第四章 回归分析

4.1 一元线性回归分析

回归分析　回归分析是用来寻找若干变量之间**统计联系**（关系）的一种方法。它是一种统计模型，分为**线性回归**和**非线性回归**，线性回归在气象中最为常用（解释性好，物理机理较为清晰）。
利用回归分析得到的统计关系对某一变量作出未来时刻的估计，称为**预报值(量)**。**前期**（也可以是同期因子）已发生的多个与之有关的气象要素称为**预报因子**。

案例分析

为了预报某地某月平均气温或降水量情况（预报量），选择预报前期已发生的多个有关的气象要素（预报因子），利用回归分析方法分析多个预报因子和预报变量之间的相互关系，建立统计关系的方程式，最后利用其对未来时刻的气温或降水量作出预报估计。

一元回归　一元回归分析处理的是**两个变量**之间的关系，即一个预报量和一个预报因子之间的关系。

4.1.1 回归模型

基本原理　对抽取容量为 n 的预报量 y 与预报因子 x 的一组样本（必须保证样本个数一致），**若认为 y 与 x 是一元线性统计关系**，则线性回归方程为：$y_i = b_0 + bx_i + \varepsilon_i$，$i = 1,2,\cdots,n$（$\varepsilon_i$ 为残差项，我们希望它越小越好），那么预报量的估计量 $\hat{y}$ 与 x 有如下关系：

$$\hat{y}_i = b_0 + bx_i \quad i = 1,2,\cdots,n$$

或写为一般的回归方程：$\hat{y} = b_0 + bx$，其中b_0为截距，b为斜率。

最小二乘法　对所有的 x_i，**若 $\hat{y}_i$ 与 y_i 的偏差最小**，就认为所确定的直线能最好地代表所有实测点的散布规律。
为了消除偏差符号的影响，可以用**偏差的平方**来反映偏差的绝对值偏离情况。
全部观测值与回归直线的**离差平方和**记为：

$$Q(b_0,b) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

它刻画了全部观测值与回归直线的偏离程度。显然Q值越小越好，Q是待定系数b_0和b的函数。

标准方程组　根据**极值原理**，要求：$\frac{\partial Q}{\partial b_0} = 0$，$\frac{\partial Q}{\partial b} = 0$。整理得到求回归系数$b_0$、$b$的方程组：

$$\begin{cases} nb_0 + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ b_0\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \end{cases}$$

称为求回归系数的**标准方程组**。

具体求解

① $\frac{\partial Q}{\partial b_0} = \frac{\partial}{\partial b_0}\left(\sum_{i=1}^{n}(y_i - b_0 - bx_i)^2\right) = \sum_{i=1}^{n} -2(y_i - b_0 - bx_i) = 0 \Rightarrow \sum_{i=1}^{n}(-y_i - b_0 - bx_i) = 0 \Rightarrow$

$nb_0 + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$　其中　$\frac{\partial(y_i - b_0 + bx_i)^2}{\partial b_0} = 2(y_i - b_0 + bx_i)\cdot(-1)$。

② $\frac{\partial Q}{\partial b} = \frac{\partial}{\partial b}\left(\sum_{i=1}^{n}(y_i - b_0 - bx_i)^2\right) = -2\sum_{i=1}^{n} x_i(y_i - b_0 - bx_i) = 0 \Rightarrow b_0\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$

回归系数 $\boxed{b_0=\bar{y}-b\bar{x}}$ $\boxed{b=\frac{\sum_{i=1}^n x_iy_i-n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2-n\bar{x}^2}=\frac{S_{xy}}{S_x^2}}$

距平形式 将 $b_0=\bar{y}-b\bar{x}$ 代入回归方程 $\hat{y}_i=b_0+bx_i$，得到 $\hat{y}_i-\bar{y}=b(x_i-\bar{x})$ 或 $\hat{y}_{di}=bx_{di}$

标准化形式 发现有关系：$b=\frac{S_{xy}}{S_x^2}=\frac{S_{xy}}{S_xS_y}\cdot\frac{S_y}{S_x}=r_{xy}\cdot\frac{S_y}{S_x}$，由此 $\hat{y}_{zi}=r_{xy}x_{zi}$ （这里的x,y都是标准化后的变量）

相关系数 回归系数b与相关系数之间的关系：$b=\frac{S_{xy}}{S_x^2}=r_{xy}\cdot\frac{S_y}{S_x}$ 相关系数r与回归系数b同号

当$b<0$，回归直线斜率为负，预报量 y 随预报因子 x 增加而减少，反映预报量与因子是负相关。
当$b>0$，回归直线斜率为正，预报量 y 随预报因子 x 增加而增加，反映预报量与因子是正相关。

4.1.2 回归问题的方差分析

意义 评价回归方程的优劣

预报量方差 **预报量方差**可以表示成**回归估计值的方差**（回归方差）和**误差**（残差）方差之和：

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

即 $s_y^2=s_{\hat{y}}^2+s_e^2$。

评估分析 方差分析表明，预报量 y 的变化可以看成由前期因子 x 的变化所引起的，同时加上随机因素 e 变化的影响，这种前期因子 x 的变化影响可以用回归方差的大小来衡量。如果**回归方差大/残差方差小**，表明用线性关系解释y与x的关系比较符合实际情况，**回归模型比较好**。

离差平方和 $\frac{1}{n}\sum_{i=1}^n(y_i-\bar{y})^2=\frac{1}{n}\sum_{i=1}^n(\hat{y}_i-\bar{y})^2+\frac{1}{n}\sum_{i=1}^n(y_i-\hat{y}_i)^2$ 两边同时乘以n变成各变量离差平方和的关系。

总离差平方和：$s_{yy}=U+Q=\sum_{i=1}^n(y_i-\bar{y})^2$ 反映**因变量 y 的 n 个观测值与其均值的总离差**
回归平方和：$U=\sum_{i=1}^n(\hat{y}_i-\bar{y})^2$ 反映**回归值的分散程度**
残差平方和：$Q=\sum_{i=1}^n(y_i-\hat{y})^2$ 反映**观测值偏离回归直线的程度**

4.1.3 相关系数与线性回归

判决系数 因为回归方差不可能大于预报量的方差，可以用它们的比值来衡量方程的拟合效果。即：

$$\frac{s_{\hat{y}}^2}{s_y^2}=\frac{\frac{1}{n}\sum_{i=1}^n(\hat{y}_i-\bar{y})^2}{\frac{1}{n}\sum_{i=1}^n(y_i-\bar{y})^2}=\frac{U}{s_{yy}}=r_{xy}^2$$

上式表明预报因子x对预报量y的方差的线性关系程度，这一比值又称为**回归方程判决系数/解释方差**。
判决系数是衡量两个变量线性关系密切程度的量，等于两变量相关系数的平方。
如果是多元线性回归，合理猜想，是复相关系数的平方。

物理含义
① **回归平方和占总离差平方和的比例** ② 反映回归直线的**拟合程度**
③ 取值范围在[0,1] ④ 判决系数**等于相关系数的平方**
⑤ $r^2\to1$ 说明回归方程拟合的越好，$r^2\to0$ 说明回归方程拟合的越差

4.1.4 回归方程的显著性检验

中心思想 显著性检验的主要思想是检验预报因子与预报量**是否有线性关系**。

统计量 可以证明在**原假设总体回归系数为 0** 的条件下，统计量：

$$F=\frac{U/1}{Q/(n-2)}$$

遵从**分子自由度为1，分母自由度为$(n-2)$**的F分布。

显著性检验 查 F 的分布表，在 $\alpha=0.05$ 下，若 $F>F_\alpha$ 则认为回归方程是显著的。反之，则不显著。

相关系数 统计量F也可以写为：$F=\frac{U/1}{Q/(n-2)}=\frac{s_{\hat{y}}^2/1}{s_e^2/(n-2)}=\frac{r^2}{(1-r^2)/(n-2)}$，与 $t=\frac{r}{\sqrt{1-r^2}}\sqrt{n-2}$ 比较，发现二者等价。

注意 F的相关系数表达式开方就是相关系数t检验的表达式，故**一元回归方程的检验与相关系数的检验一致**。

4.1.5 回归系数的显著性检验

说明 气象中使用最多的是回归方程的**距平形式**，所以对回归方程的显著性检验可以**只对因子的回归系数**进行检验。

统计量 在**原假设H_0: 回归系数 $\beta=0$**的条件下

① 统计量 $t=\dfrac{b-\beta}{\sqrt{\dfrac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}}=\dfrac{b/\sqrt{c}}{\sqrt{Q/(n-2)}}$ 遵从自由度为 $n-2$ 的 t 分布。

其中：$\hat{\sigma}^2=\dfrac{1}{n-2}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=\dfrac{Q}{n-2}$　　$c=\left[\sum_{i=1}^{n}(x_i-\bar{x})^2\right]^{-1}$

② 或者根据F分布与t分布的关系，统计量 $F=\dfrac{U/1}{Q/(n-2)}=\dfrac{b^2/c}{Q/(n-2)}$ 遵从分子自由度为1，分母自由度为$n-2$的 F 分布。其中 $U=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2=\sum_{i=1}^{n}(b_0-bx_i-b_0-b\bar{x})^2=b^2\sum_{i=1}^{n}(x_i-\bar{x})^2=\dfrac{b^2}{c}$

4.1.6 预报值的置信区间

置信区间 因为 $e_i=y_i-E(y_i)$ 可以看成遵从 **$N(0,\sigma^2)$ 的正态分布**，所以其 **95% 的置信区间为 $E(y_i)\pm1.96\sigma$**

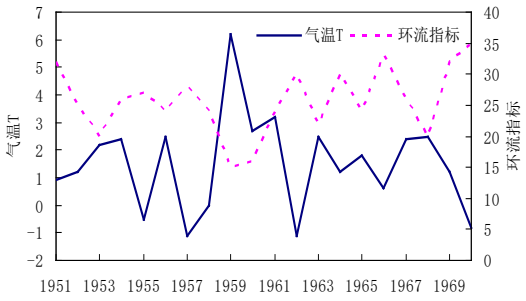
$E(y_i)$ 可用 $b_0+bx_i=\hat{y}_i$ 估计，σ 可用无偏估计量 $\hat{\sigma}=\sqrt{\dfrac{Q}{n-2}}$ 估计，$\hat{\sigma}^2=\dfrac{1}{n-2}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$

预报值的 95% 置信区间可近似估计为 $[\hat{y}_i-1.96\hat{\sigma},\hat{y}_i+1.96\hat{\sigma}]$。

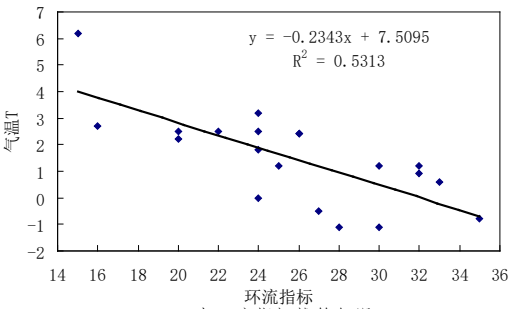
每一个点的置信区间都不一样，置信区间上下界是一个曲线。

4.1.7 一元线性回归分析预测步骤

分析数据



直观看到负相关关系



由环流指标推算气温

第一步 **计算回归系数，确定方程**。对上述资料，容易算得 $n=20$，$\sum_{i=1}^{\infty}x_i=513$，$\sum_{i=1}^{\infty}y_i=30.0$，$\sum_{i=1}^{1}x_i^2=13721$，$\sum_{i=1}^{2}x_iy_i=637$　根据 $b_0=\bar{y}-b\bar{x}$，$b=\dfrac{\sum_{i=1}^{n}x_iy_i-n\bar{x}\bar{y}}{\sum_{i=1}^{n}x_i^2-n\bar{x}^2}=\dfrac{S_{xy}}{S_x^2}$　可以解得：

$b_0=7.5$，$b=-0.23$　　最终得到回归方程：$\hat{y}=7.5-0.23x$

第二步 **回归方程显著性检验**。

再次计算得到：$\sum_{i=1}^{1}y_i^2=103.12$ 于是 $r_{xy}=\sqrt{\dfrac{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2}}=\sqrt{\dfrac{13721-\frac{1}{20}\times(513)^2}{103.12-\frac{1}{20}\times(30.0)^2}}\times(-0.23)=-0.727$

最终得到：$F=\dfrac{(-0.727)^2}{[1-(-0.727)^2]/(20-2)}=20.18$ 查询F分布表，在$\alpha=0.05$，分子自由度为 1，分母自由度为 18 时，$F_\alpha=4.41$ 由于 $F>F_\alpha$ 认为回归方程是显著的。（考试可以灵活应用t检验）

第三步 **计算预报值的置信区间，作出预测**。

将$x=24$代入回归方程，计算出预报值为 $y_{24}=1.98$℃，又有 $Q=s_{yy}-U=s_{yy}-s_{yy}r^2=s_{yy}(1-r^2)$

算出：$\hat{\sigma}=\sqrt{\frac{1}{20-2}\times58.12(1-0.727^2)}=1.11$ ，用 $E(y_i)\pm1.96\sigma$ 得到置信区间。

所以 1971 年北京 3 月下旬气温的 95%置信区间为−0.2~4℃。

4.2 多元线性回归分析

4.2.1 多元回归模型

模型定义 描述因变量 y 如何依赖于多个自变量 $x_1, x_2, \dots, x_p$ 和误差项 ε 的方程，称为**多元回归模型**。

预报因子的选择在实际研究中十分困难，而且多个物理因子之间的协同是非线性的。

涉及 p 个自变量的**多元线性回归模型**可表示为：

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$$

其中 $\beta_0, \beta_1, \beta_2, \dots, \beta_p$ 是**参数**，ε 是被称为**误差项**的随机变量，包含在 y 里面但不能被 p 个自变量的线性关系所解释的变异性。

结构表达式 假定预报量 y 与 p 个预报因子关系是线性（先画图观察），为研究它们之间的联系作 $\boldsymbol{n}$ **次抽样**，则可得到如下结构表达式：$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \cdots + \beta_p x_{pt} + e_t,\ t = 1,2,\cdots,n$ 展开：

$$\begin{cases} y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{21} + \cdots + \beta_p x_{p1} + e_1 \\ y_2 = \beta_0 + \beta_1 x_{12} + \beta_2 x_{22} + \cdots + \beta_p x_{p2} + e_2 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ y_n = \beta_0 + \beta_1 x_{1n} + \beta_2 x_{2n} + \cdots + \beta_p x_{pn} + e_n \end{cases} \quad ①$$

也可以写成**矩阵形式**：$\boldsymbol{Y = X\beta + e}$ ②

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \quad e = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} \quad X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{p1} \\ 1 & x_{12} & \cdots & x_{p2} \\ \cdots & \cdots & \cdots & \cdots \\ 1 & x_{1n} & \cdots & x_{pn} \end{bmatrix}$$

回归方程 我们得到的是一组实测 p 个变量的样本，利用这组样本（n次抽样）对上述回归模型进行估计，得到的估计方程（没有ε）为多元线性回归方程，记为：

$$\boldsymbol{\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p}$$

其中b_i 是 β_i 的估计值，下面讨论如何确定它们。

4.2.2 回归系数最小二乘估计

回归方程 和一元线性回归类似，在样本容量为 n 的 y 预报量和因子变量 x 的实测值中，满足线性回归方程：

$$\hat{y}_t = b_0 + b_1 x_{1t} + b_2 x_{2t} + \cdots + b_p x_{pt} \qquad t = 1,2,\cdots,n$$

计算方法 要求回归系数，应使**全部的预报量观测值**与**回归估计值**的**残差平方和**$\boldsymbol{Q}$达到最小。

$$Q = \sum_{t=1}^{n} (y_t - \hat{y}_t)^2 = \sum_{t=1}^{n} (y_t - b_0 - b_1 x_{1t} - b_2 x_{2t} - \cdots - b_p x_{pt})^2$$

由**极值定理**：

$$\begin{cases} \frac{\partial Q}{\partial b_0} = 0 \\ \frac{\partial Q}{\partial b_1} = 0 \\ \vdots \\ \frac{\partial Q}{\partial b_p} = 0 \end{cases} \quad \begin{cases} nb_0 + b_1 \sum_t x_{1t} + b_2 \sum_t x_{2t} + \cdots + b_p \sum_t x_{pt} = \sum_t y_t \\ b_0 \sum_t x_{1t} + b_1 \sum_t x_{1t}^2 + b_2 \sum_t x_{2t} x_{1t} + \cdots + b_p \sum_t x_{pt} x_{1t} = \sum_t y_t x_{1t} \\ b_0 \sum_t x_{2t} + b_1 \sum_t x_{1t} x_{2t} + b_2 \sum_t x_{2t}^2 + \cdots + b_p \sum_t x_{pt} x_{2t} = \sum_t y_t x_{2t} \\ \cdots\cdots \\ b_0 \sum_t x_{pt} + b_1 \sum_t x_{1t} x_{pt} + b_2 \sum_t x_{2t} x_{pt} + \cdots + b_p \sum_t x_{pt}^2 = \sum_t y_t x_{pt} \end{cases}$$

4.2.3 线性回归模型的其他形式

4.2.3.1 距平形式的多元回归方程

距平形式 我们先看第一行的方程：$\frac{\partial Q}{\partial b_0} = 0 \Rightarrow b_0 = \bar{y} - b_1 \bar{x}_1 - b_2 \bar{x}_2 - \cdots - b_p \bar{x}_p$ 得到b_0，将b_0代入待求方程：

$\hat{y}_t = b_0 + b_1 x_{1t} + b_2 x_{2t} + \cdots + b_p x_{pt} \Rightarrow \hat{y}_t - \bar{y} = b_1(x_{1t} - \bar{x}_1) + b_2(x_{2t} - \bar{x}_2) + \cdots + b_p(x_{pt} - \bar{x}_p)$

$\Rightarrow \boldsymbol{\hat{y}_d = b_1 x_{d1} + b_2 x_{d2} + \cdots + b_p x_{dp}}$ 即为**距平形式的回归方程**（d表示距平的下标）

根据距平形式的方程，由于各个回归系数相同，且能够计算出b_0，故可以得到原始方程。

求解 残差平方和为 $Q = \sum_{t=1}^{n} (y_{dt} - \hat{y}_{dt})^2 = \sum_{t=1}^{n} (y_{dt} - b_1 x_{d1t} - b_2 x_{d2t} - \cdots - b_p x_{dpt})^2$

从距平变量的观测值求回归系数，同样用最小二乘法导出求回归系数的标准方程组：

$$\begin{cases} b_1\sum_t x_{d1t}^2+b_2\sum_t x_{d2t}x_{d1t}+\cdots+b_p\sum_t x_{dpt}x_{d1t}=\sum_t y_{dt}x_{d1t} \\ b_1\sum_t x_{d1t}x_{d2t}+b_2\sum_t x_{d2t}^2+\cdots+b_p\sum_t x_{dpt}x_{d2t}=\sum_t y_{dt}x_{d2t} \\ \cdots\cdots \\ b_1\sum_t x_{d1t}x_{dpt}+b_2\sum_t x_{d2t}x_{dpt}+\cdots+b_p\sum_t x_{dpt}^2=\sum_t y_{dt}x_{dpt} \end{cases}$$ 发现中间各项是协方差形式

为了得到协方差矩阵形式，上式两边乘上$1/n$，变成各变量的**协方差形式**，相应的方程组写为：

$$\begin{cases} b_1s_{11}+b_2s_{12}+\cdots+b_ps_{1p}=s_{1y} \\ b_1s_{21}+b_2s_{22}+\cdots+b_ps_{2p}=s_{2y} \\ \cdots\cdots \\ b_1s_{p1}+b_2s_{p2}+\cdots+b_ps_{pp}=s_{py} \end{cases} \quad s_{kl}=\frac{1}{n}\sum_{i=1}^n x_{dik}x_{dil} \quad s_{ky}=\frac{1}{n}\sum_{i=1}^n x_{dik}y_{di} \quad k,l=1,2,\cdots,p$$

4.2.3.2 标准化形式的多元回归方程

标准化形式 对**距平变量**多元线性回归方程两边除以**预报量 y 的标准差 s_y**，得到：

距平形式的回归方程为 $\hat{y}_d=b_1x_{d1}+b_2x_{d2}+\cdots+b_px_{dp}$ 将其除以s_y得到：

$$\frac{\hat{y}_d}{s_y}=\frac{b_1x_{d1}}{s_y}+\frac{b_2x_{d2}}{s_y}+\cdots+\frac{b_px_{dp}}{s_y} \Rightarrow \frac{\hat{y}_d}{s_y}=b_1\frac{s_1}{s_y}\frac{x_{d1}}{s_1}+b_2\frac{s_2}{s_y}\frac{x_{d2}}{s_2}+\cdots+b_p\frac{s_p}{s_y}\frac{x_{dp}}{s_p}$$ 系数改变

令标准化回归系数为：$\boldsymbol{b_{zk}=b_k\frac{s_k}{s_y}}$ $(k=1,2,\cdots p)$

$\Rightarrow \boldsymbol{\hat{y}_z=b_{z1}x_{z1}+\cdots\cdots+b_{zp}x_{zp}}$ **标准化形式的回归方程**（z表示标准化的下标）

根据标准化形式的方程，由于能够计算出s_k,s_y，故可以得到距平方程，因此三个方程互通。

求解 残差平方和为 $Q=\sum_{t=1}^n(y_{zt}-\hat{y}_{zt})^2=\sum_{t=1}^n(y_{zt}-b_1x_{z1t}-b_2x_{z2t}-\cdots-b_px_{zpt})^2$

从标准化变量的观测值求回归系数，同样用最小二乘法导出求回归系数的标准方程组：

$$\begin{cases} b_{z1}\sum_t x_{z1t}^2+b_{z2}\sum_t x_{z2t}x_{z1t}+\cdots+b_{zp}\sum_t x_{zpt}x_{z1t}=\sum_t y_{zt}x_{z1t} \\ b_{z1}\sum_t x_{z1t}x_{z2t}+b_{z2}\sum_t x_{z2t}^2+\cdots+b_{zp}\sum_t x_{zpt}x_{z2t}=\sum_t y_{zt}x_{z2t} \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ b_{z1}\sum_t x_{z1t}x_{zpt}+b_{z2}\sum_t x_{z2t}x_{zpt}+\cdots+b_{zp}\sum_t x_{zpt}^2=\sum_t y_{zt}x_{zpt} \end{cases}$$ 发现中间各项是相关系数形式

上式两边乘上$1/n$，变成各变量的**相关系数形式**，相应的方程组写为：

$$\begin{cases} r_{11}b_{z1}+r_{12}b_{z2}+\cdots+r_{1p}b_{zp}=r_{1y} \\ r_{21}b_{z1}+r_{22}b_{z2}+\cdots+r_{2p}b_{zp}=r_{2y} \\ \cdots\cdots \\ r_{p1}b_{z1}+r_{p2}b_{z2}+\cdots+r_{pp}b_{zp}=r_{py} \end{cases}$$

4.2.4 回归问题的方差分析

回归方差 **回归方差**可表示为：$\boldsymbol{s_{\hat{y}}^2=\frac{1}{n}U=\sum_{k=1}^p b_ks_{ky}}$ 回归系数与ky的协方差

对于**标准化变量**而言，**回归方差**为：$\boldsymbol{s_{\hat{y}_z}^2=\sum_{k=1}^p b_{zk}r_{ky}}$ 关系好不代表关系显著

如果**回归方差大**，表明用线性关系解释y与x的关系比较符合实际情况，**回归模型比较好**。

推导

回归平方和为：$U=\sum_{i=1}^n(\hat{y}_i-\bar{y})^2$ 将其展开：

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

发现 $\sum_{i=1}^n(\hat{y}_i-\bar{y})(y_i-\hat{y}_i)=0$，因此 $U=n\sum_{k=1}^p b_ks_{ky}$。

4.2.5 复相关系数

复相关系数 衡量**一个预报量**与**多个变量**之间线性关系程度的量，即衡量预报量y与估计量$\hat{y}$之间线性相关程度的量：

$$R=\frac{\sum_{i=1}^n(y_i-\bar{y})(\hat{y}_i-\bar{y})}{\sqrt{\sum_{i=1}^n(y_i-\bar{y})^2\sum_{i=1}^n(\hat{y}_i-\bar{y})^2}}=\sqrt{\frac{U}{S_{yy}}}, \quad R^2=1-\frac{Q}{S_{yy}}$$

称为多元回归方程的**可解释系数**。

4.2.6 回归方程的显著性检验

总体检验 回归方程的显著性检验和一元回归类似：**假设总体回归系数为 0 时 $H_0:\beta_1=\beta_2=\cdots=\beta_p=0$**

$$F=\frac{U/p}{Q/(n-p-1)}=\frac{\frac{U}{S_{yy}}/p}{\frac{Q}{S_{yy}}/(n-p-1)}=\frac{\frac{R^2}{p}}{\frac{1-R^2}{n-p-1}}$$

遵从**分子自由度为p，分母自由度为$n-p-1$**的F分布。

显著性检验 在显著性水平下 $\alpha=0.05$，若 $F>F_\alpha$ 则认为回归方程是显著的。反之，则不显著。

注意 方程显著，不代表每个回归系数都是显著的。

4.2.7 预报值的置信区间

置信区间 因为 $e=y_i-E(y_i)\sim N(0,\sigma^2)$ 的正态分布，所以其95% 的置信区间为 $E(y_i)\pm 1.96\sigma$

$E(y_i)$可用$\hat{y}_i$估计，σ可用**无偏估计量** $\hat{\sigma}=\sqrt{\frac{Q}{n-p-1}}$ 估计，$\hat{\sigma}^2=\frac{1}{n-p-1}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$

预报值的 95% 置信区间可近似估计为 $[\hat{y}_i-1.96\hat{\sigma},\hat{y}_i+1.96\hat{\sigma}]$

4.2.8 气象应用与实例

基本步骤

① 确定预报量并选择恰当的因子。

② 根据数据计算回归系数标准方程组所包含的有关统计量(因子的交叉积、协方差阵或相关阵，以及因子与预报量交叉积、协方差或相关系数)。

③ 解线性方程组求出回归系数。

④ 建立回归方程并进行统计显著性检验。

⑤ 利用已出现的因子值代入回归方程作出预报量的估计，求出预报值的置信区间。

实例分析

设对某一预报量y，选择 4 个因子作预报，样本容量$n=13$。

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

为了说明问题，我们选取x_1,x_2,x_4作为因子，使用标准化变量的回归方程，求标准回归系数的方程组为：

$$\begin{cases}b_1+0.2286b_2-0.2455b_4=0.7307\\0.2286b_1+b_2-0.9730b_4=0.8163\\-0.2455b_1-0.9730b_2+b_4=-0.8213\end{cases}$$

略去下标z

上式系数都是相关系数。得出回归方程为：$\hat{y}=0.5679x_1+0.4323x_2-0.2613x_4$。

计算回归方差：$s_{\hat{y}}{}^2=0.5679\times0.7307+\cdots+0.2613\times0.8213=0.9823$（已知 $s_{\hat{y}_z}{}^2=\sum_{k=1}^{p}r_{ky}b_{zk}$），得到残差方差 $s_e^2=1-s_{\hat{y}}^2$=0.0177。随后对回归方程进行统计检验，计算 ，$F=\frac{U/p}{Q/n-p-1}=\frac{0.9823/3}{0.0177/(13-3-1)}$=166.4。

在显著水平 $\alpha=0.05$ 下，$F>F_\alpha$，说明该方程是显著的。

以上用的是多元线性回归方法。但是这是否说明三个因子对预报量都有显著影响呢?

对回归系数检验，利用 $F_k=\frac{b_k{}^2/c_{kk}}{Q/n-p-1}$ 发现 b_1 是显著的，而 b_2 和 b_4 是不显著的。

通过例子说明，尽管回归方程是显著的，并不能说明方程中所有因子都对预报量有显著影响。因此上述回归方程不是最优的。我们下面通过逐步回归方法来得到最优的回归方程。

4.3 逐步回归方法

小节引入

在气象预报中，对预报量的预报常常需要从可能影响预报 y 的诸多因素中挑选一批关系较好的作为预报因子，应用多元线性回归的方法建立回归方程来做预报。但如何才能保证在已选定的一批因子中得到最优的回归方程呢？逐步回归分析方法就是针对这一问题提出的一种常用方法。

4.3.1 预报因子(回归系数)的显著性检验

方差贡献　若在 p 个预报因子中去掉一个因子 k，再建立它们对 y 的预报方程，则此时回归平方和、残差平方和分别记为 $U^{(p-1)}$，$Q^{(p-1)}$，定义单个预报因子的方差贡献：

$$V_k = U^{(p)} - U^{(p-1)} = Q^{(p-1)} - Q^{(p)} = \frac{b_k^2}{C_{kk}}, \quad k = 1,2,\cdots,p$$

其中 C_{kk} 是因子离差矩阵 $C = (X'X)^{-1}$ 的对角线上的元素。我们利用**方差贡献判断因子的重要性**。

有公式：$s_{\hat{y}}{}^2 = \frac{1}{n}U = \sum_{k=1}^{p} b_k s_{ky} = \sum_{k=1}^{p} \frac{{b_k}^2}{C_{kk}}$　　$C_{kk} = [\sum_{i=1}^{n}(x_{ki} - \bar{x}_k)^2]^{-1}$

假设检验　在多元线性回归方程的建立中，尽管最后都作了方程的统计检验，但并不意味着在 p 个因子中，每个因子对预报量 y 的影响都是重要的。需要对每个因子进行考察，若某个因子对预报量 y 的作用不显著，那么在多元线性回归方程中它前面的系数就可能近似为0。

因此，检验**某一因子是否显著**等价于**检验假设** $H_0: \beta_k = 0$。

统计量的确定

要对 β_k 作假设检验，自然就要寻找它的样本统计量 b_k 和与它有关的统计量的分布。因为最小二乘估计的 b_k 是随机变量 y_i 的线性函数，由于这些随机变量是遵从正态分布，则 b_k 也遵从正态分布。

统计量　$F_k = \frac{V_k}{\frac{Q}{(n-p-1)}} = \boxed{\frac{b_k^2/C_{kk}}{Q/(n-p-1)}}$　符合**自由度为(1，$n-p-1$)的F分布**。给定信度以后，查表求出标准值，若 $F_k \geq F_\alpha$，说明该因子方差贡献显著，保留该因子，否则可以考虑从回归方程中剔除出去。

4.3.2 预报因子数目对回归方程的影响

具体影响　① 一般而言，回归方程中包含的因子个数越多，回归平方和就越大，残差平方和越小。但是当因子增加到一定数目，残差平方和下降的幅度就很小了。一般回归方程的因子数目最多在 **5-6 个**左右为宜。

② 如果因子过多，则一方面对方程所起的贡献已不很大，另一方面会带来因子本身的各种随机因素，影响回归方程的稳定性，反而使预报效果下降。

③ 选择因子时要使**因子之间**的相关系数**越小越好**，而因子**各自与预报量**之间的相关系数**越大越好**。

关键问题　既要选择对预报量影响显著的因子，又要使回归方程的残差方差估计很小，这样才有利于气象预报。

4.3.3 逐步回归的三种方案

总体方案　**逐步剔除方案、逐步引进方案、双重检验的逐步回归方案**

4.3.3.1 逐步剔除方案

基本思想　从包含**全部变量**的回归方程中逐步剔除不显著的因子。

方案　假定有 p 个预报因子，首先用这 p 个因子建立回归方程，然后检查每个因子的方差贡献大小：

$V_k = \frac{b_k^2}{C_{kk}}$　$k = 1,2,\cdots,p$　从 V_k 中选出**方差贡献最小者**记为 V_{min}，使用**统计量** $F_k = \frac{V_{\min}}{Q^{(l)}/n-l-1}$ 检验：

若显著，则其余因子也是显著的。若不显著，则剔除这一因子，对该因子对应的列进行消去后重复上面的步骤。其中 $Q^{(l)}$ 表示回归方程含 l 个因子时的残差平方和。

问题

（1）因子的方差贡献代表什么意义？

（2）为何不同时把几个不显著的因子从方程中剔除出去，而是要每次剔除一个？

问题一 回归平方和是**所有因子**对预报量的**总贡献**。所考虑的因子越多，回归平方和越大，若去掉一个因子，回归平方和只会减小，不会增加。减少的数值越大，说明该因子在回归中所起的作用越大，表明该因子越重要，所以，可用此衡量该因子的方差贡献大小。

$$V_k = U^{(l)} - U^{(l-1)}$$

V_k就是去掉**第k个**因子后，回归平方和的减少量，这部分也叫做**偏回归平方和**，其衡量每个因子对回归方程所起作用的大小。

问题二 在剔除因子过程中，假如x_1、x_2方差贡献都比较小（甚至相等），我们也只能剔除其中的**最小者**，而不应该全部去掉。因为这两个因子之间可能存在密切相关关系，剔除 x_1 后，其对 y 的影响很大部分可以转加到 x_2 对 y 的影响上，所以回归平方和不会因此减小很多。但如果同时去掉两个因子，就会比较多的减少回归平方和，从而影响回归的精度。

4.3.3.2 逐步引进方案

基本思想 在一批待选的因子中，考查他们对预报量 y 的方差贡献，挑选所有因子中方差贡献最大者，经统计检验是显著后，进入回归方程。

方案 如从 $x_1, x_2, \ldots, x_p$ 等因子中考察哪个因子方差在一元回归方程中贡献最大，首先计算：

$$V_k^{(1)} = U^{(1)} - U^{(0)} = U^{(1)} \quad (k = 1,2,\cdots,p)$$

$U^{(0)}$ 为回归方程中无任何因子时的回归平方和，此时为 0。

假如在 p 个因子中，x_k 的方差贡献最大，记为 V_{max}，则据回归系数的检验公式遵从***F*分布的统计量**进行检验：

$$F = \frac{V_{\max}}{\frac{Q^{(1)}}{n-1-1}}$$

若显著，则该因子引进。设到 l 步，方程已有 l 个因子。若考虑从 $p-l$ 个因子中引进哪个变量时，还是要考察他们各个因子引进后的方差贡献，仍选取最大者，记为V_{max}，使用统计量：

$$F = \frac{V_{\max}}{\frac{Q^{(l+1)}}{n-(l+1)-1}}$$

作检验，其中 Q^{l+1} 表示在将要引入回归方程的 $l+1$ 个因子时，回归方程的残差平方和。如此在方程中逐个地引入因子。

注意 这样得到的方程**并不能保证其中所有因子都是显著的**。因为各因子之间可能存在相关关系，引入新变量后，原有的变量就不一定仍然显著。所以，逐步引入方案不一定保证最后的回归方程是最优的。

4.3.3.3 双重检验的逐步回归方案

基本思想 将因子一个个引入，引入因子的条件是该因子的方差贡献显著；同时，每引入一个新因子，要对已引入的老因子逐个检验，将方差贡献变为不显著的因子剔除。

因此双重检验的逐步回归能使最后组成的方程**只含有重要的变量**，所建立的回归方程也称为**最优回归方程**。这一方法在目前气象统计预报中所常用。

方法 利用求解线性方程组求解求逆并行(紧凑)方案，使得计算因子方差贡献和求解回归系数**同时进行**。

优点 计算简便，由于每步都作检验，保证了最后所得方程中所有因子都是显著的。

第一步 **准备工作：**首先从**标准化变量**出发，利用**标准回归方程组**，建立**相关系数增广矩阵**，如下：

$$R^{(0)} = \begin{Bmatrix} r_{11} & r_{12} & \cdots & r_{1p} & r_{1y} \\ r_{21} & r_{22} & \cdots & r_{2p} & r_{2y} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ r_{p1} & r_{p2} & \cdots & r_{pp} & r_{py} \\ r_{y1} & r_{y2} & \cdots & r_{yp} & r_{yy} \end{Bmatrix}$$

第二步 **引进因子：**从 p 个待选的标准化因子 $x_{z1}, x_{z2}, x_{z3}, \cdots, x_{zp}$ 中，考虑引进第一个因子时，建立引进因子的回归方程：

$$\hat{y}_{zk} = b_{zk} x_{zk} \qquad (k = 1,2\cdots p)$$

引进方差贡献最大的那个因子。

$$V_k^{(1)} = U^{(1)} - U^{(0)} = U^{(1)} = b_{zk}^{(1)} r_{ky}^{(0)} = r_{ky}^{(1)} r_{ky}^{(0)} = \frac{\left[r_{ky}^{(0)}\right]^2}{r_{kk}^{(0)}} \quad (k = 1,2,\cdots,p)$$

括号表示第几步，注意为计算方便，式中回归（误差）平方和符号均用回归（误差）方差代替。
假如在p个因子中，x_{zk}的方差贡献最大，记为V_{max}，据回归系数遵从**F分布的统计量**进行检验：

$$F = \frac{V_{\max}}{\frac{Q^{(1)}}{n-1-1}} \qquad Q^{(0)} = s_{yy} = 1 = r_{yy}^{(0)} \qquad Q^{(1)} = Q^{(0)} - V_k^{(1)} = r_{yy}^{(0)} - \frac{\left[r_{ky}^{(0)}\right]^2}{r_{kk}^{(0)}} = r_{yy}^{(1)}$$

若显著，则将第 k 个因子引进方程。这相当于对$R^{(0)}$阵中第k列进行消去，变成$R^{(1)}$。
假定在前 l 步中已引入 l 个因子，考虑 $p-l$ 个未引入的因子中的方差贡献时，计算第 k 个因子方差贡献的公式：

$$V_k^{(l+1)} = \frac{\left[r_{ky}^{(l)}\right]^2}{r_{kk}^{(l)}} \qquad V_k^{(1)} = \frac{\left[r_{ky}^{(0)}\right]^2}{r_{kk}^{(0)}}$$

计算中可利用前 l 步消去求逆的结果，即用在 $R^{(0)}$ 作 l 次消去求逆变成 $R^{(l)}$ 矩阵后阵中的元素。这样可以简化过程的计算量。

其中 $V_{max} = V_k^{(l+1)}$，如果发现第 k 个因子方差贡献最大，则用它进一步作下面的显著性检验，这时利用下面统计量作检验。

$$F = \frac{V_k^{(l+1)}}{\frac{Q^{(l+1)}}{n-(l+1)-1}}$$

$$Q^{(0)} = s_{yy} = 1 = r_{yy}^{(0)}$$
$$Q^{(1)} = = Q^{(0)} - V_k^{(1)} = r_{yy}^{(0)} - \left[r_{ky}^{(0)}\right]^2 / r_{kk}^{(0)} = r_{yy}^{(1)}$$
$$Q^{(l)} = r_y^{(l)}$$
$$Q^{(l+1)} = r_{yy}^{(l)} - V_k^{(l+1)}$$

在显著性水平 α 下，若 $F > F_\alpha$，则认为该因子方差贡献显著，引入该因子。

第三步 **剔除因子：**当因子引入后，原来已引入的因子方差贡献会发生变化，可能变为不显著，因此要进行剔除，剔除的标准是进行统计检验。可以证明，在逐步回归中，**仅在第三个因子引入后才考虑剔除**。设已引进了 l 个因子，考虑其中第 k 个因子的方差贡献，使用如下公式：

$$V_k^{(l)} = \frac{\left[r_{ky}^{(l-1)}\right]^2}{r_{kk}^{(l-1)}} = \frac{\left[r_{ky}^{(l)} r_{kk}^{(l-1)}\right]^2}{r_{kk}^{(l-1)}} = \frac{\left[r_{ky}^{(l)}\right]^2}{r_{kk}^{(l)}}$$

找出其中最小者，进行统计检验：

$$F = \frac{V_k^{(l)}}{r_{yy}^{(l)}/n-l-1}$$

若该因子不显著，则剔除。再对该因子所对应的列进行消去，就当该因子从未进入过方程一样。自此，每一步**首先考虑**有无因子需要剔除，若有就进行剔除，直到没有可剔除的因子时再考虑引入新因子，如此进行下去，直到既无因子剔除又无因子可引入为止。

第四步 **计算结果：** 设最后引入了 l 个因子进入回归方程，$R^{(0)}$ 变到 $R^{(l)}$，则回归方程为：

$$\hat{\boldsymbol{y}}_{\boldsymbol{z}} = \boldsymbol{b}_{\boldsymbol{z1}}\boldsymbol{x}_{\boldsymbol{z1}} + \boldsymbol{b}_{\boldsymbol{z2}}\boldsymbol{x}_{\boldsymbol{z2}} + \cdots + \boldsymbol{b}_{\boldsymbol{zl}}\boldsymbol{x}_{\boldsymbol{zl}}$$

其中：

标准化数据回归系数为：$\boldsymbol{b_{zk}} = \boldsymbol{r_{ky}}^{(\boldsymbol{l})}$

原始数据回归系数为：$\boldsymbol{b_k} = \frac{s_y}{s_k}\boldsymbol{b_{zk}} = \frac{s_y}{s_k}\boldsymbol{r_{ky}^{(l)}}$ $\qquad b_0 = \bar{y} - b_1\bar{x}_1 - b_2\bar{x}_2 - \cdots - b_k\bar{x}_k$

回归方程的残差平方和为：$\boldsymbol{Q} = \boldsymbol{S_{yy}}\boldsymbol{Q}^{(\boldsymbol{l})} = \boldsymbol{S_{yy}}\boldsymbol{r_{yy}}^{(\boldsymbol{l})}$

回归平方和为：$\boldsymbol{U} = \boldsymbol{S_{yy}} - \boldsymbol{Q} = \boldsymbol{S_{yy}}\left(\boldsymbol{1} - \boldsymbol{r_{yy}^{(l)}}\right)$

复相关系数为：$R = \sqrt{\frac{U}{S_{yy}}} = \sqrt{1 - r_{yy}{}^{(l)}}$ 判断方程是否显著，使用t检验。

回归方程的剩余标准差无偏估计量为：$\hat{\sigma} = \sqrt{\frac{S_{yy}Q^{(l)}}{n-l-1}} = \sqrt{\frac{S_{yy}r_{yy}^{(l)}}{n-l-1}}$ 可进行预报值的置信区间估计

引入公式 $V_k^{(l+1)}=\frac{\left[r_{ky}^{(l)}\right]^2}{r_{kk}^{(l)}}$ $F=\frac{V_k^{(l+1)}}{\frac{r_{yy}^{(l)}-V_k^{(l+1)}}{n-(l+1)-1}}$ **引进 3 个后考虑剔除**

剔除公式 $V_k^{(l)}=\frac{\left[r_{ky}^{(l)}\right]^2}{r_{kk}^{(l)}}$ $F=\frac{V_k^{(l)}}{\frac{r_{yy}^{(l)}}{n-l-1}}$ **每一步中剔除结束后再引入**

矩阵变换 这种变换通常被称为旋转变换：

$$a_{ij}^{(l+1)}=\begin{cases}\frac{1}{a_{kk}^{(l)}}(i=k,j=k) & \text{① 枢轴元素}\\ \frac{a_{kj}^{(l)}}{a_{kk}^{(l)}}(i=k,j\neq k) & \text{② 对应行}\\ -\frac{a_{ik}^{(l)}}{a_{kk}^{(l)}}(i\neq k,j=k) & \text{③ 对应列}\\ a_{ij}^{(l)}-\frac{a_{kj}^{(l)}a_{ik}^{(l)}}{a_{kk}^{(l)}}(i\neq k,j\neq k) & \text{④ 其余项}\end{cases}$$

注意
① 上一步刚引入的变量下一步不可能剔除
② 上一步刚剔除的变量下一步不可能引入
③ 连续引入三个变量后考虑剔除

4.3.4 逐步回归的一些注意点与实例

注意点
① 用逐步回归方法选出 p 个重要因子，在理论上和实践上**都不能证明**它们都是最优的。是经验的。
② 用逐步回归方法选取重要因子时，被选中的因子数目 p 的多少是值得注意的。通常 F_α **一般选取 4**，作为否定域的临界值。
③ 逐步回归模型是**正态线性回归模型**。
④ 回归方程的稳定性：是否残差方差小，而且还要注意所得到的规律性在未来时间的样本内是否还存在。统计预报最基本的假设。
⑤ 逐步回归可以与一般回归混合使用。

优点
① 和逐步剔除法相比，**计算量较小**。
② 逐步引入法虽然计算量小些，但是**不一定保证**最后的方程是最优的。
③ 逐步回归方法最后能得到一个**较合理的最优回归方程**。

缺点
① 该方法最终**只提供**一个最优回归方程，而无其它选择的余地。
② 其次，需要**计算机**解决较大阶数的矩阵，对于手算有较大的工作量。

应用 **应用广泛**，工农业生产和科学研究工作中的许多问题都可以用这种方法得到解决。如今，在实验数据的一般处理，经验公式的求得，因素分析，产品质量的控制，某些新标准的制定，气象及地震预报以及许多场合中，都会用到这种分析方法。

现况 目前，由于计算机的普及，逐步回归分析方法已经十分普遍。许多现成的计算软件可供使用。

实例分析

题目：设对某一预报量 y，选择 4 个因子作预报，样本容量 $n=13$，它们的资料见表 1。

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

对上面的资料作逐步回归，具体步骤如下：

准备阶段：构造相关系数的增广矩阵（5 个变量） 这一步不可能存在大于一的值

$$R^{(0)}=\begin{bmatrix}1 & 0.2286 & -0.8241 & -0.2455 & 0.7307\\ 0.2286 & 1 & -0.1392 & -0.9730 & 0.8163\\ -0.8241 & -0.1392 & 1 & 0.0295 & -0.5347\\ -0.2455 & -0.9730 & 0.0295 & 1 & -0.8213\\ 0.7307 & 0.8163 & -0.5347 & -0.8213 & 1\end{bmatrix}$$

下面进行逐步回归计算：

第一步：计算各因子的方差贡献

$V_1{}^{(1)}=\frac{[r_{1y}{}^{(0)}]^2}{r_{11}{}^{(0)}}=\frac{(0.7307)^2}{1}=0.5339$ $V_2{}^{(1)}=0.6663$ $V_3{}^{(1)}=0.2859$ $V_4{}^{(1)}=0.6745=V_{\max}$

$F=\frac{V_{max}}{\frac{r_{yy}^{(0)}-V_{max}}{13-(1+0)-1}}=\frac{0.6745}{\frac{1-0.6745}{13-2}}=22.80>F_a$ 消去第四列，得：

$$R^{(1)}=\begin{bmatrix}0.9397 & -0.0103 & -0.8169 & 0.2455 & 0.5291\\ -0.0103 & 0.0533 & -0.1105 & 0.9730 & 0.0172\\ -0.8169 & -0.1105 & 0.9991 & -0.0295 & -0.5105\\ -0.2455 & -0.9730 & 0.0295 & 1 & -0.8213\\ 0.5291 & 0.0172 & -0.5105 & 0.8213 & 0.3255\end{bmatrix}$$

第二步：计算余下各因子的方差贡献：

$$V_1{}^{(2)}=\frac{[r_{1y}{}^{(1)}]^2}{r_{11}{}^{(1)}}=\frac{(0.5291)^2}{0.9397}=0.2979=V_{\max}\qquad V_2{}^{(2)}=0.0055\qquad V_3{}^{(2)}=0.2609$$

进行检验：$F=\frac{0.2979}{\frac{0.3255-0.2979}{13-2-1}}=107.93>F_a$，消去第一列得：

$$R^{(2)}=\begin{bmatrix}1.0642 & -0.0110 & -0.8693 & 0.2613 & 0.5631\\ 0.0110 & 0.0532 & -0.1195 & 0.9757 & 0.0230\\ 0.8693 & -0.1195 & 0.2890 & 0.1893 & -0.0505\\ 0.2613 & -0.9757 & -0.1893 & 1.0642 & -0.6831\\ -0.5631 & 0.0230 & -0.0505 & 0.6831 & 0.0275\end{bmatrix}$$

第三步：计算余下各因子的方差贡献：

$$V_2^{(3)}=\frac{\left[r_{2y}^{(2)}\right]^2}{r_{22}^{(2)}}=\frac{(0.0230)^2}{0.0532}=0.0099=V_{\max}\qquad V_3^{(3)}=0.0088$$

进行检验：$F=\frac{0.0099}{\frac{0.0275-0.0099}{13-3-1}}=5.034>F_a$，消去第二列得：

$$R^{(3)}=\begin{bmatrix}1.0665 & 0.2068 & -0.8940 & 0.4630 & 0.5679\\ 0.2068 & 18.7970 & -2.2462 & 18.3402 & 0.4323\\ 0.8940 & 2.2462 & 0.0206 & 2.3756 & 0.0012\\ 0.4631 & 18.3402 & 2.3756 & 18.9577 & -0.2613\\ -0.5679 & -0.4323 & 0.0012 & 0.2613 & 0.0177\end{bmatrix}$$

第四步：计算已引入方程中的因子的方差贡献：

$$V_1^{(3)}=\frac{\left[r_{1y}^{(3)}\right]^2}{r_{11}^{(3)}}=\frac{(0.5679)^2}{1.0665}=0.3024\qquad V_4^{(3)}=0.0037=V_{\min}$$

作剔除的检验：$F=\frac{V_k^{(l)}}{\frac{r_{yy}^{(l)}}{n-l-1}}=\frac{0.0037}{\frac{0.0177}{13-3-1}}=1.893<F_a$ 故剔除 x_{z4}，消去第四列。

得到的矩阵相当于仅消去过第 1，2 列的矩阵：

$$R^{(2)}=\begin{bmatrix}1.0552 & -0.2411 & -0.8360 & -0.0245 & 0.5743\\ -0.2411 & 1.0547 & 0.0522 & -0.9684 & 0.6854\\ 0.8360 & -0.0522 & 0.3183 & -0.1254 & 0.0340\\ 0.0245 & 0.9684 & -0.1252 & 0.0528 & -0.0138\\ -0.5743 & -0.6854 & 0.0340 & -0.0138 & 0.0213\end{bmatrix}$$

第五步：计算余下因子x_{z3}的方差贡献：

$$V_3^{(3)}=\frac{\left[r_{3y}^{(2)}\right]^2}{r_{33}^{(2)}}=\frac{(0.0340)^2}{0.3183}=0.0036$$

进行F检验，计算得：$F=\frac{0.0036}{\frac{0.0213-0.0036}{13-3-1}}=1.83<F_a$，认为$x_{z3}$不能引进，又无剔除，逐步回归到此结束。

第六步：计算最后结果：得到标准化变量的回归方程，如下 $\hat{y}_z=0.5743x_{z1}+0.6854x_{z2}$。

利用资料计算得到原始数据的回归方程为：

$$y=52.58+1.468x_1+0.662x_2$$

复相关系数：$R=\sqrt{\frac{U}{S_{yy}}}=\sqrt{1-r_{yy}^{(l)}}=\sqrt{1-0.0213}=0.989$，剩余标准差：$\hat{\sigma}=\sqrt{\frac{S_{yy}r_{yy}^{(l)}}{n-l-1}}=2.39$

消元的具体规则

基于上述案例，我们讨论：

$$R^{(0)}=\begin{bmatrix}1 & 0.2286 & -0.8241 & -0.2455 & 0.7307\\ 0.2286 & 1 & -0.1392 & -0.9730 & 0.8163\\ -0.8241 & -0.1392 & 1 & 0.0295 & -0.5347\\ -0.2455 & -0.9730 & 0.0295 & 1 & -0.8213\\ 0.7307 & 0.8163 & -0.5347 & -0.8213 & 1\end{bmatrix}\xrightarrow{\text{消去第四列}}R^{(1)}=\begin{bmatrix}0.9397 & -0.0103 & -0.8169 & 0.2455 & 0.5291\\ -0.0103 & 0.0533 & -0.1105 & 0.9730 & 0.0172\\ -0.8169 & -0.1105 & 0.9991 & -0.0295 & -0.5105\\ -0.2455 & -0.9730 & 0.0295 & 1 & -0.8213\\ 0.5291 & 0.0172 & -0.5105 & 0.8213 & 0.3255\end{bmatrix}$$

因为选择了 x_4，所以 $k=4$，此时分母总是枢轴元素$a_{kk}^{(l)}=a_{44}^{(0)}=1$。

① 计算枢轴点：利用公式 $1/a_{kk}^{(l)}(i=k,j=k)$，故矩阵中(4,4)位置为 1。

② 计算第四行：利用公式 $\frac{a_{kj}^{(l)}}{a_{kk}^{(l)}}(i=k,j\neq k)$，即新第 4 行 $=\frac{\text{旧第 4 行}}{\text{枢轴元素}}$，由于$a_{kk}^{(l)}=1$，数值不变。

③ 计算第四列：利用公式 $-\frac{a_{ik}^{(l)}}{a_{kk}^{(l)}}(i\neq k,j=k)$，即新第 4 列 $=-\frac{\text{旧第 4 列}}{\text{枢轴元素}}$，这一步需要变号。

④ 计算其他所有元素：较为复杂 $a_{ij}^{(l)}-\frac{a_{kj}^{(l)}a_{ik}^{(l)}}{a_{kk}^{(l)}}(i\neq k,j\neq k)$，有：

$$\text{新元素}=\text{旧元素}-\frac{(\text{该元素所在行的第}k\text{列元素}\times\text{该元素所在列的第}k\text{行元素})}{\text{枢轴元素}}$$

例如，对于(1,2)位置的元素，有：

$$0.2286-\frac{-0.2455\times-0.9730}{1}=0.2286-0.23887=-0.0103$$

依次计算可得。