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

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

样本协方差矩阵及样本相关系数矩阵分别为

$$C_1 = (t_{ij})_{m\times m} = \frac{1}{n-1}\sum_{k=1}^{n}(e_k - \bar{x})(e_k - \bar{x})^T$$

$$C_2 = (r_{ij})_{m\times m} = \left(\frac{t_{ij}}{\sqrt{t_{ii}t_{jj}}}\right)$$

其中

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

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

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

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

$$x_{ij}^{*}=x_{ij}/\max_i\{x_{ij}\},\ \ x_{ij}^{*}=x_{ij}/\min_i\{x_{ij}\}$$

$$x_{ij}^{*}=x_{ij}/\bar{x}_j,\ \ x_{ij}^{*}=x_{ij}/(\max_i\{x_{ij}\}-\min_i\{x_{ij}\})$$

（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

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

§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

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

这 n 对观测值之间的关系符合模型

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

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

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

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

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

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

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

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 $\sigma^2 = \mathrm{Var}(\varepsilon_i)$ 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 $SSE = SST$ 。

判定系数是一个很有趣的指标：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

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

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

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，则 SSE/σ^2 与 SSR/σ^2 是独立的随机变量，且

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

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

$$H_0:\beta_1=0，\ H_1:\beta_1\ne 0$$

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，这两个检验的意义是不同的。从逻辑上说，一般常在 F 检验通过后，再进一步进行 t 检验。

2.4.2　回归系数的显著性检验

回归参数的检验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$\beta_0$显著为零；

若$|t_0| > t_{\frac{\alpha}{2}}(n-2)$，则拒绝$H_0$，认为$\beta_0$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 $\beta_j=\hat{\beta}_j$，$j=0,1,2,\cdots,m$ 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

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

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）。

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

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

$R = \sqrt{R^2}$ 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$m+1$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$\sigma^2$，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。

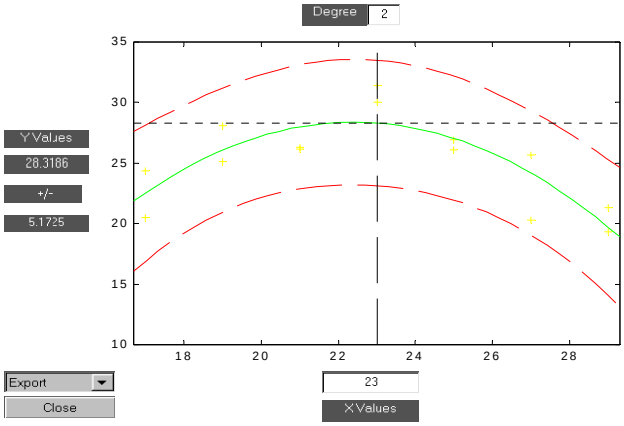

图1　拟合的交互式画面

用polytool(x0,y0,2)，可以得到一个如图1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）：$y=\beta_0+\beta_1x_1+\cdots+\beta_mx_m+\sum_{1\le j<k\le m}\beta_{jk}x_jx_k$

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

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')
```

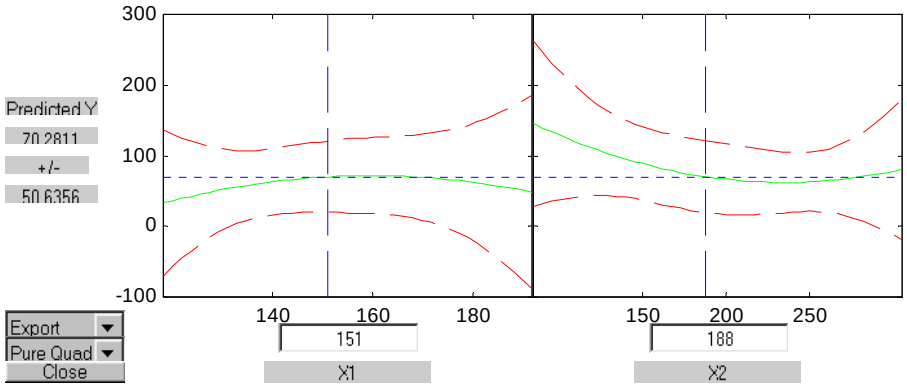

图2　拟合的交互式画面

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（44）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（44）的rmse=16.6436最小。

注意本例子在Matlab中完全二次模型的形式为

$$y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_1 x_2 + b_4 x_1^2 + b_5 x_2^2 \qquad (45)$$

§5　偏相关系数

在研究两个变量之间的线性相关程度时，可考察这两个变量的简单相关系数。但在研究多个变量之间的线性相关程度时，单纯使用两两变量的简单相关系数常具有虚假性。因为简单相关系数只考虑了两个变量之间的相互作用，而没有考虑其它变量对这两个变量的影响。为了更准确、真实地反映变量之间的相关关系，统计学中定义了偏相关系数（又称净相关系数）。

5.1　偏相关系数的定义

如果有因变量 y 和自变量 $x_1, x_2, \cdots, x_m$，怎样定义 y 与 x_1 的偏相关系数？一个直观的想法是在除掉 $x_2, x_3, \cdots, x_m$ 的影响之后，再考虑 y 与 x_1 的相关程度。

如果有 n 个样本 $i = 1,2,\cdots,n$，考虑下面两个回归模型

$$y_i = c_0 + c_2 x_{i2} + \cdots + c_m x_{im} + \varepsilon'_i$$

$$x_{i1} = d_0 + d_2 x_{i2} + \cdots + d_m x_{im} + \varepsilon''_i$$

利用最小二乘法可求得这两个模型的拟合模型，并分别求出它们的残差为

$$u_i = y_i - (\hat{c}_0 + \hat{c}_2 x_{i2} + \cdots + \hat{c}_m x_{im})$$

$$v_i = x_{i1} - (\hat{d}_0 + \hat{d}_2 x_{i2} + \cdots + \hat{d}_m x_{im})$$

求这两个残差向量 $u = (u_1, u_2, \cdots, u_n)^T$ 与 $v = (v_1, v_2, \cdots, v_n)^T$ 的简单相关系数，记为 $r_{y1\bullet 2,\cdots,m}$，称它为 y 与 x_1 的偏相关系数。

例如只有两个自变量 x_1, x_2 的情形。为方便起见，不失一般性，设 y, x_1, x_2 均为中心化变量，则有

$u_i = y_i - \hat{c}_2 x_{i2}$，$v_i = x_{i1} - \hat{d}_2 x_{i2}$，$i = 1,2,\cdots,n$

由于是中心化变量，所以两个模型的常数项均为零，即

$$\mathrm{Var}(u) = \sum_{i=1}^{n} u_i^2 = \sum_{i=1}^{n} (y_i - \hat{c}_2 x_{i2})^2 = \sum_{i=1}^{n} y_i^2 - \hat{c}_2^2 \sum_{i=1}^{n} x_{i2}^2$$

$$= \sum_{i=1}^{n} y_i^2 - \left(\frac{\sum_{i=1}^{n} y_i^2}{\sum_{i=1}^{n} x_{i2}^2} r_{y2}^2 \right) \sum_{i=1}^{n} x_{i2}^2 = (1 - r_{y2}^2) \sum_{i=1}^{n} y_i^2$$

同理

$$\mathrm{Var}(v) = (1 - r_{12}^2) \sum_{i=1}^{n} x_{i1}^2$$

$$\mathrm{Cov}(u, v) = (r_{y1} - r_{y2} r_{12}) \sqrt{\sum_{i=1}^{n} y_i^2 \sum_{i=1}^{n} x_{i1}^2}$$

这里，r_{12} 是 x_1 与 x_2 的简单相关系数，r_{y1} 和 r_{y2} 分别是 y 与 x_1 及 x_2 的简单相关系数。所以

$$r_{y1\bullet 2} = r(u, v) = \frac{\mathrm{Cov}(u, v)}{\sqrt{\mathrm{Var}(u)\mathrm{Var}(v)}} = \frac{r_{y1} - r_{y2} r_{12}}{\sqrt{(1 - r_{y2}^2)(1 - r_{12}^2)}}$$

在更一般的情况下， 有

$$r_{y1\bullet 2,\cdots,m} = \frac{t_1}{\sqrt{t_1^2 + n - m - 1}}$$

其中 t_1 是回归方程 $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_m x_m$ 中， x_1 的 t 检验值。

5.2　偏相关系数的检验

设总体的偏相关系数为 $\rho_{ij\bullet}$，样本偏相关系数为 $r_{ij\bullet}$， n 为样本容量， p 为随机变量个数。

$$H_0 : \rho_{ij\bullet} = 0$$

当 H_0 成立时，检验统计量

$$F = \frac{r_{ij\bullet}^2 (n - p)}{1 - r_{ij\bullet}^2} \sim F(1, n - p)$$

给定显著水平α，可查表得到临界值$F_\alpha(1,n-p)$。决定准则为，对于统计量的值F：

若$F>F_\alpha(1,n-p)$，则否定H_0，说明x_i与x_j之间存在显著的净相关关系；

若$F\le F_\alpha(1,n-p)$，则肯定H_0，说明x_i与x_j之间不存在显著的净相关关系。

§6 变量筛选方法

在实际工作中使用过多元回归分析的人都知道，用回归建模首先遇到的难题，就是选择哪些变量作为因变量的解释变量。在我们选择自变量时，一方面希望尽可能不遗漏重要的解释变量；另一方面，又要遵循参数节省原则，使自变量的个数尽可能少。因为当自变量数目过大时，模型计算复杂，且往往会扩大估计方差，降低模型精度。

在确定自变量系统时，一是采用穷举法，列举出所有可能的潜在自变量；再根据自变量的不同组合，选取最合适的模型。由于每个变量都有可能被选用或不被选用，所以，穷举法要拟合与比较的方程个数为2^m（m为潜在自变量的个数）。

当备选的潜在自变量数目很大时，则采用穷举方法就完全不现实了。下面我们介绍一些有效的变量筛选方法，向前选择变量法、向后删除变量法和逐步回归法。

6.1 偏F检验

在决定一个新的变量是否有必要进入模型，或者判断某个变量是否可以从模型中删除时，我们首先要问的问题是：这个变量能否对y提供显著的附加解释信息？回答这个问题的方法是采用偏F检验。

设有m个自变量$x_1,x_2,\cdots,x_m$，采用这m个自变量拟合的模型称为全模型，即

$$y=\beta_0+\beta_1x_1+\cdots+\beta_mx_m+\varepsilon$$

从这m个变量中删去自变量x_j,这时用$m-1$个自变量拟合模型称为减模型，即

$$y=\beta_0+\beta_1x_1+\cdots+\beta_{j-1}x_{j-1}+\beta_{j+1}x_{j+1}+\cdots+\beta_mx_m+\varepsilon$$

全模型的复判定系数为R^2，减模型的复判定系数记为R_j^2。定义

$$\Delta R_j^2=R^2-R_j^2$$

由于在全模型中多一个自变量x_j，所以，若ΔR_j^2几乎为零，说明增加x_j，对y的解释能力没有显著提高；否则，若ΔR_j^2显著不为零，则x_j就可以为回归模型提供显著的解释信息。

给出统计假设$H_0:\Delta R_j^2=0$，$H_1:\Delta R_j^2\ne0$

统计检验量为

$$F_j=\frac{Q_j-Q}{Q/(n-m-1)}$$

式中，Q_j是减模型的残差平方和，Q为全模型的残差平方和。

在H_0假设成立的条件下，F_j服从F分布，第一个自由度为1，第二个自由度为$n-m-1$。

根据检验水平α查F分布表（$n_1=1,n_2=n-m-1$），得到拒绝域的临界值F_α，则决策准则如下：

（1）当 $F_j > F_\alpha$ 时，拒绝 H_0，说明 ΔR_j^2 显著不为零，这说明在 $x_1,\cdots,x_{j-1},x_{j+1},\cdots,x_m$ 变量已进入模型后，引入 x_j 会显著提高对 y 的解释能力；

（2）当 $F_j \le F_\alpha$ 时，接受 H_0，ΔR_j^2 显著为零，所以，从全模型中删除 x_j，对 y 的解释能力无明显的减弱变化。

上述检验就称为偏 F 检验。偏 F 检验就是变量筛选的统计依据。

6.2　向前选择变量法

向前选择变量法在起始时，模型中没有任何变量。然后，分别考虑 y 与每一个自变量的一元线性回归模型。对所有的这 m 个模型进行 F 检验，选择 F 值最高者作为第一个进入模型的自变量（记为 x_{i_1}）。

然后，对剩下的 $m-1$ 个变量分别进行偏 F 检验（即以 y 与 x_{i_1} 的模型为减模型，以 y 与 x_{i_1} 以及另一个自变量 x_j 的模型为全模型）。如果至少有一个 x_i 通过了偏 F 检验，则在所有通过偏 F 检验的变量中，选择 F_j 值最大者作为第二个被选的自变量，进入模型（记为 x_{i_2}）。

继续上述步骤，直到在模型之外的自变量均不能通过偏 F 检验，则算法终止。

6.3　向后删除变量法

向后删除变量法的工作方法正好与向前选择变量法完全相反。在算法的起步，所有的自变量都被包含在模型之中（这是起始的全模型）。然后，依次对每一个自变量做偏 F 检验（以去掉变量 x_j 的模型为减模型）。如果所有的自变量都通过了偏 F 检验，则计算停止，所有自变量被包含在模型中。如果有若干自变量未能通过偏 F 检验，则选择出 F_j 值最小的自变量，将它从模型中删除。

对剩下的 $(m-1)$ 个自变量拟合一个全模型。然后，重新对每一个模型中的自变量进行偏 F 检验。在没有通过检验的自变量中，选择 F_j 值最小者，将它从模型中删除。

重复以上步骤，直到模型中包含的所有自变量都能通过偏 F 检验，则算法终止。

6.4　逐步回归

逐步回归法是人们最常选用的变量筛选方法。它是向前选择变量法和向后删除变量法的一种结合。在向前选择变量法中，一旦某个自变量被选入模型，它就永远留在模型之中。然而，随着其它变量的引入，由于变量之间相互传递的相关关系，一些先进入模型的变量的解释作用可能会变得不再显著。而对于向后删除变量法，一旦某个自变量被删除后，它就永远被排斥在模型之外。但是，随着其它变量的被删除，它对 y 的解释作用也可能会显著起来。

所以，逐步回归法采取边进边退的方法。对于模型外部的变量，只要它还可提供显著的解释信息，就可以再次进入模型；而对于已在内部的变量，只要它的偏 F 检验不能通过，则还可能从模型中被删除。

模型的起始与向前选择变量法一样。首先，求 y 与每一个 x_i 的一元线性回归方程，选择 F 值最大的变量进入模型。然后，对剩下的 $m-1$ 个模型外的变量进行偏 F 检验（设定 x_{i_1} 已在模型中），在若干通过偏 F 检验的变量中，选择 F_j 值最大者进入模型。

再对模型外的 $m-2$ 个自变量做偏 F 检验。在通过偏 F 检验的变量中选择 F_j 值最大者进入模型。接着对模型中的三个自变量分别进行偏 F 检验，如果三个自变量都通

过了偏F检验，则接着选择第四个变量。但如果有某一个变量没有通过偏F检验，则将其从模型中删除。

重复上述步骤，直到所有模型外的变量都不能通过偏F检验，则算法终止。为了避免变量的进出循环，一般取偏F检验拒绝域的临界值为

$$F_{进} > F_{出}$$

式中，$F_{进}$为选入变量时的临界值；$F_{出}$为删除变量时的临界值。在所有标准的统计软件中都有逐步回归的程序。$F_{进}$和$F_{出}$的检验水平值可以自定，也可以是备择的。常见的检验水平值为$\alpha_{进}=0.05$，$\alpha_{出}=0.1$。

统计学家主张在回归建模时，应采用尽可能少的自变量，不要盲目地追求复判定系数R^2的提高。其实，当变量增加时，残差项的自由度就会减少（$df_E=n-m-1$）。当$m=n-1$时，$df_E=0$。而自由度越小，数据的统计趋势就越不容易显现。为此，又定义一个调整复判定系数

$$\overline{R}^2=1-\frac{Q/(n-m-1)}{SST/(n-1)} \tag{46}$$

可见，在调整复判定系数中考虑了自由度的因素。$\overline{R}^2$与R^2的关系是

$$\overline{R}^2=1-(1-R^2)\frac{n-1}{n-m-1} \tag{47}$$

当n很大、m很少时，$\overline{R}^2$与R^2之间的差别不是很大；但是，当n较少，而m又较大时，$\overline{R}^2$就会远小于R^2。

在一般的统计软件中，常在输出中同时给出R^2和$\overline{R}^2$。如果它们相差过大，则应考虑减少或调整变量。

另外，有关$\overline{R}^2$的比较，还可以用于判断是否可以再增加新的变量。如果增加一个变量后，$\overline{R}^2$的改观不大，则这个变量的增加，意义就不大。所以，只有当$\overline{R}^2$明显增加时，才考虑增加此变量。

在SPSS等统计软件的工作过程中，则采用一种更直观的操作方法。其主要理论依据是当模型中已经包含了k个自变量$x_1,x_2,\cdots,x_k$，如果要再增加一个新自变量x_j，则这时的偏F检验值为

$$F_j=t_j^2$$

这里，t_j是以$x_1,x_2,\cdots,x_k,x_j$为自变量时回归模型中x_j的t检验值。

下面通过一个例题来看一下SPSS软件中变量筛选的工作过程。

例4　某产品的销售额y与部门的全部市场销售额x_1，给批发商的优惠x_2，价格x_3，开发预算x_4，投资x_5，广告x_6，销售费用x_7，部门全部广告的预算x_8有关。为预测未来的销售量，收集了38个样本点的有关数据见表4，试建立y的经验公式。

表4　原始数据表

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	y
398	138	56.205	12.112	49.895	76.862	228.90	98.205	5540.39
369	118	59.044	9.330	16.595	88.805	177.45	224.953	5439.04
268	129	56.723	28.748	89.182	51.297	166.40	263.032	4290.00
484	111	57.862	12.891	106.738	39.747	285.05	320.928	5502.34

394	146	59.117	13.381	142.552	51.651	209.30	406.989	4871.77
332	140	60.111	11.085	61.287	20.547	180.05	246.996	4708.08
336	136	59.839	24.957	-30.385	40.153	213.20	328.436	4627.81
383	104	60.052	20.809	-44.586	31.645	200.85	298.456	4110.24
285	105	63.141	8.485	-28.373	12.457	176.15	218.110	4122.69
277	135	62.302	10.730	75.723	68.307	174.85	410.467	4842.25
456	128	64.922	21.874	144.030	52.453	252.85	93.006	5740.65
355	131	64.857	23.506	112.904	76.677	208.00	307.226	5094.10
364	120	63.591	13.894	128.347	96.067	195.00	106.792	5383.20
320	147	65.614	14.865	10.097	47.979	154.05	304.921	4888.17
311	143	67.022	22.494	-24.760	27.231	180.70	59.612	4003.13
362	145	66.904	23.269	116.748	72.668	219.70	238.986	4941.96
408	131	66.184	13.035	120.406	62.312	234.65	141.074	5312.80
433	124	67.865	8.033	121.823	24.712	258.05	290.832	5139.87
359	106	68.889	27.048	71.055	73.912	196.30	413.636	5397.36
476	138	71.417	18.220	4.186	63.273	278.85	206.454	5149.47
415	148	69.277	7.742	46.935	28.676	207.35	79.566	5150.83
420	136	69.733	10.136	7.621	91.363	213.20	428.982	4989.02
536	111	73.162	27.370	127.509	74.016	296.40	273.072	5926.86
432	152	73.365	15.528	-49.574	16.162	245.05	309.422	4703.88
436	123	73.050	32.491	100.098	42.998	275.60	280.139	5365.59
415	119	74.910	19.712	-40.183	41.134	211.25	314.548	4630.09
462	112	73.200	14.835	68.153	92.518	282.75	212.058	5711.86
429	125	74.161	11.369	87.963	83.287	217.75	118.065	5095.48
517	142	74.283	26.751	27.098	74.892	306.90	344.553	6124.37
328	123	77.140	19.603	59.343	87.510	210.60	140.872	4787.34
418	135	78.591	34.688	141.969	74.471	269.75	82.855	5035.62
515	120	77.093	23.202	126.420	21.271	328.25	398.425	5288.01
412	149	78.231	35.739	29.558	26.494	258.05	124.027	4647.01
455	126	77.929	21.589	18.007	94.631	232.70	117.911	5315.63
554	138	81.039	19.569	42.352	92.544	323.70	161.250	6180.06
441	120	79.848	15.503	-21.558	50.048	267.15	405.088	4800.97
417	120	80.639	34.923	148.450	83.180	257.40	110.740	5512.13
461	132	82.284	26.549	-17.584	91.221	266.50	170.392	5272.21

1．全模型

首先，以 $x_1 \sim x_8$ 为全部自变量，采用最小二乘法拟合一个多元回归模型，有

$$\hat{y} = 3086.2941 + 4.4862x_1 + 1.7618x_2 - 12.9833x_3 - 3.6697x_4$$
$$+ 1.9442x_5 + 8.5707x_6 + 1.3531x_7 - 0.0086x_8$$

这个回归模型的复判定系数 $R^2 = 0.8048$，调整复判定系数 $\overline{R}^2 = 0.7509$。模型的剩余标准差为257.8049。

对模型进行 F 检验：$F = 14.9424$。

对各参数进行 t 检验的结果见表5。

表5　8个自变量模型的 t 检验结果

t	常量	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8
t 检验值	4.868	2.7119	0.5314	-1.5893	-0.5565	2.4588	4.612	0.487	-0.0214

计算的Matlab程序如下：

```
clc,clear
load data.txt    %表中的数据按照原来的排列存放在纯文本文件data.txt中
[n,m]=size(data);m=m-1;
x=[ones(38,1),data(:,1:8)]; y=data(:,9);
[b,bint,r,rint,stats]=regress(y,x)  %stats(4)返回的是残差的样本方差
r2=stats(1)   %提出复判定系数
ad_r2=1-(1-r2)*(n-1)/(n-m-1)   %计算调整复判断系数
f=stats(2)      %提出F统计量
tm=inv(x'*x);   %计算X'*X的逆矩阵
tm=diag(tm);    %提出逆矩阵的对角线元素
rmse=sqrt(stats(4))  %计算剩余标准差（残差的样本标准差）
t=b./sqrt(tm)/rmse   %求t统计量的值
```

从这个模型看，F 检验通过，但在t检验中有若干自变量对 y 的解释作用不明显，并且复判定系数与调整复判定系数的差距也比较大，因此，可以考虑对自变量集合进行调整。

2．向前选择变量法

首先，计算销售额与8个自变量之间的简单相关系数（见表6）。选取相关系数最大的自变量x_1首先进入模型。

表6　y 与 $x_1 \sim x_8$ 的简单相关系数

变量	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
y	0.7205	-0.0849	0.2881	0.0811	0.4538	0.5662	0.6411	-0.0933

x_1首先进入模型：$\hat{y} = 2950.0685 + 5.2835x_1$

复判定系数$R^2 = 0.5191$，剩余标准差为363.1402，

F 检验值：$F = 38.8644$，t检验值：$t_0 = 8.448$，$t_1 = 6.2341$。

除x_1以外，还有7个自变量在模型之外。在考虑到x_1已经在模型之中的基础上，分别计算x_j（$j = 2,3,\cdots,8$）进入模型后（即以x_1, x_j为模型中的自变量），x_j变量的t检验值t_j，并计算$x_2 \sim x_8$与y的偏相关系数$r_{yx_j \bullet x_1}$。有关的计算结果见表7。

表7　向前选择变量法的变量选择表

待选变量	x_2	x_3	x_4	x_5	x_6	x_7	x_8
t检验值	-0.2994	-1.1665	-0.2998	3.5013	4.6982	-0.273	-0.679
偏相关系数	-0.0505	-0.1935	-0.0506	0.5093	0.6219	-0.0461	-0.114

计算的Matlab程序如下：

```
clc,clear
load data.txt    %表中的数据按照原来的排列存放在纯文本文件data.txt中
[n,m]=size(data);
y=data(:,9);
TT=[];
```

```
for i=2:8
x=[ones(38,1),data(:,[1,i])];
[b,bint,r,rint,stats]=regress(y,x)  %stats(4)返回的是残差的样本方差
tm=inv(x'*x);   %计算X'*X的逆矩阵
tm=diag(tm);    %提出逆矩阵的对角线元素
rmse=sqrt(stats(4))  %计算剩余标准差（残差的样本标准差）
t=b./sqrt(tm)/rmse   %求t统计量的值
TT=[TT,t];
end
ts=TT(3,:)      %求各个新加入变量的t检验值
pr=ts./sqrt(ts.^2+n-3)  %计算偏相关系数
```

第1步，选择偏相关系数最大的自变量 x_6 进入模型，并且在以 x_1, x_6 为自变量的模型中， x_6 的 t 检验通过。前面已经指出， t_j^2 即等于对 x_j 的偏 F 检验值。

完全同于第1步，可以进行后续变量的选择。模型在第3步终止计算，得到最终模型为

$$\hat{y} = 2721.6851 + 4.4372x_1 + 2.2726x_5 + 7.5101x_6$$

对该模型，有 $R^2 = 0.7755$， $\overline{R}^2 = 0.7557$，剩余标准差为 255.313， $F = 39.1511$。

3．向后删除变量法

此种方法更易于掌握。它第1步以全部自变量 $x_1 \sim x_8$ 作为解释变量拟合方程。然后，每一步都在未通过 t 检验的自变量中选择一个 $|t_j|$ 值最小的变量，将它从模型中删除。直至某一步，所有的自变量均通过 t 检验，则算法终止。

规定的检验水平为 $\alpha_{出} = 0.10$，所以在第5步，所有自变量的回归系数均通过 t 检验，无被删除的变量，计算终止。最终模型为

$$\hat{y} = 3293.8304 + 5.2235x_1 - 1.3261x_3 + 1.9661x_5 + 8.2469x_6$$

对该模型，有 $R^2 = 0.8003$， $\overline{R}^2 = 0.7826$，剩余标准差为 244.4437， $F = 33.0554$。

由于本例中逐步回归法的计算结果与向前选择变量法相同，因此，这里不再详述。

6.5　Matlab统计工具箱中的逐步回归命令

在Matlab统计工具箱中用作逐步回归的命令是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8，试用逐步回归来确定一个线性模型

表8

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,[1:4])
```

运行上述程序，得到图3所示的图形界面。

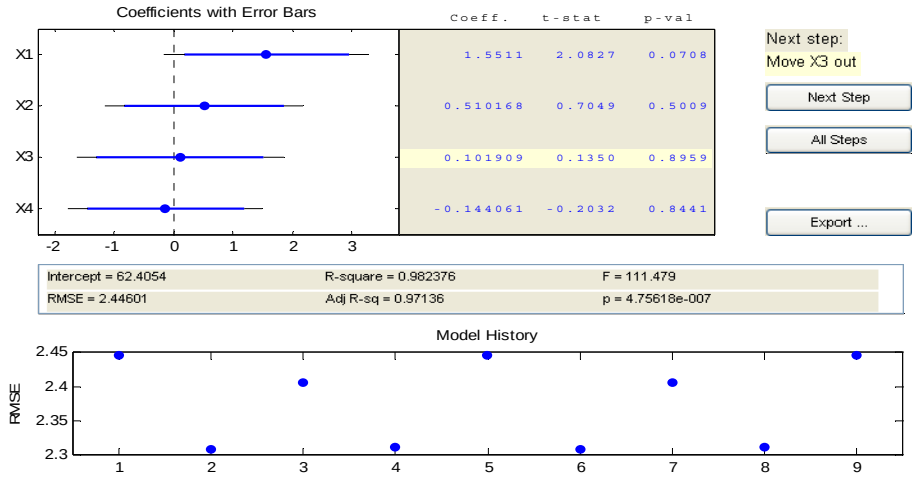

图3　逐步回归交互式画面

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。

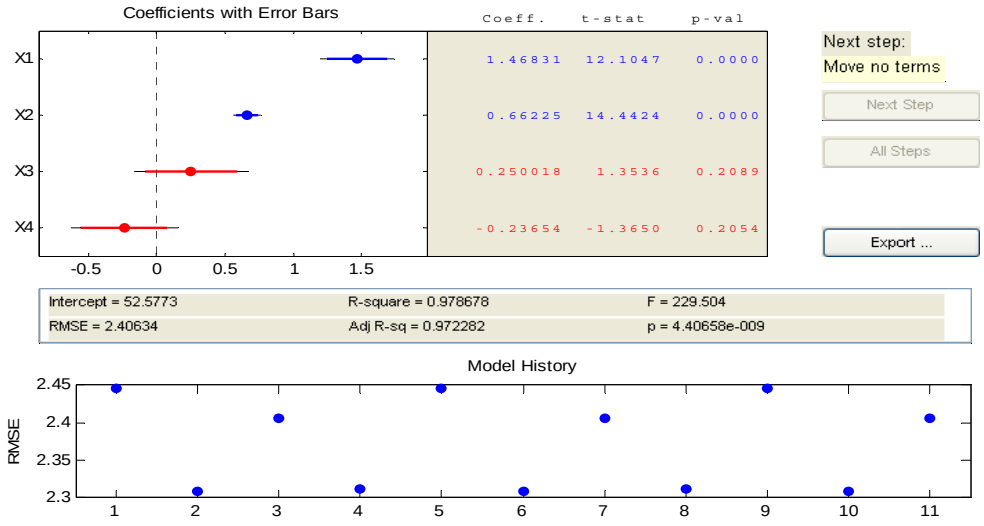

图4　逐步回归交互式画面

图4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

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

§7　复共线性与有偏估计方法

前面我们详细讨论了回归系数的最小二乘估计，并且证明了它的许多优良性质。随着电子计算机技术的飞速发展，人们愈来愈多地有能力去处理含较多回归自变量的大型回归问题。许多应用实践表明，在这些大型线性回归问题中，最小二乘估计不是总令人满意。例如，有时某些回归系数的估计值的绝对值差异较大，有时回归系数的估计值的符号与问题的实际意义相违背等。研究结果表明，产生这些问题的原因之一是回归自变量之间存在着近似线性关系，称为复共线性（Multicollinearity）。

7.1　复共线性

以下为方便起见，对 n 个独立观测数据 $(y_i, x_{i1}, \cdots, x_{im})$，$i = 1, \cdots, n, n > m$，进行了标准化，即转化成 $(y_i^*, x_{i1}^*, \cdots, x_{im}^*)$，其中

$$y_i^* = \frac{y_i - \bar{y}}{s_y}，\text{这里 } \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i，\quad s_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$x_{ij}^* = \frac{x_{ij} - \bar{x}_j}{s_j}，\text{这里 } \bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}，\quad s_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}$$

对应的标准化变量记为 $y^*, x_1^*, \cdots, x_n^*$。

数据标准化后，回归模型中常数项为零。不失一般性，我们研究以下回归模型

$$\begin{cases} Y = X\beta + \varepsilon \\ \varepsilon \sim N(0, \sigma^2 E_n) \end{cases} \tag{48}$$

其中模型的设计矩阵

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

一般假设 $\mathrm{rank}(X) = m$，即为满秩矩阵，则正规方程组的系数矩阵 $X^T X$ 为满秩阵。如果用 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$ 表示 $X^T X$ 的 m 个特征值，且当 $\left|X^T X\right| = \lambda_1 \lambda_2 \cdots \lambda_m$ 很小，即至少有一个特征根接近于0（例如 λ_m 接近于0），但不等于0时，则使正规方程组

$$X^T X\hat{\beta} = X^T Y$$

变成为一种病态方程。

虽然 $\hat{\beta}$ 是 β 的无偏估计，即 $E(\hat{\beta}) = \beta$，但其均方误差

$$E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)] = \sigma^2 (\frac{1}{\lambda_1} + \frac{1}{\lambda_2} + \cdots + \frac{1}{\lambda_m})$$

充分的大，即使 β 的估计值 $\hat{\beta}$ 的误差太大，无实用价值。此时称 m 个解释变量之间具有复共线性，也就是说设计矩阵 X 的列向量之间有近似的线性关系，但非绝对的线性关系。衡量复共线性程度的量用

$$K = \frac{\text{最大特征值}}{\text{最小特征值}} = \frac{\lambda_1}{\lambda_m}$$

来表示.

（1）当 $K < 100$ 时，则不存在复共线性；

（2）当 $100 \leq K \leq 1000$ 时，则存在较强的复共线性；

（3）当 $K > 1000$ 时，则存在严重的复共线性。

我们先引进评价一个估计优劣的标准－均方误差（Mean Squared Errors，以下简记为 mse）。

设 θ 为 $m \times 1$ 未知参数向量，$\tilde{\theta}$ 为 θ 的一个估计，定义 $\tilde{\theta}$ 的均方误差为

$$mse(\tilde{\theta}) = E\left\|\tilde{\theta} - \theta\right\|^2 = E[(\tilde{\theta} - \theta)^T (\tilde{\theta} - \theta)] \tag{49}$$

它度量了估计 $\tilde{\theta}$ 跟未知参数向量 θ 平均偏离的大小。一个好的估计应该有较小的均方误差。

7.2　岭估计

统计学界由A. E. Hoerl在1962年提出并和R. W. Kennard在1970年系统发展的岭回归（Ridge Regression）方法，可以显著改善设计矩阵列复共线性时最小二乘估计量的均方误差，增强估计的稳定性。这个方法在计算数学称为阻尼最小二乘，出现的较早一些。

岭回归方法主要就是在病态的矩阵 $(X^T X)$ 中沿主对角线人为地加进正数，从而使 λ_m 稍大一些。我们知道模型（48）中 β 的最小二乘估计为

$$\hat{\beta}_L = (X^T X)^{-1} X^T Y \tag{50}$$

则β的岭估计定义为

$$\hat{\beta}(k)=(X^TX+kE_m)^{-1}X^TY \tag{51}$$

从上式直接看出，当$k=0$时，它就是最小二乘估计，最有无偏性；当$k\to+\infty$，$\hat{\beta}(k)\to 0$，就没有意义了。k究竟取多大值为好？显然应该是尽可能小的k能使$\hat{\beta}(k)$尽可能地稳定下来。同时我们需要知道$\hat{\beta}(k)$的统计性质究竟如何。

性质1　岭估计不再是无偏估计量，即$E(\hat{\beta}(k))\neq\beta$。

因为

$$\begin{aligned}E(\hat{\beta}(k))&=E[(X^TX+kE_m)^{-1}X^TY]=(X^TX+kE_m)^{-1}X^TX\beta\\&=[(X^TX+kE_m)^{-1}((X^TX)^{-1})^{-1}]\beta=[E_m+k(X^TX)^{-1}]^{-1}\beta\end{aligned}$$

无偏性一直被认为是一个好的统计量所必须具有的基本性质，但是在现在所讨论的问题场合，我们只好牺牲无偏性，以改善估计的稳定性。

性质2　记$S=X^TX$，$Z_k=(E_m+kS^{-1})^{-1}$，则Z_k的特征值都在$(0,1)$内。

设有正交矩阵P，使得

$$P^TSP=\begin{bmatrix}\lambda_1&&\\&\ddots&\\&&\lambda_m\end{bmatrix}=\Lambda=\mathrm{diag}(\lambda_1,\cdots,\lambda_m) \tag{52}$$

则

$$\begin{aligned}P^TZ_kP&=P^T(E_m+kS^{-1})^{-1}P=(P^T(E_m+kS^{-1})P)^{-1}=(E_m+k\Lambda^{-1})^{-1}\\&=\begin{bmatrix}1+\dfrac{k}{\lambda_1}&&\\&\ddots&\\&&1+\dfrac{k}{\lambda_m}\end{bmatrix}^{-1}=\begin{bmatrix}\dfrac{\lambda_1}{\lambda_1+k}&&\\&\ddots&\\&&\dfrac{\lambda_m}{\lambda_m+k}\end{bmatrix}\hat{=}\Lambda(k)\end{aligned}$$

故知Z_k的特征根分别为$\dfrac{\lambda_i}{\lambda_i+k}$，都在$(0,1)$内。

性质3　岭估计是压缩估计，即$\left\|\hat{\beta}(k)\right\|\le\left\|\hat{\beta}\right\|$。

性质4　存在常数k，使得

$$E\left\|\hat{\beta}(k)-\beta\right\|^2\le E\left\|\hat{\beta}_L-\beta\right\|^2 \tag{53}$$

因为岭估计$\hat{\beta}(k)=(X^TX+kE_m)^{-1}X^TY$是$k$的函数，所以二维坐标平面上若以横轴为$k$，纵轴为$\hat{\beta}(k)$，它将画出$m$条曲线。这些曲线称之为岭迹。

下面我们介绍几种岭参数选择的方法。

1．岭迹稳定

观察岭迹曲线，原则上应该选取使$\hat{\beta}(k)$稳定的最小k值，同时残差平方和也不增加太多。

2．均方误差小

岭估计的均方误差 $mse(\hat{\beta}(k)) = E\left\|\hat{\beta}(k) - \beta\right\|^2$ 是 k 的函数，可以证明它能在某处取得最小值。计算并观察 $mse(\hat{\beta}(k))$，开始它将下降，到达最小值后开始上升。取它最小处的 k 作为岭参数。

3．Hoerl—Kennard公式

设 P 为正交方阵，使式（52）成立，记 $\alpha = P^T\beta$，α 称为典则参数，$Z = XP$，则原模型（48）式变为

$$Y = Z\alpha + \varepsilon \tag{54}$$

这个形式被称为线性回归的典则形式。此时 α 的最小二乘估计与岭回归估计为

$$\hat{\alpha} = (Z^TZ)^{-1}Z^TY = \Lambda^{-1}Z^TY \tag{55}$$

$$\hat{\alpha}(k) = (Z^TZ + kE_m)^{-1}Z^TY = (\Lambda + kE_m)^{-1}Z^TY \tag{56}$$

于是，

$$\hat{\alpha} = \Lambda^{-1}Z^TY = (\hat{\alpha}_1, \cdots, \hat{\alpha}_m)^T,$$

$$\hat{\sigma}^2 = \frac{1}{n-m}(\hat{Y} - Y)^T(\hat{Y} - Y) = \frac{1}{n-m}Y^T(E_n - Z\Lambda^{-1}Z^T)Y$$

都是可计算的，从而选取岭参数 $k = \hat{\sigma}^2 / \max\hat{\alpha}_i^2$。

4．$k = m\hat{\sigma}^2 / \sum_{j=1}^{m}\lambda_j\hat{\alpha}_j^2$

这是Bayes原理推出的法则。

5．$k = m\hat{\sigma}^2 / \sum_{j=1}^{m}\hat{\alpha}_j^2$

直观考虑是，当 $X^TX = E_m$ 时，取 $k = m\hat{\sigma}^2 / \sum_{j=1}^{m}\hat{\alpha}_j^2$ 可使岭估计具有最小的均方误差。

例6　对外贸的进口总额 y 进行研究，并考虑有关的3个因素：国内总产值 x_1，存贮量 x_2，总消费量 x_3，收集了11组数据，见表9。试建立 y 与 x_1, x_2, x_3 的回归方程。

表9　外贸数据

序号	国内总产值 x_1	存贮量 x_2	总消费量 x_3	进口总额 y
1	149.3	4.2	108.1	15.9
2	161.2	4.1	114.8	16.4
3	171.5	3.1	123.2	19.0
4	175.5	3.1	126.9	19.1
5	180.8	1.1	132.1	18.8
6	190.7	2.2	137.7	20.4
7	202.1	2.1	146.0	22.7
8	212.4	5.6	154.1	26.5
9	226.1	5.0	162.3	28.1
10	231.9	5.1	164.3	27.6
11	239	0.7	167.6	26.3

解　将原始数据标准化，计算得到

$$X^{*T}X^{*}=\begin{bmatrix}10 & 0.2585 & 9.9726\\ 0.2585 & 10 & 0.3567\\ 9.9726 & 0.3567 & 10\end{bmatrix}$$

再计算出它的三个特征值，分别为 $\lambda_1=19.9915$，$\lambda_2=9.9815$，$\lambda_3=0.0269$。于是 $X^{*T}X^{*}$ 的条件数 $\lambda_1/\lambda_3=742.9346$，可见设计矩阵存在中等程度的复共线性。视 $\lambda_3\approx 0$，对应的特征向量为

$$\varphi_3=[0.707 \quad 0.007 \quad -0.7072]^T$$

由于 $\varphi_3^T X^{*T}X^{*}\varphi_3=\lambda_3\varphi_3^T\varphi_3=\lambda_3\|\varphi_3\|^2=\lambda_3$，即 $\|X^{*}\varphi_3\|^2=\lambda_3\approx 0$，所以三个标准化变量之间存在复共线性关系

$$0.707x_1^{*}+0.007x_2^{*}-0.7072x_3^{*}=0$$

注意到，自变量 x_2^{*} 的系数绝对值相对非常小，可视为零，而 x_1^{*} 和 x_3^{*} 的系数又近似相等，因此自变量之间的复共线性关系可近似地写为 $x_1^{*}=x_3^{*}$。

下面进行岭回归计算。表10列出了不同 k 值的岭回归系数，表10的最后一列是岭估计对应的残差平方和。我们看到，随着 k 的增加，岭估计的残差平方和也随着增加，所以残差平方和是岭参数 k 的单调增函数。这是很自然的，因为最小二乘估计是使残差平方和达到最小的估计。对应的岭迹图由图5给出。从岭迹图可以看出选 $k=0.04$ 较好。对应的岭回归方程为

$$y=-9.4667+0.0227x_1+0.598x_2+0.1787x_3$$

表10　外贸数据的岭回归分析

k	$\beta_0(k)$	$\beta_1(k)$	$\beta_2(k)$	$\beta_3(k)$	$Q(k)$
0	-10.128	-0.0514	0.5869	0.2868	1.6729
0.01	-9.8414	-0.0178	0.5924	0.2379	1.7278
0.02	-9.6702	0.00148	0.5953	0.2097	1.809
0.03	-9.5534	0.014	0.597	0.1915	1.881
0.04	-9.4667	0.0227	0.598	0.1787	1.9409
0.05	-9.3984	0.0292	0.5986	0.1692	1.9901
0.06	-9.3421	0.0341	0.5989	0.1619	2.0311
0.07	-9.2941	0.038	0.5991	0.156	2.0655
0.08	-9.2521	0.0412	0.5991	0.1513	2.0949
0.09	-9.2146	0.0439	0.599	0.1474	2.1203
0.1	-9.1805	0.0461	0.5989	0.144	2.1424

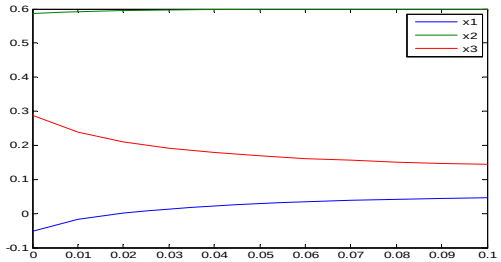

图5　外贸数据回归的岭迹

计算的Matlab程序如下：

```
clc,clear
load txt1.txt
x=txt1(:,1:3);y=txt1(:,4);
k=0:0.01:0.1;
b1=ridge(y,x,k,0), yhat=repmat(b1(1,:),[11,1])+x*b1(2:4,:);
Q=(dist(y',yhat)).^2  %计算残差平方和
plot(k,b1(2:4,:)')
legend('x1','x2','x3')
```

或者我们直接使用 $k=\hat{\sigma}^2/\max\hat{\alpha}_i^2$，确定岭参数 $k=0.0075$，最后得到岭回归方程为

$$y=-9.8976-0.0243x_1+0.5914x_2+0.2473x_3$$

计算的Matlab程序如下：

```
clc,clear
load txt1.txt
txt2=zscore(txt1);     %原始数据进行标准化
m=3;x=txt2(:,1:m);y=txt2(:,end);
[P,A]=eig(x'*x);
lamda=diag(A);Z=x*P;
[a,aci,r,rci,st]=regress(y,Z);     %求典则形式的最小二乘估计和样本方差
k=st(4)/max(a.^2)        %求岭参数的值
b1=ridge(txt1(:,end),txt1(:,1:end-1),k,0)       %使用原始数据进行岭回归分析
```

7.3　主成分估计

主成分回归的理论见第二十六章，下面我们给出例6的主成分回归。

例7　用主成分回归求解例6。

解　首先把设计矩阵 X 标准化为 X^*，对应的标准化变量记作 x_1^*, x_2^*, x_3^*。

$(X^{*T}X^*)/(n-1)$（$n=11$）的三个特征值分别为

$$\lambda_1=1.9992,\ \lambda_2=0.9982,\ \lambda_3=0.003$$

它们对应的三个标准正交化特征向量分别为

$$\varphi_1=[0.7076\quad 0.0435\quad 0.7065],\ \varphi_2=[-0.0357\quad 0.999\quad -0.0258]$$
$$\varphi_3=[-0.707\quad -0.007\quad 0.7072]$$

三个主成分分别为

$$z_1=0.7076x_1^*+0.0435x_2^*+0.7065x_3^*$$
$$z_2=-0.0357x_1^*+0.999x_2^*-0.0258x_3^*$$
$$z_1=-0.707x_1^*-0.007x_2^*+0.7072x_3^*$$

因为 $\lambda_3\approx 0$，且前两个主成分的贡献率

$$\sum_{i=1}^{2}\lambda_i/\sum_{i=1}^{3}\lambda_i=0.9991=99.91\%$$

因此，我们剔除第三个主成分，只保留前两个主成分，得到关于主成分的回归方程

$$\hat{y}^*=0.69z_1+0.1913z_2$$

化成关于标准化变量的回归方程

$$\hat{y}^* = 0.4805x_1^* + 0.2211x_2^* + 0.4826x_3^*$$

最后得到关于原始变量的回归方程为

$$\hat{y} = -9.1301 + 0.728x_1 + 0.6092x_2 + 0.1063x_3$$

计算的Matlab程序如下：

```
clc,clear
load txt1.txt
x0=txt1(:,1:3);y0=txt1(:,4);mu=mean(x0);sig=std(x0);
muy=mean(y0);sigy=std(y0);
x=zscore(x0); y=zscore(y0)
[c,s,t]=princomp(x)
m=2;          %取前两个主成分
a=s(:,1:m)\y          %主成分的回归方程系数
ab=c(:,1:m)*a     %标准化变量的回归方程的系数
b=[muy-sigy*(mu./sig)*ab,sigy*ab'./sig]   %原始变量的回归方程的系数
```

直接用最小二乘法拟合得到的回归方程

$$\hat{y} = -10.128 - 0.0514x_1 + 0.5869x_2 + 0.2868x_3$$

的拟合效果也很好，那么主成分回归好在哪里呢？最主要的好处是，主成分回归舍掉了一个约等于0的特征数，也就是去掉了一个复共线性关系

$$0.707x_1^* + 0.007x_2^* - 0.7072x_3^* = 0$$

这就使得主成分回归模型在预测未来的数据时，将表现较好的稳定性。

§8 非线性回归

本节介绍怎样用Matlab统计工具箱实现非线性回归。

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6　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11，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11

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,可以任意取，这里是给定的
[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。

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12，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12

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，得高程如表13，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13

y \ x	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 的一组数据如表14，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14

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	

4. 做了10次试验得观测数据如表15。

表15

y	16.3	16.8	19.2	18.0	19.5	20.9	21.1	20.9	20.3	22.0
x_1	1.1	1.4	1.7	1.7	1.8	1.8	1.9	2.0	2.3	2.4
x_2	1.1	1.5	1.8	1.7	1.9	1.8	1.8	2.1	2.4	2.5

（1）若以 x_1, x_2 为回归自变量，问它们之间是否存在复共线性关系？

（2）试用岭迹法求 y 关于 x_1, x_2 的岭回归方程，并划出岭迹图。

5. 对某种商品的销量 y 进行调查，并考虑有关的四个因素：x_1－居民可支配收入，x_2－该商品的平均价格指数，x_3－该商品的社会保有量，x_4－其它消费品平均价格指数。表16是调查数据。利用主成分方法建立 y 与 x_1, x_2, x_3, x_4 的回归方程。

表16

序号	x_1	x_2	x_3	x_4	y
1	82.9	92	17.1	94	8.4
2	88.0	93	21.3	96	9.6
3	99.9	96	25.1	97	10.4
4	105.3	94	29.0	97	11.4
5	117.7	100	34.0	100	12.2
6	131.0	101	40.0	101	14.2
7	148.2	105	44.0	104	15.8
8	161.8	112	49.0	109	17.9
9	174.2	112	51.0	111	19.6
10	184.7	112	53.0	111	20.8

6．表17给出10名中学生体重、胸围、胸围之呼吸差及肺活量数据，试用向前选择变量法、向后删除变量法和逐步回归法建模回归模型，其中 y：肺活量（ml），x_1：体重（kg），x_2：胸围（cm），x_3：胸围之呼吸差（cm）。

表17

y	x_1	x_2	x_3
1600	35	69	0.7
2600	40	74	2.5
2100	40	64	2.0
2650	42	74	3.0
2400	37	72	1.1
2200	45	68	1.5
2750	43	78	4.3
1600	37	66	2.0
2750	44	70	3.2
2500	42	65	3.0