MATHEMATICA MODEL

气象观测站的调整

制作：刘琼荪

问题

某地区内有12个气象观测站，位置如下图。

十年来各站测得的年降水量如下表所示：

地点 / 年	X_1	X_2	X_3	…	X_{10}	X_{11}	X_{12}
1981	276.2	324.5	158.6	… …	243.2	159.7	331.2
1982	251.6						455.1
1983							
…							
1988							
1989							
1990	324.8	406.5	235.7	… …	281.2	243.7	411.1

欲解决的问题是

为了节省开支，想要适当减少气象观测站。问减少哪些观测站可以使所得到的降水量的信息量仍然足够大？

方法概述：

① 方差、相关性和回归分析相结合；

② 主成分分析或因子分析；

③ 信息熵函数；……

1、问题分析：

欲达到的目的是站数较少，同时得到的信息量仍足够大。在站数与信息量之间，信息量是主要因素。

由于不同站之间年降水量的相关性不仅取决于地理位置的远近，而且取决于地理气象环境的相似性。

而该问题的主要依据是降水量信息，而各站的地理位置信息仅作为参考信息。

问题分解成以下几个问题：

1）删除哪几个站？

2）如何预测被删除站的降水量信息？

3）如何度量上述预测值中所包含的降水量信息是否足够大？

假设：

1）降水量信息反映了各观测站地理气象环境之间的相似性；

2）各站每年的降水量视为一个围绕其均值上下波动的正态随机变量；

模型建立与求解：

首先考虑相关系数矩阵。(xgxs.m)

A=[……];

R=corrcoef(A)

$$R=\begin{bmatrix} 1 & r_{x_1x_2} & \cdots & r_{x_1x_{12}} \\ r_{x_2x_1} & 1 & \vdots & r_{x_2x_{12}} \\ \vdots & \vdots & \ddots & \vdots \\ r_{x_{12}x_1} & r_{x_{12}x_2} & \cdots & 1 \end{bmatrix}, \qquad r_{x_ix_j}=\frac{\operatorname{cov}(x_i,x_j)}{\sqrt{Dx_i\,Dx_j}}$$

具体计算

$$r_{x_ix_j}=\frac{\sum_{k=1}^{10}(x_{ki}-\bar{x}_i)(x_{kj}-\bar{x}_j)}{\sqrt{\sum_{k=1}^{10}(x_{ki}-\bar{x}_i)^2\sum_{k=1}^{10}(x_{kj}-\bar{x}_j)}}$$

R =	x1	x2	x3	x4	x5	x6	x7
1	1.0000	-0.2456	-0.2206	-0.1082	-0.3312	0.3481	-0.2622
2	-0.2456	1.0000	0.0777	-0.3965	0.1085	0.0103	0.1775
3	-0.2206	0.0777	1.0000	-0.5031	-0.2676	0.4077	0.8126
4	-0.1082	-0.3965	-0.5031	1.0000	0.3158	-0.5567	-0.1523
5	-0.3312	0.1085	-0.2676	0.3158	1.0000	-0.6136	-0.0977
6	0.3481	0.0103	0.4077	-0.5567	-0.6136	1.0000	0.2236
7	-0.2622	0.1775	0.8126	-0.1523	-0.0977	0.2236	1.0000
8	0.0836	-0.1697	-0.0858	0.1384	-0.3156	0.0417	-0.3241
9	0.2928	0.1943	0.1915	-0.2029	-0.2524	0.4918	0.0531
10	-0.3098	0.2421	-0.0175	-0.2501	0.5817	-0.0106	-0.0674
11	0.1587	0.0563	0.3278	-0.4201	0.0844	0.6544	0.0635
12	0.2786	-0.3460	0.0111	-0.2597	-0.2037	0.3631	-0.5152

续上表

	x8	x9	x10	x11	x12
1	0.0836	0.2928	-0.3098	0.1587	0.2786
2	-0.1697	0.1943	0.2421	0.0563	-0.3460
3	-0.0858	0.1915	-0.0175	0.3278	0.0111
4	0.1384	-0.2029	-0.2501	-0.4201	-0.2597
5	-0.3156	-0.2524	0.5817	0.0844	-0.2037
6	0.0417	0.4918	-0.0106	0.6544	0.3631
7	-0.3241	0.0531	-0.0674	0.0635	-0.5152
8	1.0000	0.6348	-0.5946	-0.0762	0.5392
9	0.6348	1.0000	-0.4225	0.4272	0.4259
10	-0.5946	-0.4225	1.0000	0.4969	0.0026
11	-0.0762	0.4272	0.4969	1.0000	0.5652
12	0.5392	0.4259	0.0026	0.5652	1.0000

总结

强相关变量：x_3, x_7 (0.75~0.81)

中度相关变量：当$0.5<|\rho_{ij}|<0.7$时，称x_i, x_j中度相关。

中度相关的变量有：x_4, x_6, x_8, x_9, x_{10}, x_{11},x_{12}

轻度相关变量：x_1, x_2, x_5

显然，强相关变量与中度相关变量应优先考虑被删除。

b=std(A)

其次考虑标准差。将各站降水量的标准差列表：

x	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12
$\sqrt{Dx}$	100.2	81.8	108.2	64.0	94.1	94.2	38.1	85.1	109.4	57.3	86.5	36.8

对于标准差较大的站，认为它们包含的信息量比较大。所以应优先考虑删除标准差较小的观测站。因此，考虑删除的站有：(9→7)

$$D=\{x_4, x_6, x_7, x_8, x_{10}, x_{11}, x_{12}\}$$

排序：4　7　1　5　3　6　2

思考：究竟考虑删除几个观测站呢？

用线性回归方法根据保留站观测数据去预测被删除站的观测值。估计应该在集合D中删除3~5个变量。其组合数有

$$C_7^3 + C_7^4 + C_7^5 = 91$$

对每一种组合用回归方法进行线性拟合，根据剩余方差最小，每一个回归自变量显著为标准选出最佳组合。假定欲删除4个变量。

例

1)考虑删除观测站x_4, x_7, x_{10},x_{12}进行回归拟合：

$x_4=f_1(x_1,x_2,x_3,x_5,x_6,x_8,x_9, x_{11})$

$x_7=f_2(x_1,x_2,x_3,x_5,x_6,x_8,x_9, x_{11})$

$x_{10}=f_3(x_1,x_2,x_3,x_5,x_6,x_8,x_9, x_{11})$

$x_{12}=f_4(x_1,x_2,x_3,x_5,x_6,x_8,x_9, x_{11})$

其结果见huig2.m

2)考虑删除观测站x_6, x_7, x_8,x_{10}进行回归拟合：

$x_6=f_1(x_1,x_2,x_3,x_4,x_5,x_9,x_{11}, x_{12})$

$x_7=f_2(x_1,x_2,x_3,x_4,x_5,x_9,x_{11}, x_{12})$

$x_8=f_3(x_1,x_2,x_3,x_4,x_5,x_9,x_{11}, x_{12})$

$x_{10}=f_4(x_1,x_2,x_3,x_4,x_5,x_9,x_{11}, x_{12})$

其结果见huig1.m

```
[b1,bint1,r1,rint1,stats1]=regress(Y1,X);
```

两种情况对比：

删除情况	σ_1^2	σ_2^2	σ_3^2	σ_4^2
▲x_6,x_7,x_8,x_{10}	160.36	0.5855	6217.9	702.11
x_4,x_7,x_{10},x_{12}	12754	1421.1	114.82	1340.9

情形一▲	Stats1	0.988	62.129	0.0978
	Stats2	1	2781.4	0
	Stats3	0.9045	1.1845	0.615
	Stats4	0.9762	5.1262	0.3296
情形二	Stats1	0.6538	0.2360	0.9265
	Stats2	0.8909	1.021	0.6487
	Stats3	0.9961	31.986	0.1360
	Stats4	0.8902	1.0131	0.6505

当确定删除变量x_6，x_7，x_8，x_{10}的方案后，还可以通过逐步回归方法确定最佳回归方程。

stepwise(X, y, inmodel,alfha)（huig11.m)

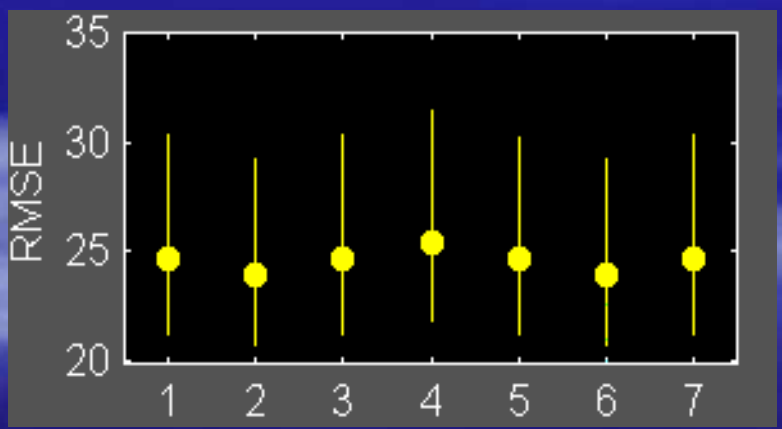
例如输入： X=[x_1,x_2,x_3,x_4];

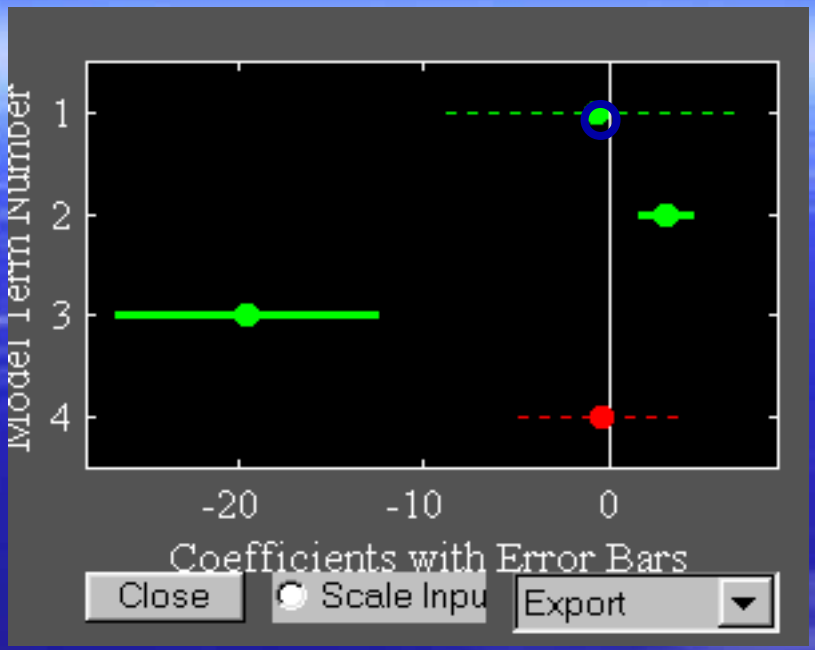
```
stepwise(X,y,[1,2,3])
```

注意：将得到一个交互式的界面

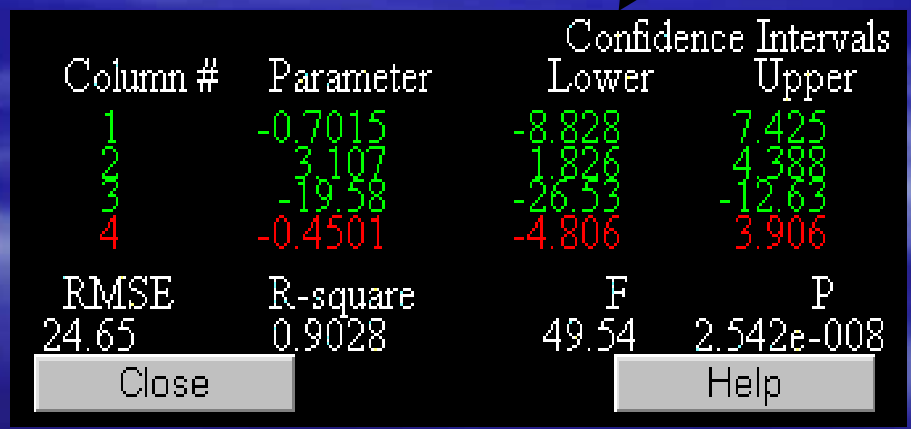
输出结果：

模型中均方差历史数据记载表





参变量数据分析表



输出结果：

$$x_6 = -0.205x_2 - 0.1632x_3 - 0.2897x_4 - 0.7537x_5 + 1.041x_{11} - 1.129x_{12} + a$$

$$x_7 = 0.3026x_3 + 0.127x_4 + 0.088x_{11} - 0.6017x_{12} + b$$

$$x_8 = -0.1695x_1 + 0.2115x_4 + 0.5543x_9 - 0.6794x_{11} + 1.67x_{12} + c$$

$$x_{10} = -0.1399x_1 - 0.1759x_3 - 0.311x_4 + 0.166x_5 - 0.2931x_9 + 0.54x_{11} - 0.29x_{12} + d$$

通过各种情况的比较，得到如下的结果：删除观测站x_6, x_7, x_8, x_{10}，得到四个相对应的回归方程（如前面所示）。四个回归方程的剩余方差分别为：

$$\sigma_6^2 = 160.36, \quad \sigma_7^2 = 0.59, \quad \sigma_8^2 = 6217.9, \quad \sigma_{10}^2 = 702.11$$

结果分析：

定义信息损失比率：

$$\delta_i = \frac{\sigma_i}{\bar{x}_i} \times 100\%, \qquad i = 6,7,8,10$$

计算出删除6，7，8，10四站的信息损失比率分别为：

dert =

4.02%　0.22%　25.96%　8.85%

平均信息损失为： 9.76%

删除4个站的降水量信息仍保留了约90%，结果比较令人满意。

思考：

除了以上分析方法以外是否还有其它统计方法可以分析呢？

如主成分分析法、定义信息熵函数等。

最大熵原理

熵(Entropy)是分子热力学中的一个概念，用以描述分子随机运动的无序程度。分子运动越是无序，则熵越大。

在信息论中，此概念被用以衡量随机试验得到的信息量的大小。熵越大，信息量越大。

定义 $H(p_1, p_2, \ldots, p_n) = -c\sum_{i=1}^{n} p_i \ln p_i$

其中c是常数，p_i=P{X=x_i}, i =1,2,…,n,

一般取 c=1. 不加任何限制，当p_i=1/n时，熵最大。

定义

假设X是连续型随机变量，X的密度函数是p(x)，则熵函数定义为：

$$H(p) = -\int_{-\infty}^{\infty} p(x)\ln p(x)dx$$

性质

1、在方差一定的连续型概率分布中，以正态分布的熵最大；

2、对于任何一个基本对称系统，其状态的概率分布应在表征这个系统状态的约束条件下，使这个分布所定义的熵最大；

例

将系统状态限制在有限区间之内时，使熵最大的分布是该区间上的均匀分布。

对气象观测站问题，可以计算在各观测站下标准差 σ_i 的值，定义 $p_i = \sigma_i / \Sigma \sigma_j$, 或

$$p_i = \overline{x}_i / \sum_{j=1}^{12} \overline{x}_j$$

可以计算当删除某个观测站，信息熵的变化值。即

$$L_1 = H(p_1, p_2, \ldots, p_n) - H(p'_2, \ldots, p'_n) \Rightarrow \min$$