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：

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

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：

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

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

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：$b = \frac{S_{xy}}{S_x^2} = \frac{S_{xy}}{S_x S_y} \cdot \frac{S_y}{S_x} = r_{xy} \cdot \frac{S_y}{S_x}$，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 $r^2 \to 1$ 说明回归方程拟合的越好，$r^2 \to 0$ 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 $t = \dfrac{b-\beta}{\sqrt{\frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}} = \dfrac{b/\sqrt{c}}{\sqrt{Q/(n-2)}}$ 遵从自由度为 $n-2$ 的 t 分布。

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

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 $U = \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 = \sum_{i=1}^{n}(b_0-bx_i-b_0-b\bar{x})^2 = b^2\sum_{i=1}^{n}(x_i-\bar{x})^2 = \frac{b^2}{c}$

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 $E(y_i) \pm 1.96\sigma$**

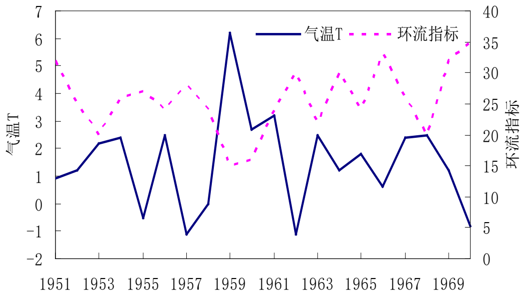
$E(y_i)$ 可用 $b_0 + bx_i = \hat{y}_i$ 估计，σ 可用无偏估计量 $\hat{\sigma} = \sqrt{\frac{Q}{n-2}}$ 估计，$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$

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

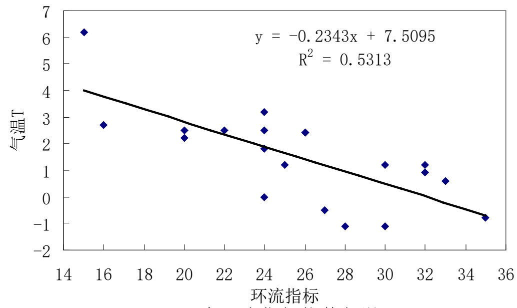
每一个点的置信区间都不一样，置信区间上下界是一个曲线。

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 $n = 20$，$\sum_{i=1}^{\infty} x_i = 513$，$\sum_{i=1}^{\infty} y_i = 30.0$，

$\sum_{i=1}^{1} x_i^2 = 13721$，$\sum_{i=1}^{2} x_i y_i = 637$　根据 $b_0 = \bar{y} - b\bar{x}$，$b = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{S_{xy}}{S_x^2}$　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 $r_{xy} = \sqrt{\frac{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2}} = \sqrt{\frac{13721-\frac{1}{20}\times(513)^2}{103.12-\frac{1}{20}\times(30.0)^2}} \times (-0.23) = -0.727$

最终得到：$F = \frac{(-0.727)^2}{[1-(-0.727)^2]/(20-2)} = 20.18$ 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：$\hat{\sigma} = \sqrt{\frac{1}{20-2} \times 58.12(1-0.727^2)} = 1.11$ ，用 $E(y_i) \pm 1.96\sigma$ 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 **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：

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

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

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：

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

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：

$$\begin{cases} b_{z1}\sum_t x_{z1t}^2 + b_{z2}\sum_t x_{z2t}x_{z1t} + \cdots + b_{zp}\sum_t x_{zpt}x_{z1t} = \sum_t y_{zt}x_{z1t} \\ b_{z1}\sum_t x_{z1t}x_{z2t} + b_{z2}\sum_t x_{z2t}^2 + \cdots + b_{zp}\sum_t x_{zpt}x_{z2t} = \sum_t y_{zt}x_{z2t} \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ b_{z1}\sum_t x_{z1t}x_{zpt} + b_{z2}\sum_t x_{z2t}x_{zpt} + \cdots + b_{zp}\sum_t x_{zpt}^2 = \sum_t y_{zt}x_{zpt} \end{cases}$$

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：

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

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

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。