第十二章　回归分析

前面我们讲过曲线拟合问题。曲线拟合问题的特点是，根据得到的若干有关变量的一组数据，寻找因变量与（一个或几个）自变量之间的一个函数，使这个函数对那组数据拟合得最好。通常，函数的形式可以由经验、先验知识或对数据的直观观察决定，要作的工作是由数据用最小二乘法计算函数中的待定系数。从计算的角度看，问题似乎已经完全解决了，还有进一步研究的必要吗?

从数理统计的观点看，这里涉及的都是随机变量，我们根据一个样本计算出的那些系数，只是它们的一个（点）估计，应该对它们作区间估计或假设检验，如果置信区间太大，甚至包含了零点，那么系数的估计值是没有多大意义的。另外也可以用方差分析方法对模型的误差进行分析，对拟合的优劣给出评价。简单地说，回归分析就是对拟合问题作的统计分析。

具体地说，回归分析在一组数据的基础上研究这样几个问题：

（i）建立因变量 y 与自变量 $x_1, x_2, \cdots, x_m$ 之间的回归模型（经验公式）；

（ii）对回归模型的可信度进行检验；

（iii）判断每个自变量 $x_i (i = 1,2,\cdots,m)$ 对 y 的影响是否显著；

（iv）诊断回归模型是否适合这组数据；

（v）利用回归模型对 y 进行预报或控制。

§1　数据表的基础知识

1.1　样本空间

在本章中，我们所涉及的均是**样本点×变量**类型的数据表。如果有 m 个变量 $x_1, x_2, \cdots, x_m$，对它们分别进行了 n 次采样（或观测），得到 n 个样本点

$$(x_{i1}, x_{i2}, \cdots, x_{im}),\quad i = 1,2,\cdots,n$$

则所构成的数据表 X 可以写成一个 $n \times m$ 维的矩阵。

$$X = (x_{ij})_{n\times m} = \begin{bmatrix} e_1^T \\ \vdots \\ e_n^T \end{bmatrix}$$

式中 $e_i = (x_{i1}, x_{i2}, \cdots, x_{im})^T \in R^m$，$i = 1,2,\cdots,n$，$e_i$ 被称为第 i 个样本点。

样本的均值为

$$\bar{x} = (\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_m),\quad \bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij},\quad j = 1,2,\cdots,m$$

样本协方差矩阵及样本相关系数矩阵分别为

$$S = (s_{ij})_{m\times m} = \frac{1}{n-1}\sum_{k=1}^{n}(e_k - \bar{x})(e_k - \bar{x})^T$$

$$R = (r_{ij})_{m\times m} = \left(\frac{s_{ij}}{\sqrt{s_{ii}s_{jj}}}\right)$$

其中

$$s_{ij}=\frac{1}{n-1}\sum_{k=1}^{n}(x_{ki}-\bar{x}_i)(x_{kj}-\bar{x}_j)$$

1.2　数据的标准化处理

（1）数据的中心化处理

数据的中心化处理是指平移变换，即

$$x_{ij}^{*}=x_{ij}-\bar{x}_j,\quad i=1,2,\cdots,n;\quad j=1,2,\cdots,m$$

该变换可以使样本的均值变为 0，而这样的变换既不改变样本点间的相互位置，也不改变变量间的相关性。但变换后，却常常有许多技术上的便利。

（2）数据的无量纲化处理

在实际问题中，不同变量的测量单位往往是不一样的。为了消除变量的量纲效应，使每个变量都具有同等的表现力，数据分析中常用的消量纲的方法，是对不同的变量进行所谓的压缩处理，即使每个变量的方差均变成 1，即

$$x_{ij}^{*}=x_{ij}/s_j$$

其中 $s_j=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2}$ 。

还可以有其它消量纲的方法，如

$$\begin{cases}x_{ij}^{*}=x_{ij}/\max\limits_{i}\{x_{ij}\},\quad x_{ij}^{*}=x_{ij}/\min\limits_{i}\{x_{ij}\}\\ x_{ij}^{*}=x_{ij}/\bar{x}_j,\quad x_{ij}^{*}=x_{ij}/(\max\limits_{i}\{x_{ij}\}-\min\limits_{i}\{x_{ij}\})\end{cases}$$

（3）标准化处理

所谓对数据的标准化处理，是指对数据同时进行中心化－压缩处理，即

$$x_{ij}^{*}=\frac{x_{ij}-\bar{x}_j}{s_j},\quad i=1,2,\cdots,n,\quad j=1,2,\cdots,m。$$

§2　一元线性回归

2.1　模型

一元线性回归的模型为

$$y=\beta_0+\beta_1x+\varepsilon,\qquad(1)$$

式中，β_0,β_1 为回归系数，ε 是随机误差项，总是假设 $\varepsilon\sim N(0,\sigma^2)$，则随机变量 $y\sim N(\beta_0+\beta_1x,\sigma^2)$。

若对 y 和 x 分别进行了 n 次独立观测，得到以下 n 对观测值

$$(y_i,x_i),\quad i=1,2,\cdots,n\qquad(2)$$

这 n 对观测值之间的关系符合模型

$$y_i=\beta_0+\beta_1x+\varepsilon_i,\quad i=1,2,\cdots,n\qquad(3)$$

这里，x_i 是自变量在第 i 次观测时的取值，它是一个非随机变量，并且没有测量误差。对应于 x_i，y_i 是一个随机变量，它的随机性是由 ε_i 造成的。$\varepsilon_i\sim N(0,\sigma^2)$，对于不同的观测，当 $i\neq j$ 时，ε_i 与 ε_j 是相互独立的。

2.2　最小二乘估计方法

2.2.1　最小二乘法

用最小二乘法估计 β_0,β_1 的值，即取 β_0,β_1 的一组估计值 $\hat{\beta}_0,\hat{\beta}_1$，使 y_i 与 $\hat{y}_i=\hat{\beta}_0+\hat{\beta}_1 x$ 的误差平方和达到最小。若记

$$Q(\beta_0,\beta_1)=\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2$$

则

$$Q(\hat{\beta}_0,\hat{\beta}_1)=\min_{\beta_0,\beta_1}Q(\beta_0,\beta_1)=\sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1 x_i)^2$$

显然 $Q(\beta_0,\beta_1)\geq 0$，且关于 β_0,β_1 可微，则由多元函数存在极值的必要条件得

$$\frac{\partial Q}{\partial \beta_0}=-2\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)=0$$

$$\frac{\partial Q}{\partial \beta_1}=-2\sum_{i=1}^{n}x_i(y_i-\beta_0-\beta_1 x_i)=0$$

整理后，得到下面的方程组

$$\begin{cases} n\beta_0+\beta_1\sum_{i=1}^{n}x_i=\sum_{i=1}^{n}y_i \\ \beta_0\sum_{i=1}^{n}x_i+\beta_1\sum_{i=1}^{n}x_i^2=\sum_{i=1}^{n}x_i y_i \end{cases} \tag{4}$$

此方程组称为正规方程组，求解可以得到

$$\begin{cases} \hat{\beta}_1=\dfrac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \\ \hat{\beta}_0=\bar{y}-\hat{\beta}_1\bar{x} \end{cases} \tag{5}$$

称 $\hat{\beta}_0,\hat{\beta}_1$ 为 β_0,β_1 的最小二乘估计，其中，$\bar{x},\bar{y}$ 分别是 x_i 与 y_i 的样本均值，即

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i\ ,\quad \bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i$$

关于 β_1 的计算公式还有一个更直观的表示方法，即

$$\hat{\beta}_1=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

$$=\frac{\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}}\cdot\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}=\frac{s_y}{s_x}r_{xy}$$

式中 $s_x^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$，$s_y^2=\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2$，$r_{xy}$ 是 x 与 y 的样本相关系数。

显然，当 x_i,y_i 都是标准化数据时，则有 $\bar{x}=0$，$\bar{y}=0$，$s_x=1$，$s_y=1$。所以，有

$$\hat{\beta}_0=0,\quad \hat{\beta}_1=r_{xy}$$

回归方程为

$$\hat{y}=r_{xy}x$$

由上可知，对标准化数据，$\hat{\beta}_1$ 可以表示 y 与 x 的相关程度。

2.2.2 $\hat{\beta}_0,\hat{\beta}_1$ 的性质

作为一个随机变量，$\hat{\beta}_1$ 有以下性质。

1．$\hat{\beta}_1$ 是 y_i 的线性组合，它可以写成

$$\hat{\beta}_1=\sum_{i=1}^{n}k_iy_i \tag{6}$$

式中，k_i 是固定的常量，$k_i=\frac{x_i-\bar{x}}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$。

证明　事实上

$$\hat{\beta}_1=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})y_i-\bar{y}\sum_{i=1}^{n}(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

由于

$$\bar{y}\sum_{i=1}^{n}(x_i-\bar{x})=\bar{y}(n\bar{x}-n\bar{x})=0$$

所以

$$\hat{\beta}_1=\sum_{i=1}^{n}\frac{x_i-\bar{x}}{\sum_{i=1}^{n}(x_i-\bar{x})^2}y_i$$

2．因为 $\hat{\beta}_1$ 是随机变量 $y_i(i=1,2,\cdots,n)$ 的线性组合，而 y_i 是相互独立、且服从正态分布的，所以，$\hat{\beta}_1$ 的抽样分布也服从正态分布。

3．点估计量 $\hat{\beta}_1$ 是总体参数 β_1 的无偏估计，有

$$E(\hat{\beta}_1)=E\left(\sum_{i=1}^{n}k_i y_i\right)=\sum_{i=1}^{n}k_i E(y_i)$$

$$=\sum_{i=1}^{n}k_i E(\beta_0+\beta_1 x_i)=\beta_0\sum_{i=1}^{n}k_i+\beta_1\sum_{i=1}^{n}k_i x_i$$

由于

$$\sum_{i=1}^{n}k_i=\sum_{i=1}^{n}\frac{x_i-\bar{x}}{\sum_{i=1}^{n}(x_i-\bar{x})^2}=0$$

$$\sum_{i=1}^{n}k_i x_i=\sum_{i=1}^{n}\frac{x_i-\bar{x}}{\sum_{i=1}^{n}(x_i-\bar{x})^2}x_i=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}=1$$

所以

$$E(\hat{\beta}_1)=\beta_1$$

4．估计量 $\hat{\beta}_1$ 的方差为

$$\mathrm{Var}(\hat{\beta}_1)=\frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \qquad (7)$$

这是因为

$$\mathrm{Var}(\hat{\beta}_1)=\mathrm{Var}\left(\sum_{i=1}^{n}k_i y_i\right)=\sum_{i=1}^{n}k_i^2\mathrm{Var}(y_i)=\sum_{i=1}^{n}k_i^2\sigma^2=\sigma^2\sum_{i=1}^{n}k_i^2$$

由于

$$\sum_{i=1}^{n}k_i^2=\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)^2=\frac{1}{\left[\sum_{i=1}^{n}(x_i-\bar{x})^2\right]^2}\sum_{i=1}^{n}(x_i-\bar{x})^2=\frac{1}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

因此，式（7）得证。

5．对于总体模型中的参数 β_1，在它的所有线性无偏估计量中，最小二乘估计量 $\hat{\beta}_1$ 具有最小的方差。

记任意一个线性估计量

$$\tilde{\beta}_1=\sum_{i=1}^{n}c_i y_i$$

式中 c_i 是任意常数，c_i 不全为零，$i=1,2,\cdots,n$。要求 $\tilde{\beta}_1$ 是 β_1 的无偏估计量，即

$$E(\tilde{\beta}_1)=\sum_{i=1}^{n}c_i E(y_i)=\beta_1$$

另一方面，由于 $E(y_i)=\beta_0+\beta_1 x_i$，所以又可以写成

$$E(\tilde{\beta}_1)=\sum_{i=1}^{n}c_i(\beta_0+\beta_1x_i)=\beta_0\sum_{i=1}^{n}c_i+\beta_1\sum_{i=1}^{n}c_ix_i$$

为保证无偏性，c_i 要满足下列限制

$$\sum_{i=1}^{n}c_i=0,\ \sum_{i=1}^{n}c_ix_i=0$$

定义 $c_i=k_i+d_i$，其中 k_i 是式（6）中的组合系数，d_i 是任意常数，则

$$\mathrm{Var}(\tilde{\beta}_1)=\sigma^2\sum_{i=1}^{n}c_i^2=\sigma^2\left(\sum_{i=1}^{n}k_i^2+\sum_{i=1}^{n}d_i^2+2\sum_{i=1}^{n}k_id_i\right)$$

由于

$$\sum_{i=1}^{n}k_id_i=\sum_{i=1}^{n}k_i(c_i-k_i)=\sum_{i=1}^{n}c_i\frac{x_i-\bar{x}}{\sum_{i=1}^{n}(x_i-\bar{x})^2}-\sum_{i=1}^{n}k_i^2$$

$$=\frac{\sum_{i=1}^{n}c_ix_i-\bar{x}\sum_{i=1}^{n}c_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2}-\sum_{i=1}^{n}k_i^2=\frac{1}{\sum_{i=1}^{n}(x_i-\bar{x})^2}-\frac{1}{\sum_{i=1}^{n}(x_i-\bar{x})^2}=0$$

而

$$\sigma^2\sum_{i=1}^{n}k_i^2=\frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}=\mathrm{Var}(\hat{\beta}_1)$$

所以

$$\mathrm{Var}(\tilde{\beta}_1)=\mathrm{Var}(\hat{\beta}_1)+\sigma^2\sum_{i=1}^{n}d_i^2$$

$\sum_{i=1}^{n}d_i^2$ 的最小值为零，所以，当 $\sum_{i=1}^{n}d_i^2=0$ 时，$\tilde{\beta}_1$ 的方差最小。但是，只有当 $d_i\equiv 0$ 时，即 $c_i\equiv k_i$ 时，才有 $\sum_{i=1}^{n}d_i^2=0$。所以，最小二乘估计量 $\hat{\beta}_1$ 在所有无偏估计量中具有最小的方差。

同理，可以得出相应于点估计量 $\hat{\beta}_0$ 的统计性质。对于一元线性正态误差回归模型来说，最小二乘估计量 $\hat{\beta}_0$ 是 y_i 的线性组合，所以，它的抽样分布也是正态的。它是总体参数 β_0 的无偏估计量，即

$$E(\hat{\beta}_0)=\beta_0$$

同样可以证明

$$Var(\hat{\beta}_0)=\sigma^2[\frac{1}{n}+\frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}]\qquad(8)$$

且 $\hat{\beta}_0$ 是 β_0 的线性无偏的最小方差估计量。

2.2.3 其它性质

用最小二乘法拟合的回归方程还有一些值得注意的性质：

1．残差和为零。

残差

$$e_i = y_i - \hat{y}_i,\quad i = 1,2,\cdots,n$$

由第一个正规方程，得

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_1) = 0 \tag{9}$$

2．拟合值 $\hat{y}_i$ 的平均值等于观测值 y_i 的平均值，即

$$\frac{1}{n}\sum_{i=1}^{n} \hat{y}_i = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y} \tag{10}$$

按照第一正规方程，有

$$\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

所以

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} (\hat{\beta}_0 + \hat{\beta}_1 x_i) = \sum_{i=1}^{n} \hat{y}_i$$

3．当第 i 次试验的残差以相应的自变量取值为权重时，其加权残差和为零，即

$$\sum_{i=1}^{n} x_i e_i = 0 \tag{11}$$

这个结论由第二个正规方程 $\sum_{i=1}^{n} x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$ 即可得出。

4．当第 i 次试验的残差以相应的因变量的拟合值为权重时，其加权残差和为零，即

$$\sum_{i=1}^{n} \hat{y}_i e_i = 0 \tag{12}$$

这是因为

$$\sum_{i=1}^{n} (\hat{\beta}_0 + \hat{\beta}_1 x_i) e_i = \hat{\beta}_0 \sum_{i=1}^{n} e_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i e_i = 0$$

5．最小二乘回归线总是通过观测数据的重心 $(\bar{x}, \bar{y})$ 的。

事实上，当自变量取值为 $\bar{x}$ 时，由式（5）

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

所以

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x} = (\bar{y} - \hat{\beta}_1 \bar{x}) + \hat{\beta}_1 \bar{x} = \bar{y}$$

2.3 拟合效果分析

当根据一组观测数据得到最小二乘拟合方程后，必须考察一下，是否真的能由所得

的模型（$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$）来较好地拟合观测值 y_i？用 $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ 能否较好地反映（或者说解释）y_i 值的取值变化？回归方程的质量如何？误差多大？对这些，都必须予以正确的评估和分析。

2.3.1　残差的样本方差

记残差

$$e_i = y_i - \hat{y}_i,\ \ i = 1,2,\cdots,n$$

残差的样本均值为

$$\bar{e} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) = 0$$

残差的样本方差为

$$MSE = \frac{1}{n-2}\sum_{i=1}^{n}(e_i - \bar{e})^2 = \frac{1}{n-2}\sum_{i=1}^{n}e_i^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

由于有 $\sum_{i=1}^{n}e_i = 0$ 和 $\sum_{i=1}^{n}x_i e_i = 0$ 的约束，所以，残差平方和有 $(n-2)$ 个自由度。可以证明，在对 $\sum_{i=1}^{n}e_i^2$ 除以其自由度 $(n-2)$ 后得到的 MSE，是总体回归模型中 $\sigma^2 = Var(\varepsilon_i)$ 的无偏估计量。记

$$S_e = \sqrt{MSE} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}e_i^2} \qquad (13)$$

一个好的拟合方程，其残差总和应越小越好。残差越小，拟合值与观测值越接近，各观测点在拟合直线周围聚集的紧密程度越高，也就是说，拟合方程 $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ 解释 y 的能力越强。

另外，当 S_e 越小时，还说明残差值 e_i 的变异程度越小。由于残差的样本均值为零，所以，其离散范围越小，拟合的模型就越为精确。

2.3.2　判定系数（拟合优度）

对应于不同的 x_i 值，观测值 y_i 的取值是不同的。建立一元线性回归模型的目的，就是试图以 x 的线性函数（$\hat{\beta}_0 + \hat{\beta}_1 x$）来解释 y 的变异。那么，回归模型 $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ 究竟能以多大的精度来解释 y 的变异呢？又有多大部分是无法用这个回归方程来解释的呢？

$y_1, y_2, \cdots, y_n$ 的变异程度可采用样本方差来测度，即

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

根据式（10），拟合值 $\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_n$ 的均值也是 $\bar{y}$，其变异程度可以用下式测度

$$\hat{s}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

下面看一下 s^2 与 $\hat{s}^2$ 之间的关系，有

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

由于

$$\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y})=\sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1x_i)(\hat{\beta}_0+\hat{\beta}_1x_i-\bar{y})$$

$$=\hat{\beta}_0\sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1x_i)+\hat{\beta}_1\sum_{i=1}^{n}x_i(y_i-\hat{\beta}_0-\hat{\beta}_1x_i)-\bar{y}\sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1x_i)=0$$

因此，得到正交分解式为

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \qquad (14)$$

记

$SST=\sum_{i=1}^{n}(y_i-\bar{y})^2$，这是原始数据 y_i 的总变异平方和，其自由度为 $df_T=n-1$；

$SSR=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$，这是用拟合直线 $\hat{y}_i=\hat{\beta}_0+\hat{\beta}_1x_i$ 可解释的变异平方和，其自由度为 $df_R=1$；

$SSE=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$，这是残差平方和，其的自由度为 $df_E=n-2$。

所以，有

$$SST=SSR+SSE,\quad df_T=df_R+df_E$$

从上式可以看出，y 的变异是由两方面的原因引起的；一是由于 x 的取值不同，而给 y 带来的系统性变异；另一个是由除 x 以外的其它因素的影响。

注意到对于一个确定的样本（一组实现的观测值），SST 是一个定值。所以，可解释变异 SSR 越大，则必然有残差 SSE 越小。这个分解式可同时从两个方面说明拟合方程的优良程度：

（1）SSR 越大，用回归方程来解释 y_i 变异的部分越大，回归方程对原数据解释得越好；

（2）SSE 越小，观测值 y_i 绕回归直线越紧密，回归方程对原数据的拟合效果越好。

因此，可以定义一个测量标准来说明回归方程对原始数据的拟合程度，这就是所谓的判定系数，有些文献上也称之为拟合优度。

判定系数是指可解释的变异占总变异的百分比，用 R^2 表示，有

$$R^2=\frac{SSR}{SST}=(1-\frac{SSE}{SST}) \qquad (15)$$

从判定系数的定义看，R^2 有以下简单性质：

（1）$0\le R^2\le 1$；

（2）当 $R^2=1$ 时，有 $SSR=SST$，也就是说，此时原数据的总变异完全可以由拟合值的变异来解释，并且残差为零（$SSE=0$），即拟合点与原数据完全吻合；

（3）当 $R^2=0$ 时，回归方程完全不能解释原数据的总变异，y 的变异完全由与 x

无关的因素引起，这时 $SSE = SST$ 。

测定系数时一个很有趣的指标：一方面它可以从数据变异的角度指出可解释的变异占总变异的百分比，从而说明回归直线拟合的优良程度；另一方面，它还可以从相关性的角度，说明原因变量 y 与拟合变量 $\hat{y}$ 的相关程度，从这个角度看，拟合变量 $\hat{y}$ 与原变量 y 的相关度越大，拟合直线的优良度就越高。

看下面的式子

$$R^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\left[\sum_{i=1}^{n}(\hat{y}_i + e_i - \bar{y})(\hat{y}_i - \bar{y})\right]^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2} = r^2(y, \hat{y}) \qquad (16)$$

在推导中，注意有

$$\sum_{i=1}^{n} e_i(\hat{y}_i - \bar{y}) = \sum_{i=1}^{n} e_i \hat{y}_i - \bar{y}\sum_{i=1}^{n} e_i = 0$$

所以， R^2 又等于 y 与拟合变量 $\hat{y}$ 的相关系数平方。

还可以证明， $\sqrt{R^2}$ 等于 y 与自变量 x 的相关系数，而相关系数的正、负号与回归系数 $\hat{\beta}_1$ 的符号相同。

2.4 显著性检验

2.4.1 回归模型的线性关系检验

在拟合回归方程之前，我们曾假设数据总体是符合线性正态误差模型的，也就是说，y 与 x 之间的关系是线性关系，即

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2), \quad i = 1,2,\cdots,n$$

然而，这种假设是否真实，还需进行检验。

对于一个实际观测的样本，虽然可以用判定系数 R^2 说明 y 与 $\hat{y}$ 的相关程度，但是，样本测度指标具有一定的随机因素，还不足以肯定 y 与 x 的线性关系。

假设 y 与 x 之间存在线性关系，则总体模型为

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1,2,\cdots,n$$

如果 $\beta_1 \neq 0$，则称这个模型为全模型。

用最小二乘法拟合全模型，并求出误差平方和为

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

现给出假设 $H_0 : \beta_1 = 0$。如果 H_0 假设成立，则

$$y_i = \beta_0 + \varepsilon_i$$

这个模型被称为选模型。用最小二乘法拟合这个模型，则有

$$\hat{\beta}_1 = 0$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_0 \bar{x} = \bar{y}$$

因此，对所有的 $i = 1,2,\cdots,n$，有

$$\hat{y}_i \equiv \bar{y}$$

该拟合模型的误差平方和为

$$\sum_{i=1}^{n}(y_i-\bar{y})^2 = SST$$

因此，有

$$SSE \le SST$$

这就是说，全模型的误差总是小于（或等于）选模型的误差的。其原因是在全模型中有较多的参数，可以更好地拟合数据。

假若在某个实际问题中，全模型的误差并不比选模型的误差小很多的话，这说明 H_0 假设成立，即 β_1 近似于零。因此，差额 $(SST-SSE)$ 很少时，表明 H_0 成立。若这个差额很大，说明增加了 x 的线性项后，拟合方程的误差大幅度减少，则应否定 H_0，认为总体参数 β_1 显著不为零。

假设检验使用的统计量为

$$F = \frac{SSR/1}{SSE/(n-2)} = \frac{MSR}{MSE}$$

式中

$$MSR = SSR/df_R = SSR/1$$
$$MSE = SSE/df_E = SSE/(n-2)$$

若假设 $H_0:\beta_1=0$ 成立，由于 $SST=SSR+SSE$，则 SSE/σ^2 与 SSR/σ^2 是独立的随机变量，且

$$SSE/\sigma^2 \sim \chi^2(n-2),\quad SSR/\sigma^2 \sim \chi^2(1)$$

这时

$$F = \frac{MSR}{MSE} \sim F(1,n-2)$$

综上所述，为了检验是否可以用 x 的线性方程式来解释 y，可以进行下面的统计检验。记 y_i 关于 x_i 的总体回归系数为 β_1，则 F 检验的原假设 H_0 与备则假设 H_1 分别是

$$H_0:\beta_1=0,\quad H_1:\beta_1\ne 0$$

检验的统计量为

$$F = \frac{MSR}{MSE} \sim F(1,n-2) \qquad (17)$$

对于检验水平 α，按自由度（$n_1=1$，$n_2=n-2$）查 F 分布表，得到拒绝域的临界值 $F_\alpha(1,n-2)$。决策规则为

若 $F \le F_\alpha(1,n-2)$，则接受 H_0 假设，这时认为 β_1 显著为零，无法用 x 的线性关系式来解释 y。

若 $F > F_\alpha(1,n-2)$，则否定 H_0，接受 H_1。这时认为 β_1 显著不为零，可以用 x 的线性关系来解释 y。习惯上说，线性回归方程的 F 检验通过了。

需要注意的是，即使 F 检验通过了，也不说明

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

就是一个恰当的回归模型，事实上，当 H_0 假设被拒绝后，只能说明 y 与 x 之间存在显著的线性关系，但很有可能在模型中还包括更多的回归变量，而不仅仅是一个回归变量 x。

一般地，回归方程的假设检验包括两个方面：一个是对模型的检验，即检验自变量与因变量之间的关系能否用一个线性模型来表示，这是由 F 检验来完成的；另一个检验是关于回归参数的检验，即当模型检验通过后，还要具体检验每一个自变量对因变量的影响程度是否显著。这就是下面要讨论的 t 检验。在一元线性分析中，由于自变量的个数只有一个，这两种检验是统一的，它们的效果完全是等价的。但是，在多元线性回归分析中，由于变量的个数只有一个，这两种检验是统一的，它们的效果完全是等价的。但是，在多元线性回归分析中，这两个建议的意义是不同的。从逻辑上说，一般常在 F 检验通过后，再进一步进行 t 建议。

2.4.2　回归系数的显著性建议

回归参数的建议是考察每一个自变量对因变量的影响是否显著。换句话说，就是要检验每一个总体参数是否显著不为零。

首先看对 $\beta_1 = 0$ 的检验。β_1 代表 x_i 变化一个单位对 y_i 的影响程度。对 β_1 的检验就是要看这种影响程度与零是否有显著差异。

由于

$$\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2})$$

$\mathrm{Var}(\hat{\beta}_1) = \dfrac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$ 的点估计为

$$S^2(\hat{\beta}_1) = \frac{MSE}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

容易证明统计量

$$\frac{\hat{\beta}_1 - \beta_1}{S(\hat{\beta}_1)} \sim t(n-2)$$

事实上，由于

$$\frac{\hat{\beta}_1 - \beta_1}{S(\hat{\beta}_1)} = \frac{(\hat{\beta}_1 - \beta_1)/\sqrt{\mathrm{Var}(\hat{\beta}_1)}}{S(\hat{\beta}_1)/\sqrt{\mathrm{Var}(\hat{\beta}_1)}}$$

其分子 $(\hat{\beta}_1 - \beta_1)/\sqrt{\mathrm{Var}(\hat{\beta}_1)}$ 服从标准正态分布，而分母项有

$$\frac{S^2(\hat{\beta}_1)}{\mathrm{Var}(\hat{\beta}_1)} = \frac{MSE/\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sigma^2/\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{MSE}{\sigma^2} = \frac{SSE}{\sigma^2(n-2)}$$

已知 $SSE/\sigma^2 \sim \chi^2(n-2)$，所以

$$\frac{\hat{\beta}_1-\beta_1}{S(\hat{\beta}_1)}\sim t(n-2)$$

$\hat{\beta}_1$ 的抽样分布清楚后，可以进行 β_1 是否显著为零的检验。

$H_0:\beta_1=0$， $H_1:\beta_1\neq 0$

检验统计量为

$$t_1=\frac{\hat{\beta}_1}{S(\hat{\beta}_1)}$$

检验统计量 t_1 在 $\beta_1=0$ 假设为真时，服从自由度为 $(n-2)$ 的 t 分布。

对于给定的检验水平 α，则通过 t 分布表可查到统计量 t_1 的临界值 $t_{\frac{\alpha}{2}}(n-2)$。决策规则是：

若 $|t_1|\le t_{\frac{\alpha}{2}}(n-2)$，则接受 H_0，认为 β_1 显著为零；

若 $|t_1|> t_{\frac{\alpha}{2}}(n-2)$，则拒绝 H_0，认为 β_1 显著不为零。

当拒绝了 H_0，认为 β_1 显著不为零时，又称 β_1 通过了 t 检验。

另一方面，由于

$$P\left\{\left|\frac{\hat{\beta}_1-\beta_1}{S(\hat{\beta}_1)}\right|<t_{\frac{\alpha}{2}}(n-2)\right\}=1-\alpha$$

还可以确定 β_1 的置信度为 $1-\alpha$ 的置信区间为

$$\hat{\beta}_1-t_{\frac{\alpha}{2}}(n-2)S(\hat{\beta}_1)\le\beta_1\le\hat{\beta}_1+t_{\frac{\alpha}{2}}(n-2)S(\hat{\beta}_1) \qquad (18)$$

同样地，也可以对总体参数 β_0 进行显著性检验，并且求出它的置信区间。它的最小二乘估计量 $\hat{\beta}_0$ 的抽样分布为正态分布，即

$$\hat{\beta}_0\sim N(\beta_0,\sigma^2[\frac{1}{n}+\frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}])$$

$\mathrm{Var}(\hat{\beta}_0)$ 的估计量为

$$S^2(\hat{\beta}_0)=MSE[\frac{1}{n}+\frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}]$$

可以推出

$$\frac{\hat{\beta}_0-\beta_0}{S(\hat{\beta}_0)}\sim t(n-2)$$

为检验 β_0 是否显著为零，提出假设

$H_0:\beta_0=0$， $H_1:\beta_0\neq 0$

检验统计量为

$$t_0 = \frac{\hat{\beta}_0}{S(\hat{\beta}_0)}$$

在 $\beta_0 = 0$ 时，检验统计量 t_0 服从自由度为 $(n-2)$ 的 t 分布。

对于给定的检验水平 α，则通过 t 分布表可查到统计量 t_0 的临界值 $t_{\frac{\alpha}{2}}(n-2)$。决策准则为：

若 $|t_0| \le t_{\frac{\alpha}{2}}(n-2)$，则接受 H_0，认为 β_0 显著为零；

若 $|t_0| > t_{\frac{\alpha}{2}}(n-2)$，则拒绝 H_0，认为 β_0 显著不为零。

此外，根据

$$P\left\{\left|\frac{\hat{\beta}_0 - \beta_0}{S(\hat{\beta}_0)}\right| < t_{\frac{\alpha}{2}}(n-2)\right\} = 1-\alpha$$

还可以确定 β_0 的置信度为 $1-\alpha$ 的置信区间为

$$\hat{\beta}_0 - t_{\frac{\alpha}{2}}(n-2)S(\hat{\beta}_0) \le \beta_0 \le \hat{\beta}_0 + t_{\frac{\alpha}{2}}(n-2)S(\hat{\beta}_0) \tag{19}$$

§3 多元线性回归

3.1 模型

多元线性回归分析的模型为

$$\begin{cases} y = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m + \varepsilon \\ \varepsilon \sim N(0, \sigma^2) \end{cases} \tag{20}$$

式中 $\beta_0, \beta_1, \cdots, \beta_m, \sigma^2$ 都是与 $x_1, x_2, \cdots, x_m$ 无关的未知参数，其中 $\beta_0, \beta_1, \cdots, \beta_m$ 称为回归系数。

现得到 n 个独立观测数据 $(y_i, x_{i1}, \cdots, x_{im})$，$i = 1, \cdots, n, n > m$，由（20）得

$$\begin{cases} y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_m x_{im} + \varepsilon_i \\ \varepsilon_i \sim N(0, \sigma^2), \quad i = 1, \cdots, n \end{cases} \tag{21}$$

记

$$X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1m} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1} & \cdots & x_{nm} \end{bmatrix}, \quad Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \tag{22}$$

$$\varepsilon = [\varepsilon_1 \quad \cdots \quad \varepsilon_n]^T, \quad \beta = [\beta_0 \quad \beta_1 \quad \cdots \quad \beta_m]^T$$

（20）表为

$$\begin{cases} Y = X\beta + \varepsilon \\ \varepsilon \sim N(0, \sigma^2 E_n) \end{cases} \tag{23}$$

其中 E_n 为 n 阶单位矩阵。

3.2　参数估计

模型（20）中的参数 $\beta_0,\beta_1,\cdots,\beta_m$ 仍用最小二乘法估计，即应选取估计值 $\hat{\beta}_j$，使当 $\beta_j=\hat{\beta}_j$ 时，$j=0,1,2,\cdots,m$ 时，误差平方和

$$Q=\sum_{i=1}^{n}\varepsilon_i^2=\sum_{i=1}^{n}(y_i-\beta_0-\beta_1x_{i1}-\cdots-\beta_mx_{im})^2 \tag{24}$$

达到最小。为此，令

$$\frac{\partial Q}{\partial\beta_j}=0,\quad j=0,1,2,\cdots,n$$

得

$$\begin{cases}\dfrac{\partial Q}{\partial\beta_0}=-2\sum\limits_{i=1}^{n}(y_i-\beta_0-\beta_1x_{i1}-\cdots-\beta_mx_{im})=0\\ \dfrac{\partial Q}{\partial\beta_j}=-2\sum\limits_{i=1}^{n}(y_i-\beta_0-\beta_1x_{i1}-\cdots-\beta_mx_{im})x_{ij}=0,\quad j=1,2,\cdots,m\end{cases} \tag{25}$$

经整理化为以下正规方程组

$$\begin{cases}\beta_0n+\beta_1\sum\limits_{i=1}^{n}x_{i1}+\beta_2\sum\limits_{i=1}^{n}x_{i2}+\cdots+\beta_m\sum\limits_{i=1}^{n}x_{im}=\sum\limits_{i=1}^{n}y_i\\ \beta_0\sum\limits_{i=1}^{n}x_{i1}+\beta_1\sum\limits_{i=1}^{n}x_{i1}^2+\beta_2\sum\limits_{i=1}^{n}x_{i1}x_{i2}+\cdots+\beta_m\sum\limits_{i=1}^{n}x_{i1}x_{im}=\sum\limits_{i=1}^{n}x_{i1}y_i\\ \beta_0\sum\limits_{i=1}^{n}x_{im}+\beta_1\sum\limits_{i=1}^{n}x_{im}x_{i1}+\beta_2\sum\limits_{i=1}^{n}x_{im}x_{i2}+\cdots+\beta_m\sum\limits_{i=1}^{n}x_{im}^2=\sum\limits_{i=1}^{n}x_{im}y_i\end{cases} \tag{26}$$

正规方程组的矩阵形式为

$$X^TX\beta=X^TY \tag{27}$$

当矩阵 X 列满秩时，X^TX 为可逆方阵，（27）式的解为

$$\hat{\beta}=(X^TX)^{-1}X^TY \tag{28}$$

将 $\hat{\beta}$ 代回原模型得到 y 的估计值

$$\hat{y}=\hat{\beta}_0+\hat{\beta}_1x_1+\cdots+\hat{\beta}_mx_m \tag{29}$$

而这组数据的拟合值为 $\hat{Y}=X\hat{\beta}$，拟合误差 $e=Y-\hat{Y}$ 称为残差，可作为随机误差 ε 的估计，而

$$Q=\sum_{i=1}^{n}e_i^2=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \tag{30}$$

为残差平方和（或剩余平方和），即 $Q(\hat{\beta})$。

3.3　统计分析

不加证明地给出以下结果：

（i）$\hat{\beta}$ 是 β 的线性无偏最小方差估计。指的是 $\hat{\beta}$ 是 Y 的线性函数；$\hat{\beta}$ 的期望等于 β；在 β 的线性无偏估计中，$\hat{\beta}$ 的方差最小。

（ii）$\hat{\beta}$ 服从正态分布

$$\hat{\beta} \sim N(\beta, \sigma^2 (X^T X)^{-1}) \tag{31}$$

记 $(X^T X)^{-1} = (c_{ij})_{n\times n}$。

（iii）对残差平方和 Q，$EQ = (n-m-1)\sigma^2$，且

$$\frac{Q}{\sigma^2} \sim \chi^2(n-m-1) \tag{32}$$

由此得到 σ^2 的无偏估计

$$s^2 = \frac{Q}{n-m-1} = \hat{\sigma}^2 \tag{33}$$

s^2 是剩余方差（残差的方差），s 称为剩余标准差。

（iv）对总平方和 $SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$ 进行分解，有

$$SST = Q + U，\ U = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \tag{34}$$

其中 Q 是由（24）定义的残差平方和，反映随机误差对 y 的影响，U 称为回归平方和，反映自变量对 y 的影响。上面的分解中利用了正规方程组。

3.4　回归模型的假设检验

因变量 y 与自变量 $x_1, \cdots, x_m$ 之间是否存在如模型（20）所示的线性关系是需要检验的，显然，如果所有的 $|\hat{\beta}_j|$ $(j = 1, \cdots, m)$ 都很小，y 与 $x_1, \cdots, x_m$ 的线性关系就不明显，所以可令原假设为

$$H_0 : \beta_j = 0 (j = 1, \cdots, m)$$

当 H_0 成立时由分解式（34）定义的 U, Q 满足

$$F = \frac{U/m}{Q/(n-m-1)} \sim F(m, n-m-1) \tag{35}$$

在显著性水平 α 下有上 α 分位数 $F_\alpha(m, n-m-1)$，若 $F < F_\alpha(m, n-m-1)$，接受 H_0；否则，拒绝。

注意　接受 H_0 只说明 y 与 $x_1, \cdots, x_m$ 的线性关系不明显，可能存在非线性关系，如平方关系。

还有一些衡量 y 与 $x_1, \cdots, x_m$ 相关程度的指标，如用回归平方和在总平方和中的比值定义复判定系数

$$R^2 = \frac{U}{S} \tag{36}$$

$R = \sqrt{R^2}$ 称为复相关系数，R 越大，y 与 $x_1, \cdots, x_m$ 相关关系越密切，通常，R 大于 0.8（或 0.9）才认为相关关系成立。

3.5　回归系数的假设检验和区间估计

当上面的 H_0 被拒绝时，β_j 不全为零，但是不排除其中若干个等于零。所以应进

一步作如下 m 个检验 $(j=0,1,\cdots,m)$：

$$H_0^{(j)}:\beta_j=0$$

由(31)式，$\hat{\beta}_j\sim N(\beta_j,\sigma^2c_{jj})$，$c_{jj}$ 是 $(X^TX)^{-1}$ 中的第 (j,j) 元素，用 s^2 代替 σ^2，由（31）~（33）式，当 $H_0^{(j)}$ 成立时

$$t_j=\frac{\hat{\beta}_j/\sqrt{c_{jj}}}{\sqrt{Q/(n-m-1)}}\sim t(n-m-1) \qquad (37)$$

对给定的 α，若 $|t_j|<t_{\frac{\alpha}{2}}(n-m-1)$，接受 $H_0^{(j)}$；否则，拒绝。

（37）式也可用于对 β_j 作区间估计（ $j=0,1,\cdots,m$ ），在置信水平 $1-\alpha$ 下，β_j 的置信区间为

$$[\hat{\beta}_j-t_{\frac{\alpha}{2}}(n-m-1)s\sqrt{c_{jj}},\hat{\beta}_j+t_{\frac{\alpha}{2}}(n-m-1)s\sqrt{c_{jj}}] \qquad (38)$$

其中 $s=\sqrt{\frac{Q}{n-m-1}}$。

3.6　利用回归模型进行预测

当回归模型和系数通过检验后，可由给定的 $x_0=(x_{01},\cdots,x_{0m})$ 预测 y_0，y_0 是随机的，显然其预测值（点估计）为

$$\hat{y}_0=\hat{\beta}_0+\hat{\beta}_1x_{01}+\cdots+\hat{\beta}_mx_{0m} \qquad (39)$$

给定 α 可以算出 y_0 的预测区间（区间估计），结果较复杂，但当 n 较大且 x_{0i} 接近平均值 $\bar{x}_i$ 时，y_0 的预测区间可简化为

$$[\hat{y}_0-z_{\frac{\alpha}{2}}s,\hat{y}_0+z_{\frac{\alpha}{2}}s] \qquad (40)$$

其中 $z_{\frac{\alpha}{2}}$ 是标准正态分布的上 $\frac{\alpha}{2}$ 分位数。

对 y_0 的区间估计方法可用于给出已知数据残差 $e_i=y_i-\hat{y}_i\ (i=1,\cdots,n)$ 的置信区间，e_i 服从均值为零的正态分布，所以若某个 e_i 的置信区间不包含零点，则认为这个数据是异常的，可予以剔除。

§4　Matlab 中的回归分析

4.1　多元线性回归

Matlab 统计工具箱用命令 regress 实现多元线性回归，用的方法是最小二乘法，用法是：

```
b=regress(Y,X)
```

其中 Y,X 为按（22）式排列的数据，b 为回归系数估计值 $\hat{\beta}_0,\hat{\beta}_1,\cdots,\hat{\beta}_m$。

```
[b,bint,r,rint,stats]=regress(Y,X,alpha)
```

这里 Y,X 同上，alpha 为显著性水平（缺省时设定为 0.05），b,bint 为回归系数估计值和它们的置信区间，r,rint 为残差（向量）及其置信区间，stats 是用于检验回归模型的统计量，有四个数值，第一个是 R^2（见（36）式），第二个是 F（见（35）式），第三个

是与 F 对应的概率 p， $p<\alpha$ 拒绝 H_0，回归模型成立，第四个是残差的方差 s^2（见（33）式）。

残差及其置信区间可以用 rcoplot(r,rint)画图。

例 1　合金的强度 y 与其中的碳含量 x 有比较密切的关系，今从生产中收集了一批数据如下表 1。

表 1

x	0.10	0.11	0.12	0.13	0.14	0.15	0.16	0.17	0.18
y	42.0	41.5	45.0	45.5	45.0	47.5	49.0	55.0	50.0

试先拟合一个函数 $y(x)$，再用回归分析对它进行检验。

解　先画出散点图：

```
x=0.1:0.01:0.18;
y=[42,41.5,45.0,45.5,45.0,47.5,49.0,55.0,50.0];
plot(x,y,'+')
```

可知 y 与 x 大致上为线性关系。

设回归模型为

$$y=\beta_0+\beta_1 x \tag{41}$$

用 regress 和 rcoplot 编程如下：

```
clc,clear
x1=[0.1:0.01:0.18]';
y=[42,41.5,45.0,45.5,45.0,47.5,49.0,55.0,50.0]';
x=[ones(9,1),x1];
[b,bint,r,rint,stats]=regress(y,x);
b,bint,stats,rcoplot(r,rint)
```

得到

```
b =27.4722   137.5000
bint =18.6851    36.2594
       75.7755   199.2245
stats =0.7985    27.7469    0.0012    4.0883
```

即 $\hat{\beta}_0=27.4722$，$\hat{\beta}_1=137.5000$，$\hat{\beta}_0$ 的置信区间是[18.6851,36.2594]，$\hat{\beta}_1$ 的置信区间是[75.7755,199.2245]；$R^2=0.7985$，$F=27.7469$，$p=0.0012$，$s^2=4.0883$。

可知模型（41）成立。

观察命令 rcoplot(r,rint)所画的残差分布，除第 8 个数据外其余残差的置信区间均包含零点，第 8 个点应视为异常点，将其剔除后重新计算，可得

```
b =30.7820    109.3985
bint =26.2805    35.2834
       76.9014   141.8955
stats =0.9188    67.8534    0.0002    0.8797
```

应该用修改后的这个结果。

表 2

x_1 元	120	140	190	130	155	175	125	145	180	150
x_2 元	100	110	90	150	210	150	250	270	300	250
y 个	102	100	120	77	46	93	26	69	65	85

例 2　某厂生产的一种电器的销售量 y 与竞争对手的价格 x_1 和本厂的价格 x_2 有关。表 2 是该商品在 10 个城市的销售记录。试根据这些数据建立 y 与 x_1 和 x_2 的关系式，对得到的模型和系数进行检验。若某市本厂产品售价 160（元），竞争对手售价 170（元），预测商品在该市的销售量。

解　分别画出 y 关于 x_1 和 y 关于 x_2 的散点图，可以看出 y 与 x_2 有较明显的线性关系，而 y 与 x_1 之间的关系则难以确定，我们将作几种尝试，用统计分析决定优劣。

设回归模型为

$$y=\beta_0+\beta_1x_1+\beta_2x_2 \tag{42}$$

编写如下程序：

```
x1=[120  140  190  130  155  175  125  145  180  150]';
x2=[100  110   90  150  210  150  250  270  300  250]';
y=[102  100   120  77   46   93   26   69   65   85]';
x=[ones(10,1),x1,x2];
[b,bint,r,rint,stats]=regress(y,x);
b,bint,stats
```

得到

```
b =66.5176      0.4139     -0.2698
bint =-32.5060   165.5411
       -0.2018     1.0296
       -0.4611    -0.0785
stats =0.6527     6.5786     0.0247     351.0445
```

可以看出结果不是太好：$p=0.0247$，取 $\alpha=0.05$ 时回归模型（42）可用，但取 $\alpha=0.01$ 则模型不能用；$R^2=0.6527$ 较小；$\hat{\beta}_0,\hat{\beta}_1$ 的置信区间包含了零点。下面将试图用 x_1,x_2 的二次函数改进它。

4.2　多项式回归

如果从数据的散点图上发现 y 与 x 呈较明显的二次（或高次）函数关系，或者用线性模型（20）的效果不太好，就可以选用多项式回归。

4.2.1　一元多项式回归

一元多项式回归可用命令 polyfit 实现。

例 3　将 17 至 29 岁的运动员每两岁一组分为 7 组，每组两人测量其旋转定向能力，以考察年龄对这种运动能力的影响。现得到一组数据如表 3。

表 3

年 龄	17	19	21	23	25	27	29
第一人	20.48	25.13	26.15	30.0	26.1	20.3	19.35
第二人	24.35	28.11	26.3	31.4	26.92	25.7	21.3

试建立二者之间的关系。

解　数据的散点图明显地呈现两端低中间高的形状，所以应拟合一条二次曲线。

选用二次模型

$$y=a_2x^2+a_1x+a_0 \tag{43}$$

编写如下程序：

```
x0=17:2:29;x0=[x0,x0];
y0=[20.48  25.13  26.15  30.0  26.1   20.3  19.35...
     24.35  28.11  26.3   31.4  26.92  25.7  21.3];
```

```
[p,s]=polyfit(x0,y0,2); p
```
得到

p =-0.2003　　8.9782　-72.2150

即 $a_2=-0.2003$，$a_1=8.9782$，$a_0=-72.2150$。

上面的s是一个数据结构，用于计算函数值，如

```
[y,delta]=polyconf(p,x0,s);y
```

得到 y 的拟合值，及预测值 y 的置信区间半径delta。

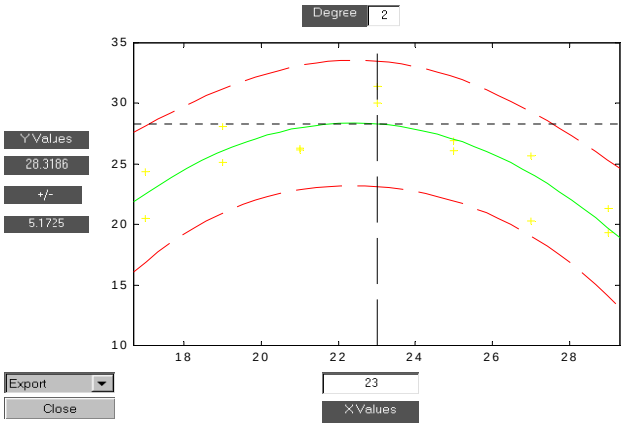

图1　拟合的交互式画面

用polytool(x0,y0,2)，可以得到一个如1图的交互式画面，在画面中绿色曲线为拟合曲线，它两侧的红线是 y 的置信区间。你可以用鼠标移动图中的十字线来改变图下方的 x 值，也可以在窗口内输入，左边就给出 y 的预测值及其置信区间。通过左下方的Export下拉式菜单，可以输出回归系数等。这个命令的用法与下面将介绍的rstool相似。

4.2.2　多元二项式回归

统计工具箱提供了一个作多元二项式回归的命令rstool，它也产生一个交互式画面，并输出有关信息，用法是

rstool(x,y,model,alpha)

其中输入数据x,y分别为 $n\times m$ 矩阵和 n 维向量，alpha为显著性水平 α（缺省时设定为0.05），model由下列4个模型中选择1个（用字符串输入，缺省时设定为线性模型）：

linear(线性)：$y=\beta_0+\beta_1x_1+\cdots+\beta_mx_m$

purequadratic(纯二次)：$y=\beta_0+\beta_1x_1+\cdots+\beta_mx_m+\sum_{j=1}^{m}\beta_{jj}x_j^2$

interaction（交叉）：$y=\beta_0+\beta_1x_1+\cdots+\beta_mx_m+\sum_{1\le j\ne k\le m}\beta_{jk}x_jx_k$

quadratic(完全二次)：$y=\beta_0+\beta_1x_1+\cdots+\beta_mx_m+\sum_{1\le j,k\le m}\beta_{jk}x_jx_k$

我们再作一遍例2 商品销售量与价格问题，选择纯二次模型，即

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_{11}x_1^2+\beta_{22}x_2^2 \qquad (44)$$

编程如下：

```
x1=[120  140  190  130  155  175  125  145  180  150]';
x2=[100  110   90  150  210  150  250  270  300  250]';
y=[102  100   120  77   46   93   26   69   65   85]';
```

```
x=[x1 x2];
rstool(x,y,'purequadratic')
```

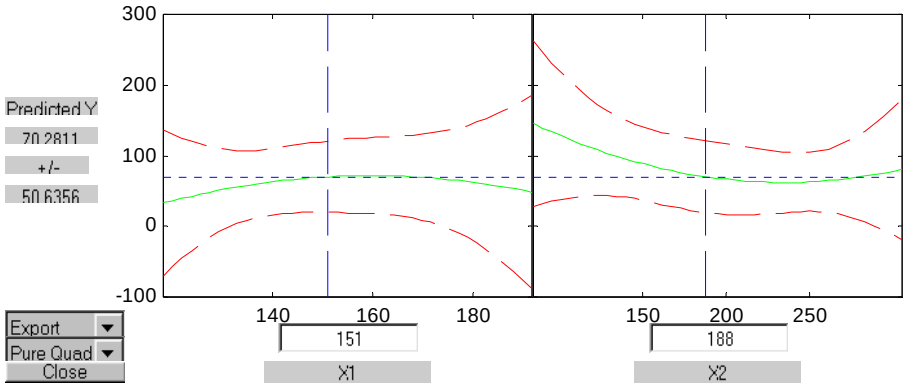

图2　拟合的交互式画面

得到一个如图2所示的交互式画面，左边是 x_1（=151）固定时的曲线 $y(x_1)$ 及其置信区间，右边是 x_2（=188）固定时的曲线 $y(x_2)$ 及其置信区间。用鼠标移动图中的十字线，或在图下方窗口内输入，可改变 x_1, x_2。图左边给出 y 的预测值及其置信区间，就用这种画面可以回答例2提出的“若某市本厂产品售价160（元），竞争对手售价170（元），预测该市的销售量”问题。

图的左下方有两个下拉式菜单，一个菜单Export用以向Matlab工作区传送数据，包括beta(回归系数)，rmse（剩余标准差），residuals(残差)。模型（41）的回归系数和剩余标准差为

beta =−312. 5871　　7. 2701　　−1. 7337　　−0. 0228　　0. 0037

rmse =16.6436

另一个菜单model用以在上述4个模型中选择，你可以比较一下它们的剩余标准差，会发现以模型（24）的rmse=16.6436最小。

注意本例子在Matlab中完全二次模型的形式为

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_1x_2 + b_4x_1^2 + b_5x_2^2 \qquad (45)$$

§5　非线性回归和逐步回归

本节介绍怎样用Matlab统计工具箱实现非线性回归和逐步回归。

5.1 非线性回归

非线性回归是指因变量 y 对回归系数 $\beta_1, \cdots, \beta_m$（而不是自变量）是非线性的。Matlab统计工具箱中的nlinfit，nlparci，nlpredci，nlintool，不仅给出拟合的回归系数，而且可以给出它的置信区间，及预测值和置信区间等。下面通过例题说明这些命令的用法。

例4　在研究化学动力学反应过程中，建立了一个反应速度和反应物含量的数学模型，形式为

$$y = \frac{\beta_4 x_2 - \dfrac{x_3}{\beta_5}}{1 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3}$$

其中 $\beta_1,\cdots,\beta_5$ 是未知的参数， x_1,x_2,x_3 是三种反应物（氢， n 戊烷，异构戊烷）的含量， y 是反应速度。今测得一组数据如表4，试由此确定参数 $\beta_1,\cdots,\beta_5$ ，并给出其置信区间。 $\beta_1,\cdots,\beta_5$ 的参考值为（0.1，0.05，0.02，1，2）。

表4

序号	反应速度 y	氢 x_1	n 戊烷 x_2	异构戊烷 x_3
1	8.55	470	300	10
2	3.79	285	80	10
3	4.82	470	300	120
4	0.02	470	80	120
5	2.75	470	80	10
6	14.39	100	190	10
7	2.54	100	80	65
8	4.35	470	190	65
9	13.00	100	300	54
10	8.50	100	300	120
11	0.05	100	80	120
12	11.32	285	300	10
13	3.13	285	190	120

解　首先，以回归系数和自变量为输入变量，将要拟合的模型写成函数文件huaxue.m：

```
function yhat=huaxue(beta,x);
yhat=(beta(4)*x(:,2)-x(:,3)/beta(5))./(1+beta(1)*x(:,1)+...
beta(2)*x(:,2)+beta(3)*x(:,3));
```

然后，用nlinfit计算回归系数，用nlparci计算回归系数的置信区间，用nlpredci计算预测值及其置信区间，编程如下：

```
clc,clear
x0=[ 1      8.55     470     300      10
 2      3.79     285     80       10
 3      4.82     470     300      120
 4      0.02     470     80       120
 5      2.75     470     80       10
 6      14.39    100     190      10
 7      2.54     100     80       65
 8      4.35     470     190      65
 9      13.00    100     300      54
 10     8.50     100     300      120
 11     0.05     100     80       120
 12     11.32    285     300      10
 13     3.13     285     190      120];
x=x0(:,3:5);
y=x0(:,2);
beta=[0.1,0.05,0.02,1,2]';  %回归系数的初值,任意取的
[betahat,r,j]=nlinfit(x,y,@huaxue,beta);  %r,j是下面命令用的信息
betaci=nlparci(betahat,r,'jacobian',j);
betaa=[betahat,betaci]   %回归系数及其置信区间
[yhat,delta]=nlpredci(@huaxue,x,betahat,r,'jacobian',j)
%y的预测值及其置信区间的半径，置信区间为yhat±delta。
```

用nlintool得到一个交互式画面，左下方的Export可向工作区传送数据，如剩余标

准差等。使用命令

```
nlintool(x,y,'huaxue',beta)
```

可看到画面，并传出剩余标准差rmse= 0.1933。

4.2　逐步回归

实际问题中影响因变量的因素可能很多，我们希望从中挑选出影响显著的自变量来建立回归模型，这就涉及到变量选择的问题，逐步回归是一种从众多变量中有效地选择重要变量的方法。以下只讨论线性回归模型（1）式的情况。

变量选择的标准，简单地说就是所有对因变量影响显著的变量都应选入模型，而影响不显著的变量都不应选入模型，从便于应用的角度应使模型中变量个数尽可能少。

若候选的自变量集合为 $S=\{x_1,\cdots,x_m\}$，从中选出一个子集 $S_1\subset S$，设 S_1 中有 l 个自变量 $(l=1,\cdots,m)$，由 S_1 和因变量 y 构造的回归模型的误差平方和为 Q，则模型的剩余标准差的平方 $s^2=\dfrac{Q}{n-l-1}$，n 为数据样本容量。所选子集 S_1 应使 s 尽量小，通常回归模型中包含的自变量越多，误差平方和 Q 越小，但若模型中包含有对 y 影响很小的变量，那么 Q 不会由于包含这些变量在内而减少多少，却因 l 的增加可能使 s 反而增大，同时这些对 y 影响不显著的变量也会影响模型的稳定性，因此可将剩余标准差 s 最小作为衡量变量选择的一个数量标准。

逐步回归是实现变量选择的一种方法，基本思路为，先确定一初始子集，然后每次从子集外影响显著的变量中引入一个对 y 影响最大的，再对原来子集中的变量进行检验，从变得不显著的变量中剔除一个影响最小的，直到不能引入和剔除为止。使用逐步回归有两点值得注意，一是要适当地选定引入变量的显著性水平 α_{in} 和剔除变量的显著性水平 α_{out}，显然，α_{in} 越大，引入的变量越多；α_{out} 越大，剔除的变量越少。二是由于各个变量之间的相关性，一个新的变量引入后，会使原来认为显著的某个变量变得不显著，从而被剔除，所以在最初选择变量时应尽量选择相互独立性强的那些。

在Matlab统计工具箱中用作逐步回归的是命令stepwise，它提供了一个交互式画面，通过这个工具你可以自由地选择变量，进行统计分析，其通常用法是：

```
stepwise(x,y,inmodel,alpha)
```

其中x是自变量数据，y是因变量数据，分别为 $n\times m$ 和 $n\times 1$ 矩阵，inmodel是矩阵x的列数的指标，给出初始模型中包括的子集（缺省时设定为空），alpha为显著性水平。

Stepwise Regression 窗口，显示回归系数及其置信区间，和其它一些统计量的信息。绿色表明在模型中的变量，红色表明从模型中移去的变量。在这个窗口中有Export按钮，点击Export产生一个菜单，表明了要传送给Matlab工作区的参数，它们给出了统计计算的一些结果。

下面通过一个例子说明stepwise的用法。

例5　水泥凝固时放出的热量 y 与水泥中4种化学成分 x_1,x_2,x_3,x_4 有关，今测得一组数据如表5，试用逐步回归来确定一个线性模型

表5

序号	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

编写程序如下：

```
clc,clear
x0=[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];
x=x0(:,2:5);
y=x0(:,6);
stepwise(x,y)
```

得到图3所示的图形界面。

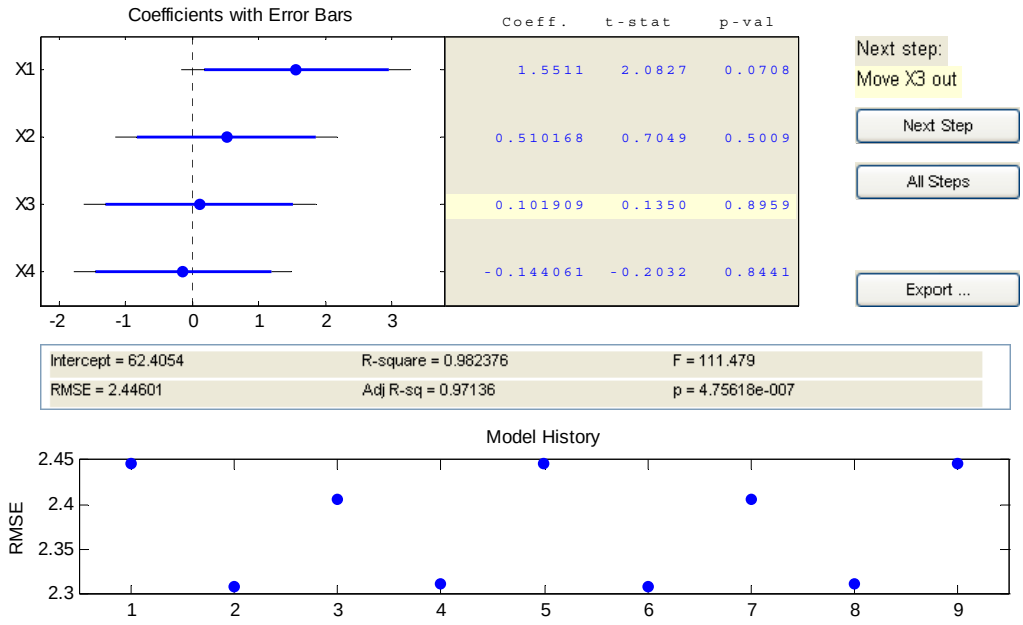

图3　逐步回归交互式画面

可以看出，x_3, x_4 不显著，移去这两个变量后的统计结果如图4。

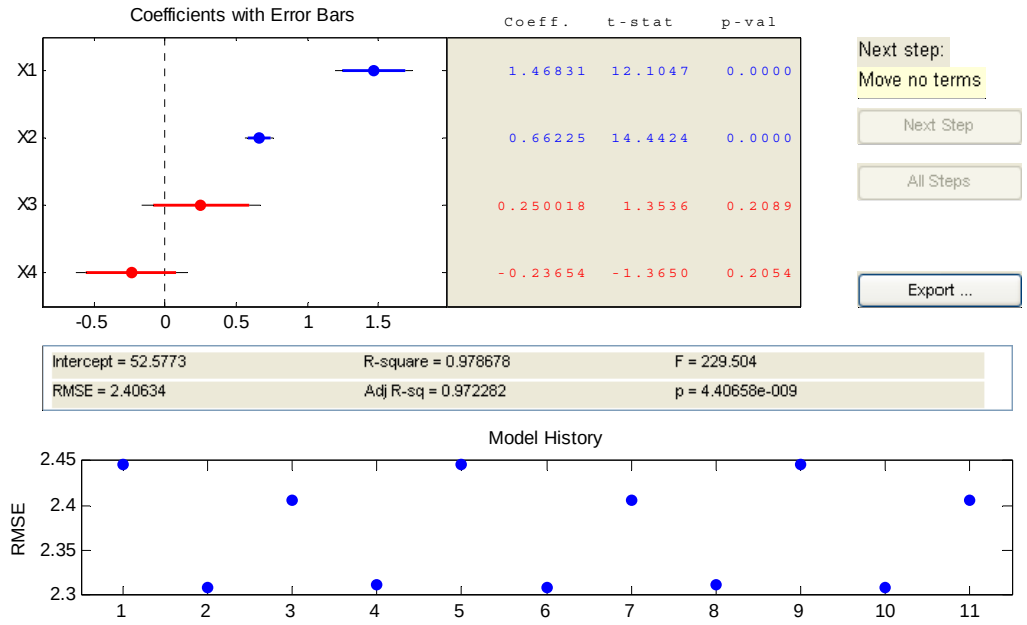

图4　逐步回归交互式画面

这个表中的 x_3, x_4 两行用红色显示，表明它们已移去。

从新的统计结果可以看出，虽然剩余标准差 s（RMSE）没有太大的变化，但是统计量 F 的值明显增大，因此新的回归模型更好一些。可以求出最终的模型为

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

习 题 十 二

1. 某人记录了21天每天使用空调器的时间和使用烘干器的次数，并监视电表以计算出每天的耗电量，数据见表6，试研究耗电量（KWH）与空调器使用的小时数(AC)和烘干器使用次数(DRYER)之间的关系，建立并检验回归模型，诊断是否有异常点。

表6

序号	1	2	3	4	5	6	7	8	9	10	11
KWH	35	63	66	17	94	79	93	66	94	82	78
AC	1.5	4.5	5.0	2.0	8.5	6.0	13.5	8.0	12.5	7.5	6.5
DRYER	1	2	2	0	3	3	1	1	1	2	3
序号	12	13	14	15	16	17	18	19	20	21	
kWH	65	77	75	62	85	43	57	33	65	33	
AC	8.0	7.5	8.0	7.5	12.0	6.0	2.5	5.0	7.5	6.0	
DRYER	1	2	2	1	1	0	3	0	1	0	

2. 在一丘陵地带测量高程，x 和 y 方向每隔100米测一个点，得高程如下表，试拟合一曲面，确定合适的模型，并由此找出最高点和该点的高程。

表7

x / y	100	200	300	400
100	636	697	624	478
200	698	712	630	478
300	680	674	598	412
400	662	626	552	334

3. 一矿脉有13个相邻样本点，人为地设定一原点，现测得各样本点对原点的距离 x，与该样本点处某种金属含量 y 的一组数据如下，画出散点图观测二者的关系，试建立合适的回归模型，如二次曲线、双曲线、对数曲线等。

表8

x	2	3	4	5	7	8	10
y	106.42	109.20	109.58	109.50	110.00	109.93	110.49
x	11	14	15	16	18	19	
y	110.59	110.60	110.90	110.76	111.00	111.20	