全为零

$$E(\varepsilon_t)=0 \qquad (t=1,2,\cdots,N) \qquad (1.3)$$

② 这些观测是相互独立的，并且具有相同的精度。因此 ε_t 的协方差可表为

$$V(\varepsilon_t, \varepsilon_\tau)=\delta_{t\tau}\sigma^2 \qquad (t,\tau=1,2,\cdots,N) \qquad (1.4)$$

这里的δ_{tt}是Kronecker记号，当$t=\tau$时，它等于1，$t\neq\tau$时等于0。σ^2是公共方差。当$t=\tau$时，$V(\varepsilon_t,\varepsilon_t)$就是$\varepsilon_t$的方差，可简记为$V(\varepsilon_t)$。

③观测误差服从正态分布。有了这个假设，就可以顺利地使用一些统计检验方法。

如果用记号$N(\mu,\sigma^2)$表示数学期望为μ，方差为σ^2的正态分布，则上述三类假设可简单地概括为"误差ε_t相互独立地服从正态分布$N(0,\sigma^2)$"。

在回归分析中，需要解决的主要问题是：

①根据给出的N组观测数据，确定各回归系数β_i的估值b_i $(i=0,1,2,\cdots,m)$。同时对各b_i作统计检验，以便指出这些估值的可靠性。

②对自变量$(X_1,X_2,\cdots,X_m)$的任一组观测值，预报量y的取值，并且指出预报的精度。这就需要给出误差ε的未知方差σ^2的估值。

回归分析在一般数据处理，经验公式，曲线拟合，以及各类预报问题中都有广泛的应用。但是随着应用领域的不同，回归计算的重点也常常有差别：例如在一般经验公式与曲线拟合中，人们往往只注意求出各b_i值；在各类预报问题中，人们显然关心回归方程在预报中的精度（主要是σ^2的大小）。此外为建立较好的回归方程，人们往往需要对所有可供使用的自变量作统计检验（即对b_i作显著性检验），以判别哪些是重要的，哪些是不重要的，从而为变量的挑选提供依据。

本文将从应用的角度简要地讨论上述各类问题，重点放在预报问题中各预报因子的挑选上，特别将介绍在筛选变量中已有广泛应用的"正交筛选法"和"逐步回归法"。在内容安排上，为使重点突出，我们将把部分内容以补充"说明"形式给出在各节之后。

本节补充说明：

〔説明1·1〕：在上述回归模型中，自变量 X_i 被認为是确定性变量，而观测誤差 ε_t 是相互独立地服从正态分佈 $N(0,\sigma^2)$，因此 y_t 也是相互独立地服从正态分佈的随机变量，其數学期望和方差分別是：

$$E(y_t)=\beta_0+\beta_1X_{t1}+\beta_2X_{t2}+\cdots\cdots+\beta_mX_{tm} \qquad (1\cdot5)$$

$$V(y_t)=\sigma^2 \qquad (1\cdot6)$$

（1·5）或

$$E(y)=\beta_0+\beta_1X_1+\beta_2X_2+\cdots\cdots+\beta_mX_m \qquad (1\cdot7)$$

可以代替（1·2）或（1·1），它们都称为回归方程。从这一形式的回归方程开始討論，就可以不涉及誤差 ε 的问题，很多統計书籍都这样做。

〔説明1·2〕：上述回归模型，严格地說，应称为"多元线性正态回归模型"。所謂"多元"，是指量 y 依赖不止一个的自变量（$m>1$）；所謂"线性"，是指回归方程是关于自变量 X_i 的线性函數；所謂"正态"，是指 ε 或 y 是正态分佈的随机变量。今后我们可以看到，线性的含义可以推广为只是关于 β_i 的線性函數，而不一定是 X_i 的线性函數。

〔説明1·3〕：在一般回归分析中，我们总是假設观测次數比回归系數的个數大（即 $N>m+1$），而且任一自变量不能用其它自变量线性表出。在正交篩选和逐步回归中，后一假設不是必需的。

§2. 回归系數的最小二乘估計

假設我们有某一种方法可以得到 b_i 的估值，则 y 的观测值可表为

$$y_t=b_0+b_1X_{t1}+b_2X_{t2}+\cdots\cdots+b_mX_{tm}+e_t \qquad (2\cdot1)$$

上式也称回归方程，或加"经验"二字，用以与（1·1）的"理論"形式相区别。这里 e_t 是误差 ε_t 的估值，常称为"残差"或"剩余"。令 y_t^* 为 y_t 的估值，即

$$y_t^* = b_0 + b_1 x_{t1} + b_2 x_{t2} + \cdots\cdots + b_m x_{tm} \tag{2·2}$$

$$e_t = y_t - y_t^* \tag{2·3}$$

b_i 的最优估值可以用最小二乘法得到，即 b_i 的确定将使残差平方和 Q 达到最小

$$Q = \sum_t e_t^2 = \sum_t (y_t - y_t^*)^2 \qquad \text{〔注〕}$$

$$= \sum_t [y_t - (b_0 + b_1 x_{t1} + b_2 x_{t2} + \cdots\cdots + b_m x_{tm})]^2 \tag{2·4}$$

由数学分析的极值原理，知各 b_i 满足如下方程组

$$\begin{cases} \dfrac{\partial Q}{\partial b_0} = 0 \\ \dfrac{\partial Q}{\partial b_1} = 0 \\ \vdots \\ \dfrac{\partial Q}{\partial b_m} = 0 \end{cases}$$

即 Q 对各 b_i 的偏微商全为零。

由 $\dfrac{\partial Q}{\partial b_0} = 0$ 得

$$\sum_t [y_t - (b_0 + b_1 x_{t1} + b_2 x_{t2} + \cdots\cdots + b_m x_{tm})](-1) = 0 \tag{2·6}$$

或 $\sum_t y_t - (N b_0 + b_1 \sum_t x_{t1} + b_2 \sum_t x_{t2} + \cdots\cdots + b_m \sum_t x_{tm}) = 0$

〔注〕：求和号 $\sum$ 下常添有下标 t 或 i，其中 t（或 t）表示从 1 到 N 的和，即 $\sum\limits_{t=1}^{N}$；i（或 j）表示从 1 到 m 的和，即 $\sum\limits_{i=1}^{m}$

记观测值的平均值为

$$\begin{cases}\overline{x}_i=\frac{1}{N}\sum_t x_{ti} & (i=1,2,\cdots,m)\\ \overline{y}=\frac{1}{N}\sum_t y_t\end{cases} \tag{2·7}$$

(2·6)可表为

$$b_0=\overline{y}-(b_1\overline{x}_1+b_2\overline{x}_2+\cdots+b_m\overline{x}_m) \tag{2·8}$$

以(2·8)代入(2·5)，并令

$$\begin{cases}x'_{ti}=x_{ti}-\overline{x}_i & (i=1,2,\cdots,m,\\ y'_t=y_t-\overline{y} & t=1,2,\cdots,N)\end{cases} \tag{2·9}$$

得 $$Q=\sum_t[y'_t-(b_1x'_{t1}+b_2x'_{t2}+\cdots+b_mx'_{tm})]^2 \tag{2·10}$$

这时(2·5)可化为

$$\begin{cases}\frac{\partial Q}{\partial b_1}=\frac{1}{2}\sum_t[y'_t-(b_1x'_{t1}+b_2x'_{t2}+\cdots+b_mx'_{tm})](-x'_{t1})=0\\ \frac{\partial Q}{\partial b_2}=\frac{1}{2}\sum_t[y'_t-(b_1x'_{t1}+b_2x'_{t2}+\cdots+b_mx'_{tm})](-x'_{t2})=0\\ -\ -\ -\ -\ -\ -\ -\ -\ -\ -\ -\ -\ -\ -\ -\\ \frac{\partial Q}{\partial b_m}=\frac{1}{2}\sum_t[y'_t-(b_1x'_{t1}+b_2x'_{t2}+\cdots+b_mx'_{tm})](-x'_{tm})=0\end{cases} \tag{2·11}$$

经整理并引入如下记号

$$\begin{cases}S_{ij}=S_{ji}=\sum_t x'_{ti}x'_{tj}=\sum_t(x_{ti}-\overline{x}_i)(x_{tj}-\overline{x}_j)\\ S_{iy}=\sum_t x'_{ti}y'_t=\sum_t(x_{ti}-\overline{x}_i)(y_t-\overline{y})\end{cases} \tag{2·12}$$

$(i,j=1,2,\cdots,m)$

(2·11)可化为

$$\begin{cases}S_{11}b_1+S_{12}b_2+\cdots+S_{1m}b_m=S_{1y}\\ S_{21}b_1+S_{22}b_2+\cdots+S_{2m}b_m=S_{2y}\\ -\ -\ -\ -\ -\ -\ -\ -\ -\ -\\ S_{m1}b_1+S_{m2}b_2+\cdots+S_{mm}b_m=S_{my}\end{cases} \tag{2·13}$$

上式常称为"正规方程"，在观测数据给定的情况下，S_{ij} 与 S_{iy} 可由（2·12）一一算出，因此正规方程是以 $b_1, b_2, \cdots\cdots b_m$ 为未知数的 m 个联立方程式，例如可用消去法求解（详见§5）。解出 $b_1, b_2, \cdots\cdots, b_m$ 后，由（2·8）便可得 b_0。

本节补充说明：

〔说明2·1〕：在上述回归模型中，b_0 与 $b_1, b_2, \cdots, b_m$ 出现的形式不一致，给处理带来麻烦。如果引入一个形式变量 x_0，其观测值恒为1，即 $x_{t0} \equiv 1$，则回归方程可写作

$$y_t = b_0 x_{t0} + b_1 x_{t1} + \cdots\cdots + b_m x_{tm} + e_t \tag{2.14}$$

类似（2·11）～（2·13）的推导，可得

$$\frac{\partial Q}{\partial b_i} = \frac{1}{2}\sum_t [y_t - (b_0 x_{t0} + b_1 x_{t1} + \cdots\cdots + b_m x_{tm})](-x_{ti}) = 0$$
$$(i = 0, 1, 2, \cdots, m) \qquad (2.15)$$

令
$$\begin{cases} a_{ij} = a_{ji} = \sum\limits_t x_{ti} x_{tj} \\ a_{iy} = \sum\limits_t x_{ti} y_t \qquad (i, j = 0, 1, 2, \cdots\cdots, m) \end{cases} \tag{2.16}$$

可得正规方程

$$\begin{cases} a_{00} b_0 + a_{01} b_1 + \cdots\cdots + a_{0m} b_m = a_{0y} \\ a_{10} b_0 + a_{11} b_1 + \cdots\cdots + a_{1m} b_m = a_{1y} \\ - - - - - - - - - - \\ a_{m0} b_0 + a_{m1} b_1 + \cdots\cdots + a_{mm} b_m = a_{my} \end{cases} \tag{2.17}$$

上式是以 $b_0, b_1, \cdots, b_m$ 为未知数的 $m+1$ 个联立方程式，解之便得到全部 b_i。显然，这样的处理方法要比本节正文中更为简便，不少统计著作就是这样做的。

〔说明2·2〕：由 $b_0 = \bar{y} - (b_1\bar{x}_1 + b_2\bar{x}_2 + \cdots + b_m\bar{x}_m)$（见(2·8)）可知当各自变量的观测数据的平均值 $\bar{x}_i$，$\bar{y}$ 全为零，则 $b_0 = 0$，这时可以认为回归方程中根本没有 b_0 这一项，即

$$y_t = b_1 x_{t1} + b_2 x_{t2} + \cdots\cdots + b_m x_{tm} + e_t \tag{2.18}$$

在实际问题中 $\overline{x_i}$，$\overline{y}$ 全为零是不可能的，但是可以通过变量变换

$$x'_{ti}=x_{ti}-\overline{x_i} \qquad y'_t=y_t-\overline{y}$$

使新的变量 x'_i，y' 的平均值全为零，于是仍然可以得到不含有 b_0 的回归方程

$$y'_t=b'_1x'_{t1}+b'_2x'_{t2}+\cdots\cdots+b'_mx'_{tm}+e_t \tag{2·19}$$

这就是正文中的处理方法，从正文的推导来看，（见(2·8)～～(2·10)）这里的 $b'_1, b'_2, \cdots\cdots, b'_m$ 分别等于原回归模型(有 b_0 的)

$$y_t=b_0+b_1x_{t1}+b_2x_{t2}+\cdots\cdots+b_mx_{tm}+e_t$$

中的 $b_1, b_2, \cdots\cdots, b_m$。

〔说明2·3〕：本节正文中把 b_0 与 $b_1, b_2, \cdots\cdots, b_m$ 分开来处理的方法，虽不及〔说明2·1〕的简便，但可以附带得到一些重要信息，因此也有很多统计书作这样处理。事实上，各变量之间的协方差（这时暂把各 x_i 看作是随机变量）可用 $\frac{1}{N-1}S_{ij}$ 估计，而相关系数 r_{ij} 可表为

$$r_{ij}=\frac{S_{ij}}{\sqrt{S_{ii}}\sqrt{S_{jj}}}=\frac{\sum_t(x_{ti}-\overline{x_i})(x_{tj}-\overline{x_j})}{\sqrt{\sum_t(x_{ti}-\overline{x_i})^2}\sqrt{\sum_t(x_{tj}-\overline{x_j})^2}} \tag{2·20}$$

在实际计算中，考虑到 r_{ij} 是无量纲的，各 r_{ij} 的差异要比 S_{ij} 小，如果用 r_{ij} 代替 S_{ij} 来求解正规方程的话，舍入误差一般要小一些，在电子计算机上计算时更是这样。要注意的是，这样得到的解 b'_i 与原来的解 b_i 不同，其关係式是

$$b'_i=\frac{\sqrt{S_{ii}}}{\sqrt{S_{yy}}}b_i \qquad (i=1,2,\cdots\cdots,m) \tag{2·21}$$

其中

$$S_{yy}=\sum_t(y'_t)^2=\sum_t(y_t-\overline{y})^2 \tag{2·22}$$

这一事实可作如下验证：

(1) 在正规方程（2·13）中，第 i 个方程的各个系数同除以 $\sqrt{S_{ii}}\sqrt{S_{yy}}$ 其解不变，这时有

$$\frac{S_{i1}}{\sqrt{S_{ii}}\sqrt{S_{yy}}}b_1+\frac{S_{i2}}{\sqrt{S_{ii}}\sqrt{S_{yy}}}b_2+\cdots\cdots+\frac{S_{im}}{\sqrt{S_{ii}}\sqrt{S_{yy}}}b_m=\frac{S_{iy}}{\sqrt{S_{ii}}\sqrt{S_{yy}}}$$

$$(i=1,2,\cdots\cdots,m)$$

(2) 上述方程中的第 i 个系数的分子分母同乘以 $\sqrt{S_{ii}}$，其解不变，这时有

$$\frac{S_{i1}}{\sqrt{S_{ii}}\sqrt{S_{11}}}\left(\frac{\sqrt{S_{11}}}{\sqrt{S_{yy}}}b_1\right)+\frac{S_{i2}}{\sqrt{S_{ii}}\sqrt{S_{22}}}\left(\frac{\sqrt{S_{22}}}{\sqrt{S_{yy}}}b_2\right)+\cdots\cdots$$

$$\cdots\cdots+\frac{S_{im}}{\sqrt{S_{ii}}\sqrt{S_{mm}}}\left(\frac{\sqrt{S_{mm}}}{\sqrt{S_{yy}}}b_m\right)=\frac{S_{iy}}{\sqrt{S_{ii}}\sqrt{S_{yy}}}$$

$$(i=1,2,\cdots\cdots,m)$$

由（2·20）和（2·21）得

$$r_{i1}b'_1+r_{i2}b'_2+\cdots\cdots+r_{im}b'_m=r_{iy} \qquad (2\cdot 23)$$

$$(i=1,2,\cdots\cdots,m)$$

或

$$\begin{cases} r_{11}b'_1+r_{12}b'_2+\cdots\cdots+r_{1m}b'_m=r_{1y} \\ r_{21}b'_1+r_{22}b'_2+\cdots\cdots+r_{2m}b'_m=r_{2y} \\ -\ -\ -\ -\ -\ -\ -\ -\ -\ - \\ r_{m1}b'_1+r_{m2}b'_2+\cdots\cdots+r_{mm}b'_m=r_{my} \end{cases} \qquad (2\cdot 24)$$

〔说明 2·4〕：关于 $\overline{x}_i$ 与 S_{ij} 的计算问题。

① $\overline{x}_i$ 的计算（包括 $\overline{y}$）：

$$\overline{x}_i=\frac{1}{N}\sum_t x_{ti}=\frac{1}{N}\sum_t(x_{ti}-C_i)+C_i \qquad (2\cdot 25)$$

其中 C_i 是与 t 无关的常量。因此当某个 x_i 的观测值的前几位几乎相同时，可先减去其公共部分 C_i 再求和，最后才把 C_i 加回去。这种计算方法对于手算非常适合，可以节省计算时间。对电子计算机来说，这一算法虽然不能起到节省计算时间的作用，却可以提高计算精度。这时作为第一步，先按下式算出 C_i，

$$C_i=\frac{1}{N}\sum_t x_{ti} \qquad (2\cdot 26)$$

C_i 已是 X_i 的均值，但可能有较大的舍入误差，为此作第二步计标

$$\overline{X}_i=\frac{1}{N}\sum_t(X_{ti}-C_i)+C_i \qquad (2.27)$$

其中右端的求和项集中了均值 C_i 的计标过程中的主要舍入误差，因此这样的"二步标法"可以得到更精确的 $\overline{X}_i$，特别当观测数据的位数较多，而机器的字长较短时更是如此。〔文献12〕、〔文献13〕

② S_{ij} 的计标（包括 S_{iy} 与 S_{yy}）：

$$S_{ij}=\sum_t(X_{ti}-\overline{X}_i)(X_{tj}-\overline{X}_j)$$

$$=\sum_t X_{ti}X_{tj}-N\overline{X}_i\overline{X}_j \qquad (2.28)$$

计标时用第二个等式可以节省时间，但是用电子计标机计标时，第一个等式的精度要比第二个高。〔文献12〕、〔文献13〕

〔说明2.5〕：引入矩阵记号

$$S=\begin{pmatrix} S_{11}S_{12}\cdots S_{1m}\\ S_{21}S_{22}\cdots S_{2m}\\ \\ S_{m1}S_{m2}\cdots S_{mm}\end{pmatrix} \quad b=\begin{pmatrix} b_1\\ b_2\\ \vdots\\ b_m\end{pmatrix} \quad g=\begin{pmatrix} S_{1y}\\ S_{2y}\\ \vdots\\ S_{my}\end{pmatrix} \qquad (2.29)$$

则正规方程（2.13）可表为

$$S\ b=g \qquad (2.30)$$

在〔说明1.3〕的假設下，S 矩阵满秩（其行列式的值不等于零），b 有唯一解，这时 S 矩阵的逆矩阵 S^{-1} 存在，记 S^{-1} 为 C，则

$$C=\begin{pmatrix} C_{11}\,C_{12}\cdots C_{1m}\\ C_{21}\,C_{22}\cdots C_{2m}\\ -\ -\ -\ -\ -\ -\\ C_{m1}C_{m2}\cdots C_{mm}\end{pmatrix} \qquad (2.31)$$

b可表为

$$b=S^{-1}g \qquad 或 \qquad b=Cg$$

即 $$b_i=\sum_j C_{ij}g_j \tag{2.32}$$

〔说明2.6〕：对于经验公式与曲线拟合一类问题，人们只关心 b_i 的计算，因此其后的内容（特别是§3）也许是多余的。从计算观点看，确定 b_i 的过程与解线性矛盾方程组的最小二乘法无异。

§3. 回归问题的方差分析与统计检验

(1) 回归效果的分析——方差分析

y 的离差平方和 S_{yy} 可表为

$$S_{yy}=\sum_t(y_t-\bar{y})^2=\sum_t[(y_t-y_t^*)+(y_t^*-\bar{y})]^2$$

$$=\sum_t(y_t-y_t^*)^2+\sum_t(y_t^*-\bar{y})^2 \quad 〔注1〕 \tag{3.1}$$

上式中右端第一部分是前面已定义过的残差平方和 Q，第二部分称为回归平方和，记为 U 〔注2〕。

〔注1〕：这是因为由（2.8），（2.9）有

$$\sum_t(y_t-y_t^*)(y_t^*-\bar{y})$$

$$=\sum_t[y'_t-(b_1x'_{t1}+\cdots\cdots+b_mx'_{tm})](b_1x'_{t1}+\cdots\cdots+b_mx'_{tm})$$

$$=b_1\sum_t[y'_t-(b_1x'_{t1}+\cdots\cdots+b_mx'_{tm})](x'_{t1})+\cdots\cdots$$
$$\cdots\cdots+b_m\sum_t[y'_t-(b_1x'_{t1}+\cdots\cdots+b_mx'_{tm})](x'_{tm})=0$$

最后一个等号是因为式中的每一个求和项都为零（见(2·11)）。

〔註2〕：記估計值 y_t^* 之均值为

$$\overline{y}^*=\frac{1}{N}\sum_t y_t^*,$$

由（2·8）易知 $\overline{y}^*=\overline{y}$，故

$$U=\sum_t(y_t^*-\overline{y})^2=\sum_t(y_t^*-\overline{y}_t^*)^2$$

是回归方程的离差平方和。

即总的离差平方和 S_{yy} 可分解为残差平方和与回归平方和两部分，可記为

$$S_{yy}=Q+U \qquad (3·2)$$

Q的計算可用前面的定义（2·4）或（2·10）进行，当 S_{yy} 已算出的情况下，用下式計算则更方便

$$Q=S_{yy}-\sum_i b_i S_{iy} \qquad (3·3)$$

这是因为

$$Q=\sum_t(y_t-y_t^*)^2=\sum_t(y_t-y_t^*)[(y_t-\overline{y})-(y_t^*-\overline{y})]$$

$$=\sum_t(y_t-y_t^*)(y_t-\overline{y})-\sum_t(y_t-y_t^*)(y_t^*-\overline{y})$$

上式第二部分由上页〔註1〕知其为零，故

$$Q=\sum_t(y_t-y_t^*)(y_t-\overline{y})=\sum_t[(y_t-\overline{y})-(y_t^*-\overline{y})](y_t-\overline{y})$$

$$=\sum_t(y_t-\overline{y})^2-\sum(b_1x'_{t1}+\cdots+b_m x'_{tm})y'_t$$

$$=S_{yy}-(b_1S_{1y}+b_2S_{2y}+\cdots\cdots+b_mS_{my})$$

对比（3·2）与（3·3），易知回归平方和也可記为

$$U=\sum_i b_iS_{iy} \qquad (3·4)$$

对给定的观测值而言，S_{yy}是不变的，因此$Q+U$是个常量，Q大则U小，反之Q小则U大，因此Q与U都可以用于衡量回归的效果，但是Q或U与S_{yy}一样，都是有量纲的，即与y的单位有关，大小不好判别。为此，可采用如下定义的无量纲指标：

$$R^2=\frac{U}{S_{yy}} \quad \text{或} \quad R=\sqrt{1-\frac{Q}{S_{yy}}} \tag{3.5}$$

回归平方和U实质上是回归方程中的全部自变量的"方差贡献"，因此R^2就是这种贡献在总和中所占的比例。R称为"复(或全)相关系数"，它是全部自变量与y的相关，为强调这一点，有些书籍把R记作$r_{y\cdot12\cdots m}$。易知$0\le R\le 1$。显然，复相关系数越接近1，则回归效果越好。

尽管我们常用R作为总的回归效果的一个重要指标，但是我们要清醒地看到，R是与回归方程中变量个数m以及观测次数N有关的。当m较大而N不那么大时，常常有较大的R（即较小的Q）。特别当$m+1=N$时，即使这m个变量与y风马牛不相干，亦必然有$R=1$（即$Q=0$）。这就是说，在实际计算中，要注意m与N的比例适当。一般人认为N至少是m的5（或10）倍以上。当然，这样的要求，并不是标R所特有的，在整个回归分析中都有此需要。

下面给出另一个衡量回归效果的指标：

$$F=\frac{U/m}{Q/(N-m-1)} \tag{3.6}$$

这是两个方差比，它考虑了m，N的作用，实际上比R更为合理。重要的是在$\beta_i\equiv0\ (i=1,2,\cdots,m)$的假设下，这个统计量服从$F$分布，自由度分别为$m$与$N-m-1$。因此可用它来检验这$m$个变量的回归效果。例如在给定显著水平$\alpha$后，可在$F$分布表中查出相应的临界值$F_\alpha$，当计算的$F>F_\alpha$，则各$\beta_i$全为零的假设不成立，即回归效果显著。

上述内容，可归纳成一个"方差分析表"

方 差 分 析 表

来源	平方和	自由度	方差	F检验
回归	$U=\sum_t(y_t^*-\bar{y})^2=\sum_i b_i S_{iy}$	m	U/m	$F=\dfrac{U/m}{Q/(N-m-1)}$
残差	$Q=\sum_t(y_t-y_t^*)^2$	$N-m-1$	$Q/(N-m-1)$	
总和	$S_{yy}=\sum_t(y_t-\bar{y})^2$	$N-1$		

(2) β_i 与 y 的估计精度——区间估计

在§2中，已给出 β_i 的估计值 b_i，并给出 y_t 的估计值 y_t^*。但是这种估计只给出一个值，并没有给出估计值的精度的知识，因此称为点估计。这里将给出 β_i 与 y 的区间估计，由估值区间的大小，便可知道它们的精度。

首先我们指出 b_i 是服从正态分布的随机变量，并有

均值 $E(b_i)=\beta_i$ (3.7)

方差 $V(b_i)=C_{ii}\sigma^2$ (3.8)

协方差 $V(b_i,b_j)=C_{ij}\sigma^2 \quad (i,j=1,2,\cdots,m)$ (3.9)

这里 $C=(C_{ij})$ 是正规方程的系数矩阵 $S=(S_{ij})$ 的逆矩阵[註]。σ^2 是 y_t（或 ε_t）的公共方差。一般 σ^2 未知，我们可以给出它的无偏估计 S_y^2，可証：

$$S_y^2=\frac{Q}{N-m-1} \tag{3.10}$$

有关 b_i 和 S_y 的一些结论的証明，这里不再给出，有兴趣的读者，可从［文献2］的前77页中看到。

下面我们将使用这些结論给出 β_i 与 y 的区间估计。

〔註〕：注意此处有关 b_i 的結論，虽然没有把 b_0 包括进去（人们往往并不太注意 b_0 的性质）。但只需作小的修改，即把 (C_{ij}) 理解为〔说明2.1〕中的 $m+1$ 阶矩阵 (a_{ij}) 的逆，则有关結論对 b_0 也是正确的。

因为 b_i 服从正态分布 $N(\beta_i, C_{ii}\sigma^2)$，因此

$$z_i=\frac{b_i-\beta_i}{\sqrt{C_{ii}}\,\sigma}$$

是标准正态量 $N(0,1)$。如果 σ 已知，则对指定的置信水平 α，可给出 β_i 的区间估計

$$\left|\frac{b_i-\beta_i}{\sqrt{C_{ii}}\;\sigma}\right|<N_{\alpha/2} \qquad (3\cdot 12)$$

或

$$b_i-N_{\alpha/2}\sqrt{C_{ii}}\,\sigma<\beta_i<b_i+N_{\alpha/2}\sqrt{C_{ii}}\,\sigma$$

式中 $N_{\alpha/2}$ 为正态分布表上相应于 $\alpha/2$ 的临界值〔註〕。对于典型值 $\alpha=0.05$，$N_{0.025}=1.96$。在实际问题中，一般 σ 是未知的；这时可用它的估值 S_y 替代，但按統計理論，

$$t_i=\frac{b_i-\beta_i}{\sqrt{C_{ii}}\,S_y} \qquad (3\cdot 13)$$

已不再是正态分布，而是 t 分布，自由度为 $N-m-1$，因此 β_i 的区间估值为

$$\left|\frac{b_i-\beta_i}{\sqrt{C_{ii}}\,S_y}\right|<t_{\alpha/2} \qquad (3\cdot 14)$$

式中 $t_{\alpha/2}$ 为 t 分布表上相应于 $\alpha/2$ 的临界值〔註〕。

但是当 t 分布的自由度数大时，t 分布很接近正态分布。因此在大子样问题中，可直接把 $(3\cdot 12)$ 式中的 σ 代之以 S_y 作 β_i 的估計。

由〔說明1·1〕知 y 服从正态分布

$$N(\beta_0+\beta_1X_1+\cdots\cdots+\beta_mX_m,\ \sigma^2)$$

类似上述 β_i 的区间估計的討論，可知：当自变量的任一組观

〔註〕：由于 $|b_i-\beta_i|$ 包括 $b_i-\beta_i>0$ 和 $b_i-\beta_i<0$ 两种情况，根据分布的对称性，因此临界值相应于 $\alpha/2$。

测值（$x_1, x_2, \cdots, x_m$）相对应的y的区间估計为

$$\left|\frac{y-(\beta_0+\beta_1 x_1+\cdots+\beta_m x_m)}{\sigma}\right|<N_{\alpha/2} \qquad (3\cdot15)$$

（显著水平为α）

使用上式时要求σ与各β_i都已知，这在实际问题中是几乎不可能的。为了給出区间估計，我們也可以按統計理論构造新的統計量，但是計算相当复杂。在实际計算中，只要样本足够大，人們往々用b_i代替β_i，用S_y代替σ，得如下近似式

$$\left|\frac{y-(b_0+b_1 x_1+\cdots+b_m x_m}{S_y}\right|<N_{\alpha/2} \qquad (3\cdot16)$$

或

$$y^*-N_{\alpha/2}S_y<y<y^*+N_{\alpha/2}S_y$$

上述分析表明，β_i与y的估計精度直接依赖于S_y，因此S_y也是常用于衡量总的回归效果的一个重要指标。由于

$$S_y^2=\frac{Q}{N-m-1}$$

要使S_y减小（即提高估計精度），则必须降低Q或m。众所周知，残差平方和Q是随着回归方程中变量个数m的增加而减少的，因此要求同时降低Q和m是矛盾的。解决矛盾的方法显然是对給出的自变量逐一鑑别，去粗取精，选用重要的变量，这正是下面的討論内容。

(3) 各变量的重要性——β_i的显著性检验。

类似本节（1）的討論，考察回归方程中的任一个变量（例如x_i）的方差貢献。为明确起见，令$Q^{(m)}$表示m个变量组成的回归方程的残差平方和，$Q^{(m-1)}$表示m个变量中去掉任一自变量（例如x_i）后所得的残差平方和。令

$$V_i=Q^{(m-1)}-Q^{(m)} \qquad (3\cdot17)$$

显然V_i就是x_i在这m个变量的回归方程中的方差貢献。仿