11 综合实例：投资额与国民生产总值和物价指数

- 建立投资额模型，研究某地区实际投资额与国民生产总值（GNP）及物价指数（PI）的关系,根据对未来GNP及PI的估计，预测未来投资额 。以下是地区连续20年的统计数据：

年份	投资额	国民生产总值	物价指数		年份	投资额	国民生产总值	物价指数
1	90.9	596.7	0.7167		11	229.8	1326.4	1.0575
2	97.4	637.7	0.7277		12	228.7	1434.2	1.1508
3	113.5	691.1	0.7436		13	206.1	1549.2	1.2508
4	125.7	756.0	0.7676		14	257.9	1718.0	1.3234
5	122.8	799.0	0.7906		15	324.1	1918.3	1.4005
6	133.3	873.4	0.8254		16	386.6	2163.9	1.5042
7	149.3	944.0	0.8679		17	423.0	2417.8	1.6342
8	144.2	992.7	0.9145		18	401.9	2631.7	1.7842
9	166.4	1077.6	0.9601		19	474.9	2954.7	1.9514
10	195.0	1185.9	1.0000		20	424.5	3073.0	2.0688

投资额与国民生产总值和物价指数

分析

许多经济数据在时间上有一定的滞后性

以时间为序的数据，称为时间序列

时间序列中同一变量的顺序观测值之间存在自相关

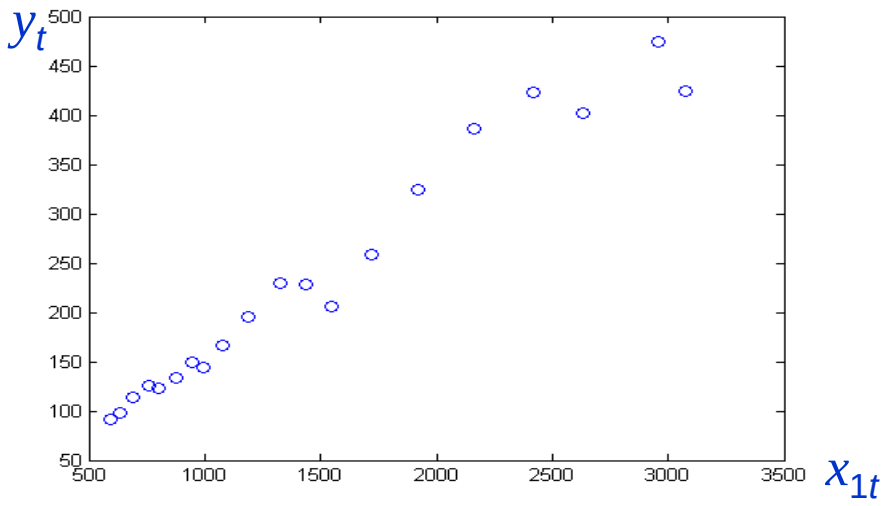
若采用普通回归模型直接处理，将会出现不良后果

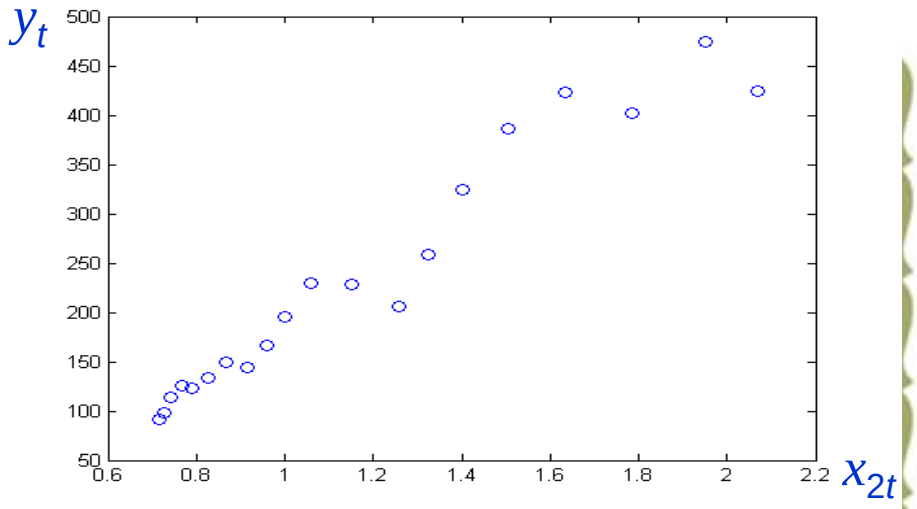
需要诊断并消除数据的自相关性，建立新的模型

年份序号	投资额	国民生产总值	物价指数	年份序号	投资额	国民生产总值	物价指数
1	90.9	596.7	0.7167	11	229.8	1326.4	1.0575
2	97.4	637.7	0.7277	12	228.7	1434.2	1.1508
3	113.5	691.1	0.7436	13	206.1	1549.2	1.2579
4	125.7	756.0	0.7676	14	257.9	1718.0	1.3234
…	…	…	…	…	…	…	…

基本回归模型

t ~年份， y_t ~ 投资额， x_{1t} ~ GNP， x_{2t} ~ 物价指数





投资额与 GNP及物价指数间均有很强的线性关系

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \varepsilon_t$$

$\beta_0, \beta_1, \beta_2$ ~回归系数

ε_t ~对t相互独立的零均值正态随机变量

基本回归模型的结果与分析

MATLAB 统计工具箱

参数	参数估计值	置信区间
β_0	322.7250	[224.3386 421.1114]
β_1	0.6185	[0.4773 0.7596]
β_2	-859.4790	[-1121.4757 -597.4823]
R^2= 0.9908	F= 919.8529	p=0.0000

$$\hat{y}_t = 322.725 + 0.6185x_{1t} - 859.479x_{2t}$$

模型优点 R^2=0.9908，拟合度高

模型缺点 没有考虑时间序列数据的滞后性影响可能忽视了随机误差存在自相关；如果存在自相关性，用此模型会有不良后果

自相关性的定性诊断

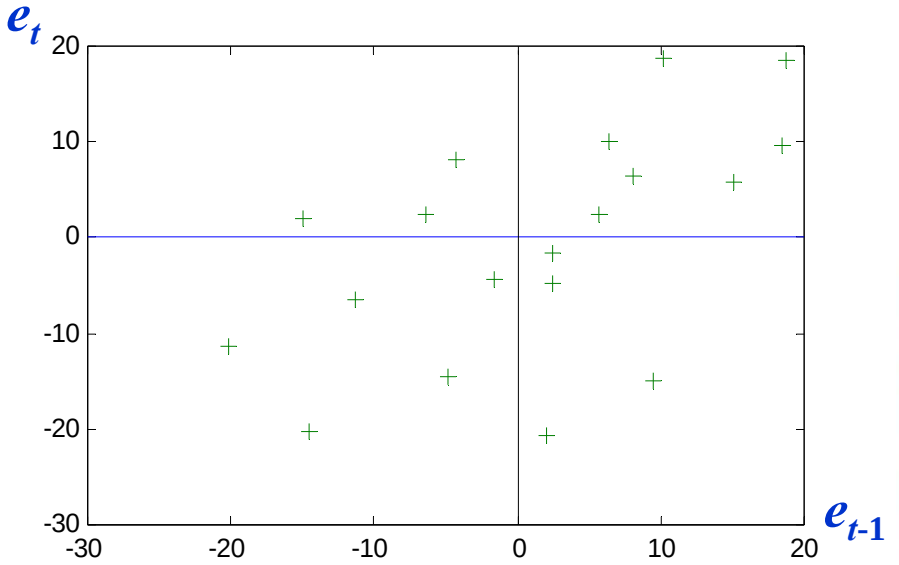
残差诊断法

模型残差 $e_t = y_t - \hat{y}_t$

e_t为随机误差ε_t的估计值

在MATLAB工作区中输出

作残差 e_t~e_{t-1} 散点图



大部分点落在第1, 3象限 ⇨ ε_t存在正的自相关

大部分点落在第2, 4象限 ⇨ ε_t存在负的自相关

自相关性直观判断 ⇨ 基本回归模型的随机误差项ε_t 存在正的自相关

自回归性的定量诊断

D-W检验

自回归模型 $y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \varepsilon_t, \quad \varepsilon_t = \rho \varepsilon_{t-1} + u_t$

$\beta_0, \beta_1, \beta_2$ ~回归系数　　ρ ~自相关系数　　$|\rho| \le 1$

u_t ~对t相互独立的零均值正态随机变量

$\rho = 0$ ⇨ 无自相关性

$\rho > 0$ ⇨ 存在正自相关性

$\rho < 0$ ⇨ 存在负自相关性

如何估计 ρ ⇔ **D-W统计量**

如何消除自相关性 ⇔ 广义差分法

D-W统计量与D-W检验

$$DW=\frac{\sum_{t=2}^{n}(e_t-e_{t-1})^2}{\sum_{t=2}^{n}e_t^2}\approx 2\left[1-\frac{\sum_{t=2}^{n}e_te_{t-1}}{\sum_{t=2}^{n}e_t^2}\right]$$

*n*较大

$$\hat{\rho}=\sum_{t=2}^{n}e_te_{t-1}/\sum_{t=2}^{n}e_t^2$$

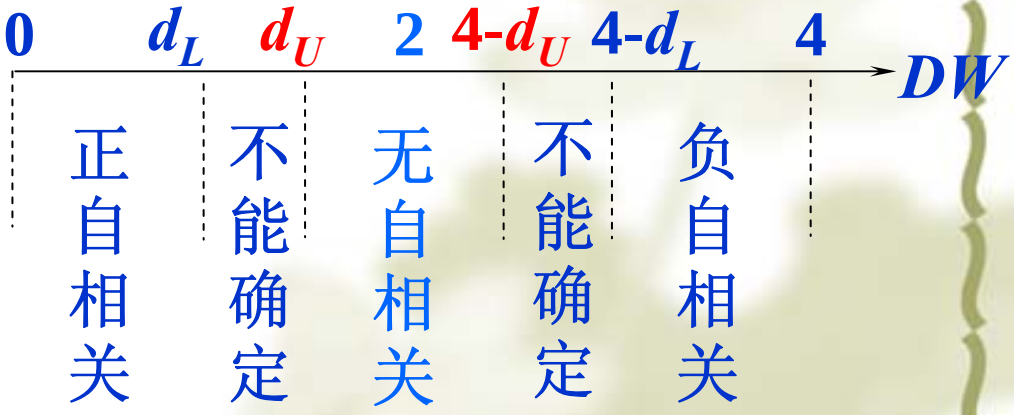
$$=2(1-\hat{\rho})$$

$$-1\le\hat{\rho}\le 1\rightarrow 0\le DW\le 4$$

$$\hat{\rho}=1\rightarrow DW=0$$

$$\hat{\rho}=-1\rightarrow DW=4$$

$$\hat{\rho}=0\rightarrow DW=2$$



检验水平, 样本容量，回归变量数目

D-W分布表

检验临界值d_L和d_U

由*DW*值的大小确定自相关性

广义差分变换

$$DW = 2(1-\hat{\rho}) \Rightarrow \hat{\rho} = 1 - \frac{DW}{2}$$

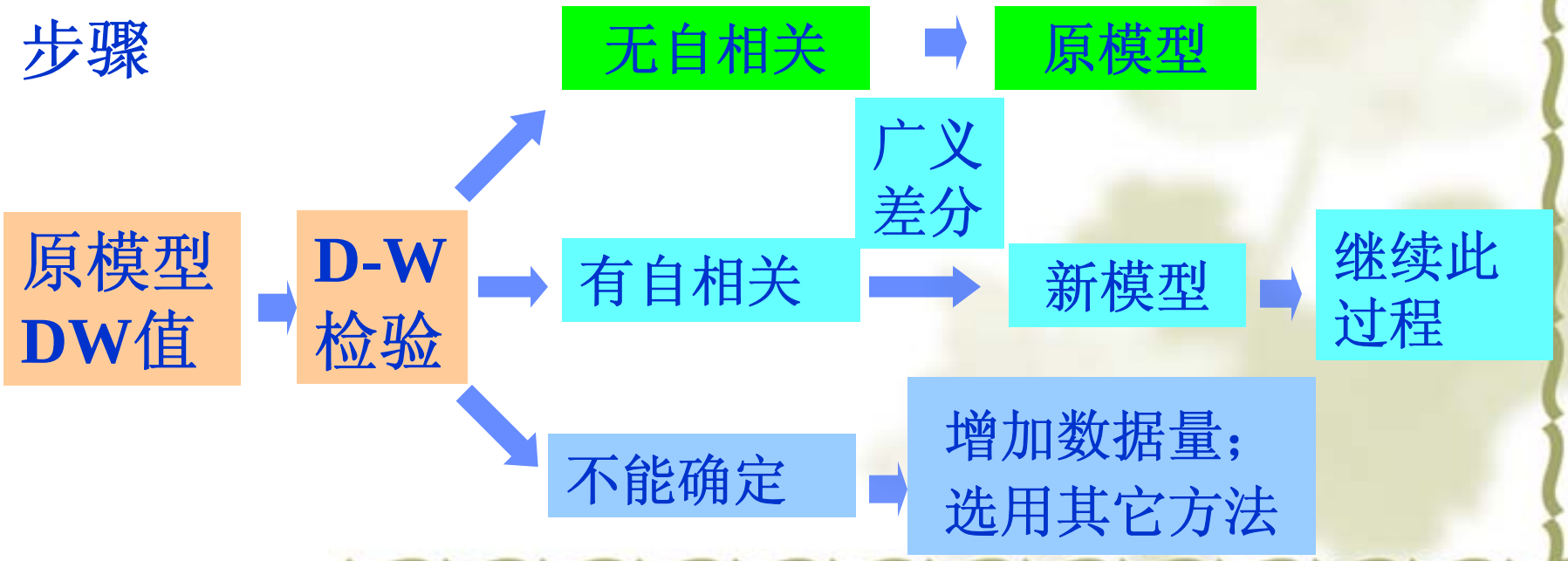
原模型 $y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \varepsilon_t, \quad \varepsilon_t = \rho\varepsilon_{t-1} + u_t$

变换 $y_t^* = y_t - \rho y_{t-1}, \; x_{it}^* = x_{it} - \rho x_{i,t-1}, \quad i = 1,2$

新模型 $y_t^* = \beta_0^* + \beta_1 x_{1t}^* + \beta_2 x_{2t}^* + u_t \quad \beta_0^* = \beta_0(1-\rho)$

以$\beta^*_0, \beta_1, \beta_2$为回归系数的普通回归模型

步骤



投资额新模型的建立

$$DW = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=2}^{n} e_t^2}$$

原模型　　DW_{old}=

残差e_t　　0.8754

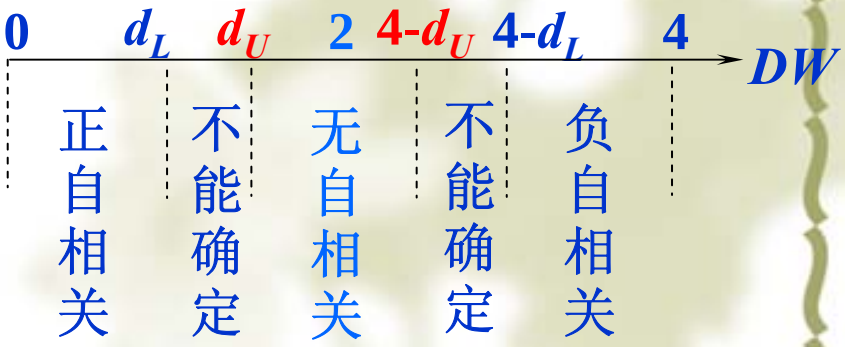
样本容量n=20，回归变量数目k=3，α=0.05

查表 ⇩

临界值d_L=1.10, d_U=1.54

$DW_{old} < d_L$ ➡ 原模型有正自相关

$$\hat{\rho} = 1 - DW / 2 = 0.5623$$



作变换

$$y_t^* = y_t - 0.5623 y_{t-1}$$

$$x_{it}^* = x_{it} - 0.5623 x_{i,t-1}, \quad i = 1,2$$

投资额新模型的建立

$$y_t^* = y_t - 0.5623 y_{t-1} \qquad x_{it}^* = x_{it} - 0.5623 x_{i,t-1}, \quad i = 1,2$$

$$y_t^* = \beta_0^* + \beta_1 x_{1t}^* + \beta_2 x_{2t}^* + u_t$$

由数据 $y_t^*, x_{1t}^*, x_{2t}^*$ 估计系数 β_0^*, β_1, β_2

参数	参数估计值	置信区间
β_0^*	163.4905	[1265.4592 2005.2178]
β_1	0.6990	[0.5751 0.8247]
β_2	-1009.0333	[-1235.9392 -782.1274]
R^2= 0.9772	F=342.8988	p=0.0000

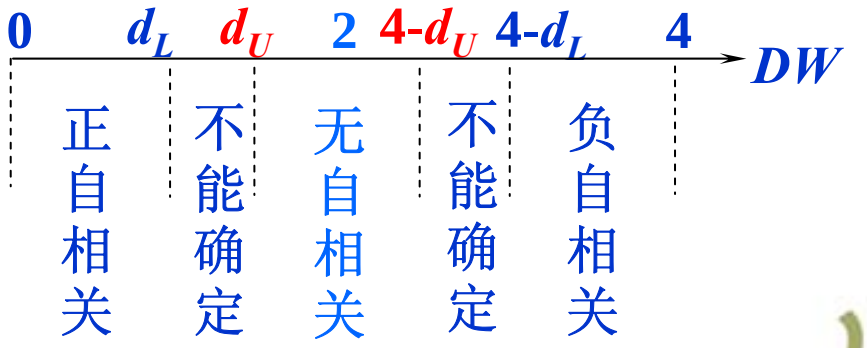
总体效果良好

剩余标准差

s_{new}= 9.8277 < s_{old}=12.7164

新模型的自相关性检验

新模型残差e_t ➡ DW_{new}= 1.5751



样本容量n=19，回归变量数目k=3，α=0.05

查表 ⇩

临界值d_L=1.08, d_U=1.53

$d_U < DW_{new} < 4-d_U$

⬇

新模型无自相关性

新模型 $\hat{y}_t^* = 163.4905 + 0.699\, x_{1t}^* - 1009.033\, x_{2t}^*$

还原为原始变量

$$\hat{y}_t = 163.4905 + 0.5623\, y_{t-1} + 0.699\, x_{1,t} - 0.3930\, x_{1,t-1} - 1009.0333\, x_{2,t} + 567.3794\, x_{2,t-1}$$

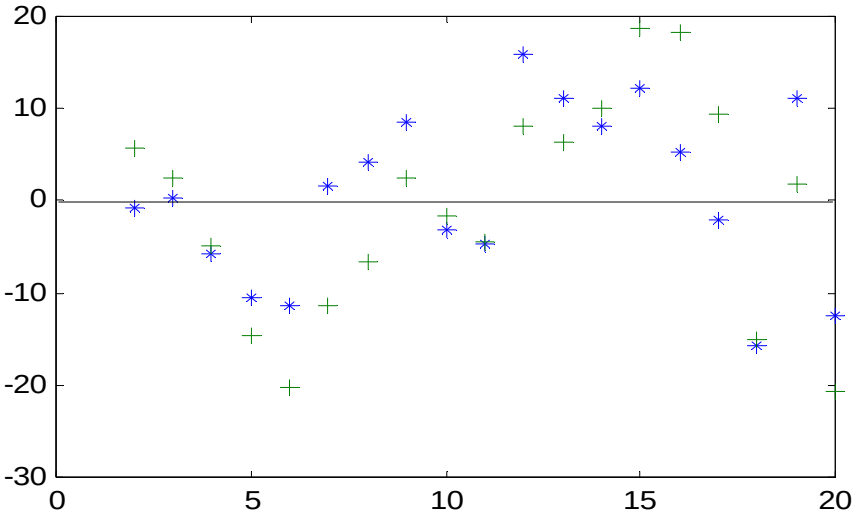
一阶自回归模型

模型结果比较

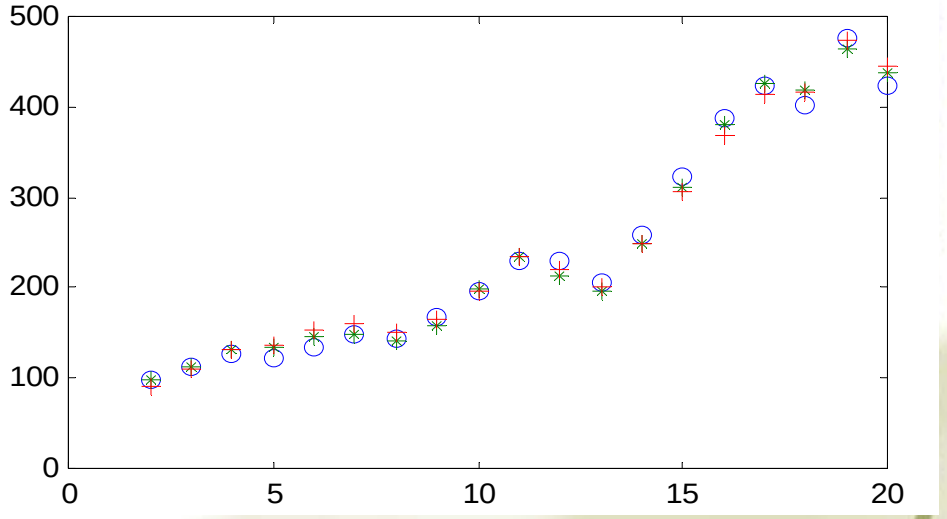
基本回归模型 $\hat{y}_t = 322.725 + 0.6185x_{1t} - 859.479x_{2t}$

一阶自回归模型 $\hat{y}_t = 163.4905 + 0.5623y_{t-1} + 0.699x_{1,t} - 0.3930x_{1,t-1}$
$-1009.0333x_{2,t} + 567.3794x_{2,t-1}$

残差图比较



拟合图比较



新模型 e_t~ *，原模型 e_t~ +

新模型 $\hat{y}_t$ ~ *，新模型 $\hat{y}_t$ ~ +

一阶自回归模型残差e_t比基本回归模型要小

投资额预测

对未来投资额y_t作预测，需先估计出未来的国民生产总值x_{1t}和物价指数x_{2t}

年份序号	投资额	国民生产总值	物价指数	年份序号	投资额	国民生产总值	物价指数
1	90.9	596.7	0.7167	18	401.9	2631.7	1.7842
2	97.4	637.7	0.7277	19	474.9	2954.7	1.9514
3	113.5	691.1	0.7436	20	424.5	3073.0	2.0688

设已知 t=21时，x_{1t}=3312，x_{2t}=2.1938

基本回归模型 $\hat{y}_t = 485.6720$

一阶自回归模型 $\hat{y}_t = 469.7638$

$\hat{y}_t$ 较小是由于y_{t-1}=424.5过小所致

6聚类分析

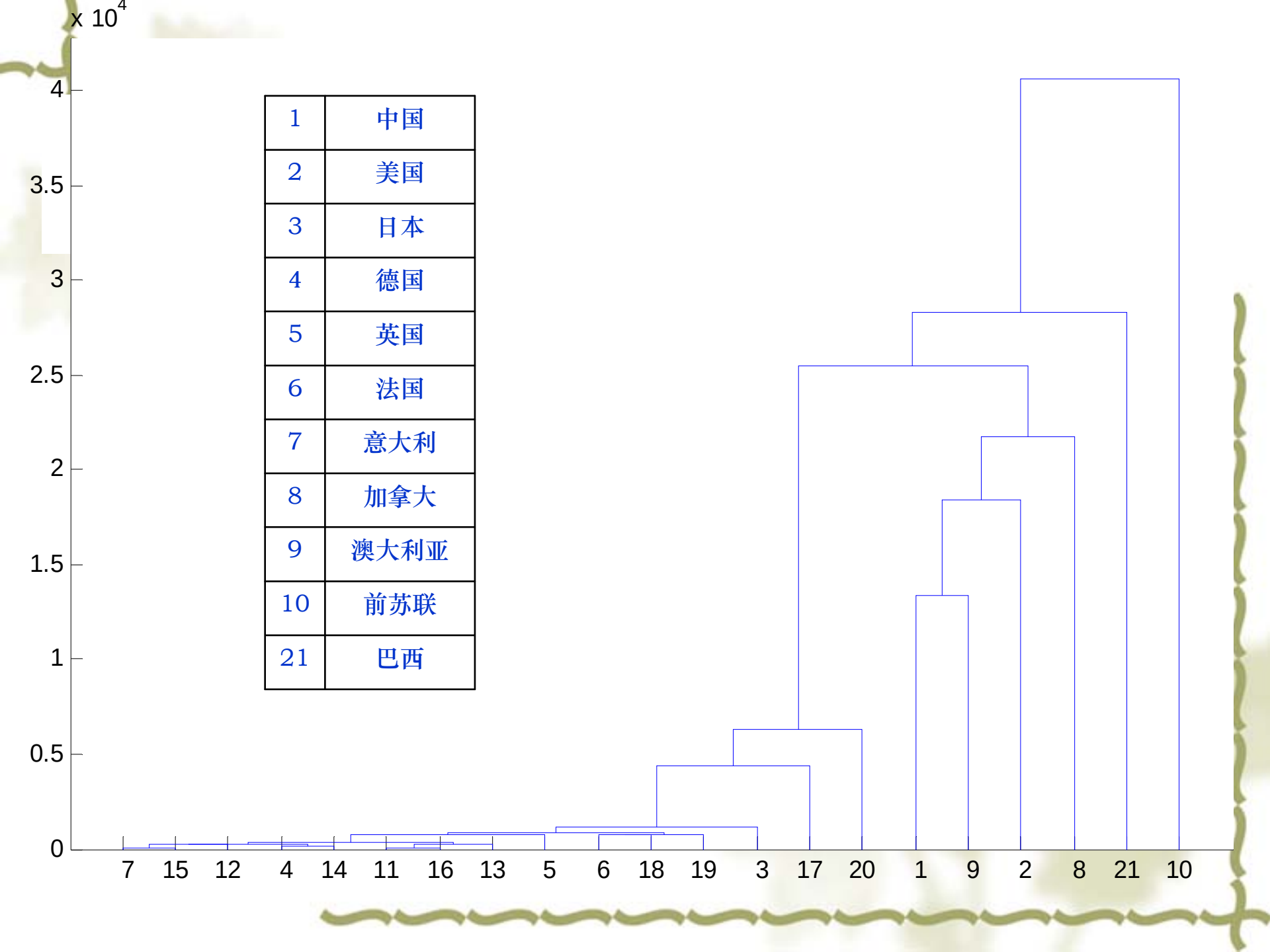
- 人类认识世界往往首先将被认识的对象进行分类，聚类分析是研究分类问题的多元数据分析方法，是数值分类学中的一支。
- 聚类分析的基本思想是在样品之间定义距离，在变量之间定义相似系数，距离或相似系数代表样品或变量之间的相似程度。按相似程度的大小，将样品（或变量）逐一归类，关系密切的类聚到一个小的分类单位，然后逐步扩大，使得关系疏远的聚合到一个大的分类单位，直到所有的样品（或变量）都聚集完毕，形成一个表示亲疏关系的谱系图，依次按照某些要求对样品（或变量）进行分类。

- 为了研究世界各国森林、草原资源的分布规律，共抽取了21个国家的数据，每个国家4项指标，原始数据见下表1。使用该原始数据对国别进行聚类分析。

国别	森林面积（万公顷）	森林覆盖率（%）	林木蓄积量（亿立方米）	草原面积（万公顷）
中国	11978	12.5	93.5	31908
美国	28446	30.4	202.0	23754
日本	2501	67.2	24.8	58
德国	1028	28.4	14.0	599
英国	210	8.6	1.5	1147
法国	1458	26.7	16.0	1288
意大利	635	21.1	3.6	514
加拿大	32613	32.7	192.8	2385
澳大利亚	10700	13.9	10.5	45190
前苏联	92000	41.1	841.5	37370

- Matlab提供了两种方法进行聚类分析。

- 一种是利用 clusterdata函数对样本数据进行一次聚类，其缺点为可供用户选择的面较窄，不能更改距离的计算方法；
- clusterdata函数
- 调用格式：T=clusterdata(X,…)
- 说明：根据数据创建分类。



1	中国
2	美国
3	日本
4	德国
5	英国
6	法国
7	意大利
8	加拿大
9	澳大利亚
10	前苏联
21	巴西

- 另一种是分步聚类：
- Step1 寻找变量之间的相似性
- 用pdist函数计算相似矩阵，有多种方法可以计算距离，进行计算之前最好先将数据用zscore函数进行标准化。
- X2=zscore(X); %标准化数据
- Y2=pdist(X2); %计算距离

‘euclidean’：欧氏距离（默认）；

‘seuclidean’：标准化欧氏距离；

‘mahalanobis’：马氏距离；

‘cityblock’：布洛克距离；

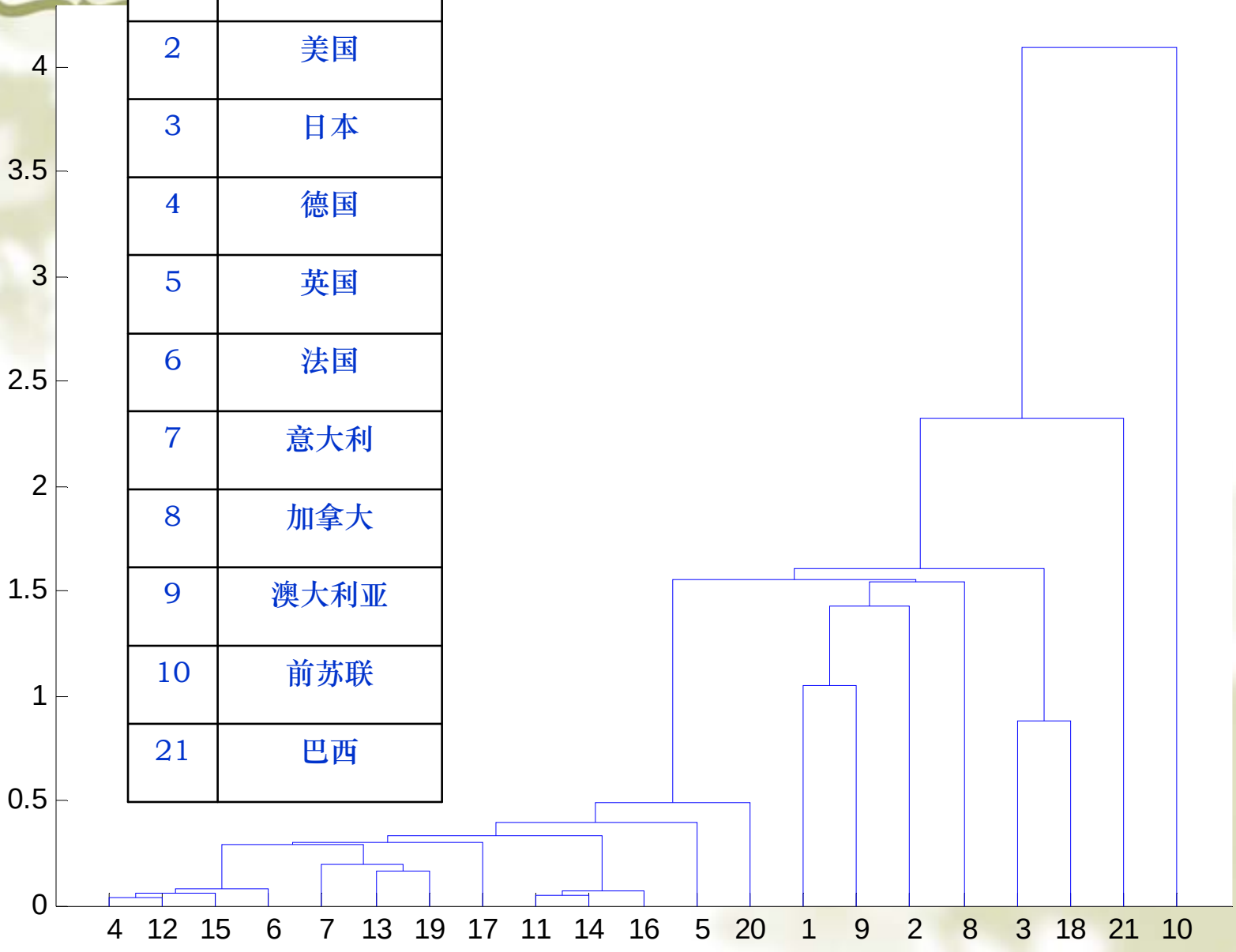
‘minkowski’：明可夫斯基距离；

- Step2　定义变量之间的连接
- Z2=linkage(Y2);

- Step3　评价聚类信息
- 　C2=cophenet(Z2,Y2);　　//0.94698

- Step4 创建聚类，并作出谱系图
- 　　T=cluster(Z2,6);
- 　　H=dendrogram(Z2);

1	中国
2	美国
3	日本
4	德国
5	英国
6	法国
7	意大利
8	加拿大
9	澳大利亚
10	前苏联
21	巴西



分类结果：{加拿大}，{中国，美国，澳大利亚}，{日本，印尼}，{巴西}，{前苏联}剩余的为一类。