第九章　统计分析

教学内容	第一节 回归分析
教学目标	**知识目标：** 理解回归分析的概念；会用 Matlab 软件处理回归分析问题。 **能力目标：**通过实例演示、问题驱动等方式激发学生学习的积极性，通过观察对比、学生交流、师生交流、小组探究等多种形式， 培养学生的逻辑思维能力、分析判断能力、和解决问题能力。 **素质目标：** 培养学生敢于质疑、善于分析、勇于创新的精神；培养学生的自主学习意识和团队协作精神；培养学生脚踏实地、不畏艰辛、锲而不舍的精神。
授课学时	2 学时
教学重点	Matlab 软件求解回归分析问题
教学难点	回归分析模型的建立。
教学方法	采用问题驱动法、案例演示法、启发式讲授法及自主学习法相结合，以教师的讲解为主，学生的课堂报告、分组讨论为辅，充分调动学生学习的主动性和思考问题的积极性。
教学手段	以课堂讲授为主，主要是多媒体课件和板书相结合的形式。同时借助在线资源，如慕课、雨课堂等平台。
教学过程	**（一）由引例出发，介绍一元线性回归模型的概念（约 15 分钟）** 【教师活动】先引入具体案例，由案例出发，引导学生预测变量间的函数关系的可能形式。 案例1：合金的强度 $y(\mathrm{kg/mm^2})$ 与其中的碳含量 $x(\%)$ 有比较密切的关系，现有一批数据如下表，试研究这些数据之间的规律性。

观测数据

x	0.10	0.11	0.12	0.13	0.14	0.15	0.16	0.17	0.18	0.20	0.22	0.24
y	42.0	42.5	45.0	45.5	45.0	47.5	49.0	51.0	50.0	55.0	57.5	59.5

解：首先利用 Matlab 软件绘出散点图如下：

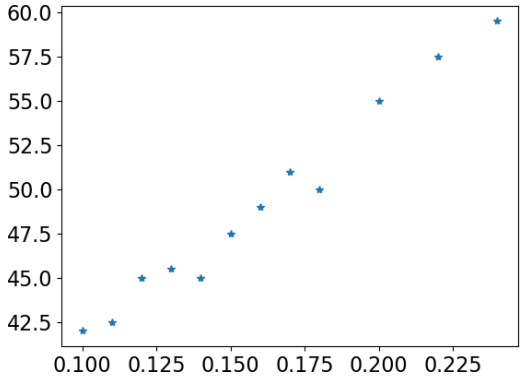


图 9.1 合金强度与碳含量关系散点图

从散点图 9.1 中可知，所有数据点基本上聚集在某一条直线附近，这说明变量 y 与 x 大致上可以看作线性关系。不过这些点又

不都在一条直线上，这表明 y 与 x 之间的关系不是确定性关系。实际上合金的强度 y 除了与碳含量 x 有一定的关系外，还受到许多其他因素的影响。因此 y 与 x 之间可以假定有如下结构式：

$$y=\beta_0+\beta_1 x+\varepsilon, \tag{9.1}$$

因此可以假定 y 与 x 大致上为线性关系。即可以建立一元线性回归模型：

$$\begin{cases} y=\beta_0+\beta_1 x+\varepsilon, \\ E(\varepsilon)=0, D(\varepsilon)=\sigma^2(\text{未知}) \end{cases} \tag{9.2}$$

其中 β_0 为回归常数， β_1 为回归系数，自变量 x 为回归变量。

ε $\sim N(0,\sigma^2)$ 为随机误差。

在（9.1）式两端同时取期望可得： $y=\beta_0+\beta_1 x$ ，称为 y 对 x 的

回归直线方程。

知识点 1：一元线性回归模型具体形式如下：

$$\begin{cases} y=\beta_0+\beta_1 x+\varepsilon, \\ E(\varepsilon)=0, D(\varepsilon)=\sigma^2 (\text{未知}) \end{cases}$$

其中 β_0 为回归常数， β_1 为回归系数，自变量 x 为回归变量。

$\varepsilon \sim N(0,\sigma^2)$ 为随机误差。

【学生活动】讨论自变量和因变量间的函数关系情况，并分析原因。

【设计意图】从学生熟知的实际问题出发，更容易切入问题， 提高学生的学习积极性。

（二）采用类比法，介绍多元线性回归的概念。（约 15 分钟）

一元线性回归是反映一个因变量与一个自变量的之间关系的回归模型，而在实际应用中，某个因变量（响应变量）通常与多个自变量（因素）之间存在相互依赖关系，这时就需要利用多元回归模型进行描述。

知识点 2：多元线性回归模型可以表示为

$$\begin{cases} y=\beta_0+\beta_1 x_1+\cdots+\beta_m x_m+\varepsilon_i, \\ \varepsilon_i \sim N(0,\sigma^2), \end{cases} \tag{9.3}$$

其中 σ 未知，$\beta=(\beta_0,\beta_1,\cdots,\beta_m)$ 称为**回归系数向量**。

现得到一个样本容量为 n 的样本，其观测数据为

$$(y_i, x_{i1}, \cdots, x_{im}), i=1,\cdots,n(n>m).$$

代入（9.3），得

$$\begin{cases} y_i=\beta_0+\beta_1 x_{i1}+\cdots+\beta_m x_{im}+\varepsilon_i, \\ \varepsilon_i \sim N(0,\sigma^2), i=1,\cdots,n. \end{cases} \tag{9.4}$$

记

$$X=\begin{bmatrix} 1 & x_{11} & \cdots & x_{1m} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{nm} \end{bmatrix}, Y=\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix},$$
$$\varepsilon=(\varepsilon_1,\cdots,\varepsilon_n)^T, \beta=(\beta_0,\beta_1,\cdots,\beta_m)^T. \tag{9.5}$$

于是（9.4）可以写成矩阵形式：

$$\begin{cases} Y = X\beta + \varepsilon, \\ \varepsilon \sim N(0, \sigma^2). \end{cases} \qquad (9.6)$$

（三）结合案例，讲授多项式回归的概念（约 15 分钟）

知识点 3：多项式回归模型是线性回归模型的一种，此时回归函数关于回归系数是线性的。由于任一函数都可以用多项式逼近，因此多项式回归有着广泛应用。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，或者用线性模型的

效果不太好，就可以选用多项式回归。在随机意义下，一元多项式回归的数学模型可以表达为

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_n x^n + \varepsilon,$$

其中ε 为随机误差，满足 $E(\varepsilon) = 0, D(\varepsilon) = \sigma^2$ 。

案例 2：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 9.2 所示，试建立二者之间的关系。

表9.1 **年龄与运动能力关系测量数据**

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

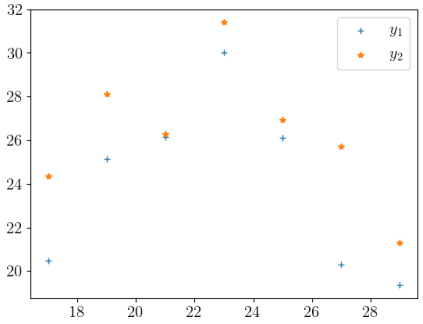
解： 先画出散点图如图 9.2 所示。

图9.2 观测结果散点图

从图中可以看出，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。建立二次多项式模型

$$y = a_2 x^2 + a_1 x + a_0 + \varepsilon .$$



输出结果中，p 返回拟合多项式的系数向量，即回归预测方程为 $y=-0.2003x^2+8.9782x-72.2150$.

（四）采用问题驱动法、案例演示法，针对实际数据，综合多种建模方法，建立多元二次多项式回归模型（40 分钟）

【教师活动】给出具体的案例数据，让学生自由讨论，并随机选择学生到台上阐述自己建立模型的思路，最后老师总结步骤。

案例 3：根据表 9.2 某猪场 25 头育肥猪 4 个胴体性状的数据资料，试进行瘦肉量 y 对眼肌面积（x_1）、腿肉量(x_2)、腰肉量(x_3) 的多元回归分析。

表 9.2 某养猪场数据资料

序号	瘦肉量 y (kg)	眼肌面积x_1 (cm^2)	腿肉量 x_2(kg)	腰肉量 x_3(kg)	序号	瘦肉量 y (kg)	眼肌面积x_1 (cm^2)	腿肉量 x_2(kg)	腰肉量 x_3(kg)
1	15.02	23.73	5.49	1.21	14	15.94	23.52	5.18	1.98
2	12.62	22.34	4.32	1.35	15	14.33	21.86	4.86	1.59
3	14.86	28.84	5.04	1.92	16	15.11	28.95	5.18	1.37
4	13.98	27.67	4.72	1.49	17	13.81	24.53	4.88	1.39
5	15.91	20.83	5.35	1.56	18	15.58	27.65	5.02	1.66
6	12.47	22.27	4.27	1.50	19	15.85	27.29	5.55	1.70
7	15.80	27.57	5.25	1.85	20	15.28	29.07	5.26	1.82
8	14.32	28.01	4.62	1.51	21	16.40	32.47	5.18	1.75
9	13.76	24.79	4.42	1.46	22	15.02	29.65	5.08	1.70
10	15.18	28.96	5.30	1.66	23	15.73	22.11	4.90	1.81
11	14.20	25.77	4.87	1.64	24	14.75	22.43	4.65	1.82
12	17.07	23.17	5.80	1.90	25	14.35	20.04	5.08	1.53
13	15.40	28.57	5.22	1.66					

要求：

（1） 求 y 关于 x_1,x_2,x_3 的线性回归方程

$$y=c_0+c_1x_1+c_2x_2+c_3x_3+\varepsilon,$$

计算 c_0,c_1,c_2,c_3 的估计值；

（2） 对上述回归模型和回归系数进行检验（要写出相关的统计量）；

（3） 试建立 y 关于 x_1,x_2,x_3 的完全二项式回归模型。

解 (1) 记 y, x_1, x_2, x_3 的观察值分别为 $b_i, a_{i1}, a_{i2}, a_{i3}$，$i = 1,2,\cdots,25$，

$$X = \begin{pmatrix} 1 & a_{11} & a_{12} & a_{13} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & a_{25,1} & a_{25,2} & a_{25,3} \end{pmatrix}, \quad Y = \begin{pmatrix} b_1 \\ \vdots \\ b_{25} \end{pmatrix}.$$

用最小二乘法求 c_0, c_1, c_2, c_3 的估计值，即应选取估计值 c_j，使当 $c_j = c_j$，$j = 0,1,2,3$ 时，误差平方和

$$Q = \sum_{i=1}^{25} (b_i - \hat{b}_i)^2 = \sum_{i=1}^{25} (b_i - c_0 - c_1 a_{i1} - c_2 a_{i2} - c_3 a_{i3})^2$$

达到最小。为此，令

$$\frac{\partial Q}{\partial c_j} = 0，\quad j = 0,1,2,3$$

得到正规方程组，求解正规方程组得 c_0, c_1, c_2, c_3 的估计值

$$[c_0, c_1, c_2, c_3] = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y.$$

求得

$$c_0 = 0.8539，\ c_1 = 0.0178，\ c_2 = 2.0782，\ c_3 = 1.9396.$$

(2) 因变量 y 与自变量 x_1, x_2, x_3 之间是否存在线性关系是需要检验的，显然，如果所有的 $|c_j|\ (j = 1,2,3)$ 都很小，y 与 x_1, x_2, x_3 的线性关系就不明显，所以可令原假设为

$$H_0 : c_j = 0, \ j = 1,2,3. \qquad (9.6)$$

记 $m = 3$，$n = 25$，$Q = e^2 = \sum_{i=1}^{n} (b_i - b_i)^2$，$U = \sum_{i=1}^{n} (b_i - b)^2$，这里 $b_i = c_0 + c_1 a_{i1} + \cdots + c_m a_{im}$（$i = 1, \cdots, n$），$b = \frac{1}{n}\sum_{i=1}^{n} b_i$。当 H_0 成立时统计量

$$F = \frac{U/m}{Q/(n-m-1)} \sim F(m, n-m-1)，$$

在显著性水平 α 下，若

$$F_{1-\alpha}(m, n-m-1) < F < F_{\alpha}(m, n-m-1)，$$

接受 H_0；否则拒绝。

求得统计量 F 37.7453 ，查表得上 /2 分位数

$$= \begin{bmatrix} \vdots & \vdots & \vdots & \vdots \end{bmatrix} \qquad = \begin{bmatrix} \vdots \end{bmatrix}$$

$$= \qquad =$$

$$= \sum \varepsilon \quad = \sum \quad - \qquad = \sum \quad - \quad - \qquad - \qquad -$$

$$\frac{\partial}{\partial} = \qquad =$$

$$- \quad - \quad - \quad - \quad = \qquad -$$

$$- = \qquad - = \qquad - = \qquad - =$$

$$- \qquad =$$

$$= \qquad =$$

$$= \qquad = \qquad = \sum \quad = \sum \quad - \qquad = \sum \quad - \quad -$$

$$= \quad + \qquad + \cdots + \qquad = \cdots \qquad = - \sum$$

$$= \frac{\qquad\qquad\qquad}{- \quad -} \qquad - \quad -$$

$$-\alpha \qquad - \quad - \quad < \quad < \quad \alpha \qquad - \quad -$$

$$= \qquad\qquad\qquad\qquad \alpha$$

$$- \quad - \qquad\qquad -$$

$F_{0.025}(3,21) = 3.8188$，因而拒绝（9.6）式的原假设，模型整体上通过了检验。

当（9.6）式的 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

$$H_0^{(j)}: c_{jj}\beta_j = 0, j = 0,1,2,3 \qquad (9.7)$$

当 $H_0^{(j)}$ 成立时

$$t_j = \frac{\hat{\beta}_j / \sqrt{c_{jj}}}{\sqrt{Q/(n-m-1)}} \sim t(n-m-1)，$$

这里 c_{jj} 是 $(X^TX)^{-1}$ 中的第 (j,j) 个元素，对给定的 α，若 $|t_j| < t_{\frac{\alpha}{2}}(n-m-1)$，接受 $H_0^{(j)}$；否则拒绝。

求得统计量

$t_0 = 0.6223$，$t_1 = 0.6090$，$t_2 = 7.7407$，$t_3 = 3.8062$，

查表得上 $\alpha/2$ 分位数 $t_{0.025}(21) = 2.0796$。

对于9.7式的检验在显著性水平 $\alpha = 0.05$ 时接受 $H_0^{(j)}: c_{jj}\beta_j = 0$（$j = 0,1$），拒绝 $H_0^{(j)}: c_{jj}\beta_j = 0$（$j = 2,3$），即变量 x_1 对模型的影响是不显著的。建立线性模型时，可以不使用 x_1。

（3）求得的完全二次项模型为

$$y = -17.0988 + 0.3611x_1 + 2.3563x_2 + 18.2730x_3 - 0.1412x_1x_2 - 0.4404x_1x_3 - 1.2754x_2x_3 + 0.0217x_1^2 + 0.5025x_2^2 + 0.3962x_3^2.$$

【学生活动】自由讨论，阐述思路。

（五）课程思政（3 分钟）

结合人的年龄与身体内脂肪含量的关系的回归分析模型，引导学生养成好的生活习惯，健康饮食，锻炼身体。

（六）课堂总结（2 分钟）

总结本次课的教学内容：

1）：回归分析模型；

2）：回归分析模型的 Matlab 求解。

课后作业	本章的课后题 1、2。

教学内容	第二节 聚类分析
教学目标	**知识目标：** 理解常用的聚类分析方法；掌握用 Matlab 软件进行聚类分析的方法。 **能力目标：**通过实例演示、问题驱动等方式激发学生学习的积极性，通过观察对比、学生交流、师生交流、小组探究等多种形式， 培养学生的逻辑思维能力、分析判断能力、和解决问题能力。 **素质目标：** 培养学生敢于质疑、善于分析、勇于创新的精神；培养学生的自主学习意识和团队协作精神；培养学生脚踏实地、不畏艰辛、锲而不舍的精神。
授课学时	2 学时
教学重点	常用的聚类分析方法。
教学难点	用 Matlab 软件进行聚类分析
教学方法	采用问题驱动法、案例演示法、启发式讲授法及自主学习法相结合，以教师的讲解为主，学生的课堂报告、分组讨论为辅，充分调动学生学习的主动性和思考问题的积极性。
教学手段	以课堂讲授为主，主要是多媒体课件和板书相结合的形式。同时借助在线资源，如慕课、雨课堂等平台。
教学过程	**（一）采用讲授法，介绍常用的多种聚类分析方法的相关概念（约 45 分钟）** 系统聚类分析法是目前使用最多的一种方法，一般方法是： 设有 N 个样品，初始时这 N 个样品各自成一类，然后计算样品之间的距离，将距离最小的类并为一新类，再计算并类后的新类与其他类的距离，又将距离最小的两类并为一新类，这样每次减少一些类，直到将 N 个样品合并成一类为止。 知识点 1：最短距离法。

最短距离法将两样品间的距离定义为一个类中所有个体与另一类中的所有个体间距离的最小者。设有 N 个样品， d_{ij} 表示第 i 个样品与第 j 个样品之间的距离，用$G_1,G_2,\cdots,G_N$表示初始类，并类的原则是：类与类之间距离最近的两类合并。用D_{pq} 表示 G_p,G_q 之间的距离，规定

$$D_{pq}=\min_{\substack{i\in G_p\\ j\in G_q}}\{d_{ij}\},\ p\neq q$$

当 $p=q$ 时， $D_{pq}=0$

最短距离法就是以 D_{pq} 准则进行聚类，其聚类步骤为：

1） 规定样品之间的距离，计算 N 个样品中两两之间的距离 $d_{ij},i,j=1,2,\cdots,N$ ，得到对称矩阵 $D(0)=\left(d_{ij}\right)$ ，初始时每个样品自成一类，故 $D_{pq}=d_{pq}$ ；

2） 选择$D(0)$ 中最小非零元素，设为d_{pq} ，于是将G_p,G_q 并类，记为$G_r=\left\{G_p,G_q\right\}$；

3） 计算新类G_r与其他类$G_k(k\neq p,q)$ 的距离

$$D_{rk}=\min_{\substack{i\in G_r\\ j\in G_k}}\{d_{ij}\}=\min\{\min_{\substack{i\in G_p\\ j\in G_k}}\{d_{ij}\},\min_{\substack{i\in Gq\\ j\in G_k}}\{d_{ij}\}\}=\min\{D_{pk},D_{qk}\},$$

将$D(0)$ 中的第 p,q 行及第 p,q 列上的元素按照步骤 2合并成一个新类，记为G_r，对应于新行、新列得到的矩阵记为 $D(1)$ ；

4） 对$D(1)$ 重复上述 1）、2）的做法，得到 $D(2)$ ；

5） 继续下去，直到所有元素并为一类为止。

如果某一步 $D(k)$ 中最小的非零元素不唯一，对应于这些最小元素的类可以同时合并。

知识点 2：最长距离法。

最长距离法中两类合并的准则是类与类之间距离最长的两类合并，即

$$D_{pq} = \max_{\substack{i\in G_p \\ j\in G_q}} \{d_{ij}\}, p \neq q$$

最长距离法与最短距离法聚类步骤完全一样，只是距离准则不同。

设某一步将G_p, G_q合并为一类，记为G_r。则G_r与其他类G_k的距离公式

$$D_{rk} = \max_{\substack{i\in G_r \\ j\in G_k}} \{d_{ij}\} = \max\{\max_{\substack{i\in G_p \\ j\in G_k}} \{d_{ij}\}, \max_{\substack{i\in G_q \\ j\in G_k}} \{d_{ij}\}\} = \max\{D_{pk}, D_{qk}\},$$

再找距离最大的并类，直到所有元素并为一类为止。

最长距离法克服了最短距离法连接聚合的缺陷，但是当数据有较大的离散程度时，易产生较多类。与最短距离法一样，受异常值影响较大。

知识点 3：中间距离法。

如果类与类之间的距离既不采用最长距离法，也不采用最短距离法，而采用介于两者之间的距离，就称为中间距离法。

设某一步将G_p, G_q合并为一类，记为G_r。则G_r与其他类G_k的距离公式：

$$D_{kr}^2 = \frac{1}{2} D_{kp}^2 + \frac{1}{2} D_{kq}^2 + \beta D_{pq}^2, -\frac{1}{4} \leq \beta \leq 0,$$

其中，β常取为$-\frac{1}{4}$。

由于距离公式是平方的形式，故只需在第一步中记$D_{ij}^2(0) = (d^2)$，其余步骤不变。

知识点 4：质心法

设G_p, G_q的重心分别是$X_{(p)}$和$X_{(q)}$，则G_p, G_q之间的距离定义为

$$D_{pq} = (X_{(p)} - X_{(q)})^T (X_{(p)} - X_{(q)}) = d_2^2 (X_{(p)}, X_{(q)})$$

用这种距离进行聚类的方法，叫做质心法（或重心法）。

设某一步将G_p, G_q合并为一类，记为G_r, G_p, G_q的质心为$\bar{X}^{(p)}$和$\bar{X}^{(q)}$，若各类的样品个数分别为N_p, N_q，则G_r类的样品个数为$N_r = N_p + N_q$，质心$\bar{X}_r$为

$$\bar{X}^{(r)} = \frac{1}{N_r}\left(N_p \bar{X}^{(p)} + N_q \bar{X}^{(q)}\right)$$

则G_r与其他类G_k的距离公式（在此采用欧氏距离）：

$$D^2_{kr} = \frac{N_p}{N_r} D^2_{kp} + \frac{N_q}{N_r} D^2_{kq} - \frac{N_p N_q}{N_r^2} D^2_{pq}$$

知识点 5：类平均法。

一个类的重心，虽然有很好的代表性，但未能充分利用各样品的信息，于是提出了利用两类元素中两两之间的距离平方的平均值作为类与类之间的距离，即

$$D^2_{pq} = \frac{1}{N_p N_q}\sum_{i\in G_p}\sum_{j\in G_q} d^2_{ij}$$

利用这种距离的聚类法叫**类平均法**。

类平均法的聚类递推公式：

$$D^2_{kr} = \frac{1}{N_k N_r}\sum_{i\in G_k}\sum_{j\in G_r} d^2_{ij} = \frac{1}{N_r}\left(\sum_{i\in G_k}\sum_{j\in G_p} d^2_{ij} + \sum_{i\in G_k}\sum_{j\in G_q} d^2_{ij}\right) = \frac{N_p}{N_r} D^2_{kp} + \frac{N_q}{N_r} D^2_{kq}$$

知识点 6：离差平方和法。

该方法的基本思想来源于方差分析，如果类分得准确，同类样品的离差平方应当较小，而类与类之间的离差平方和应当较大，从而提出了离差平方和法。

设将N个样品分成k个类：$G_1, G_2, \cdots, G_k$，用$\bar{X}^{(i)}_{(i)}$表示G_i中第i个样品，N_t表G_t中的样品个数，$\bar{X}_{(t)}$表示G_t的质心，则G_t的样品离差平方和是

$$S_t = \sum_{i=1}^{N_t}\left(\bar{X}^{(i)}_{(i)} - \bar{X}_t\right)^T\left(\bar{X}^{(i)}_{(i)} - \bar{X}_t\right)$$

k 个类的离差平方和是

$$S = \sum_{t=1}^{k} \sum_{i=1}^{N_t} (X^{(i)} - \bar{X})^T (X^{(i)} - \bar{X}) \quad t \quad (i) \quad t$$

当k 固定时，要选择使得S 达到最小值的分类结果。具体做法是：先将 N 个样品各自分成一类，然后每次缩小一类，每缩小一类后的离差平方和就要增大，选择使S 增大最小的两类合并，直到所有样品归为一类为止。

离差平方和法的聚类递推公式：

$$D^2_{kr} = \frac{N_p + N_k}{N_r + N_k} D^2_{pq} + \frac{N_q + N_k}{N_r + N_k} D^2_{kp} - \frac{N_k}{N_r + N_k} D^2_{kq}$$

（二）采用讲授法，介绍 Matlab 软件与聚类分析相关的命令。（约 15 分钟）

（三）采用案例启发法，根据若干国家和地区的部分数据，对其进行分类（约 25 分钟）

根据信息基础设施的发展状况，对世界 20 个国家和地区进行分类。

表9.3　　20 个国家和地区信息基础设施数据

序号	country and region	call	movecall	fee	computer	mips	net
1	美 国	631.60	161.90	0.36	403.00	26073.00	35.34
2	日 本	498.40	143.20	3.57	176.00	10223.00	6.26
3	德 国	557.60	70.60	2.18	199.00	11571.00	9.48
4	瑞 典	684.10	281.80	1.40	286.00	16660.00	29.39
5	瑞 士	644.00	93.50	1.98	234.00	13621.00	22.68
6	丹 麦	620.30	248.60	2.56	296.00	17210.00	21.84
7	新加坡	498.40	147.50	2.50	284.00	13578.00	13.49
8	中国台湾	469.40	56.10	3.68	119.00	6911.00	1.72
9	韩 国	434.50	73.00	3.36	99.00	5795.00	1.68
10	巴 西	81.90	16.30	3.02	19.00	876.00	0.52
11	智 利	138.60	8.20	1.40	31.00	1411.00	1.28

12	墨西哥	92.20	9.80	2.61	31.00	1751.00	0.35
13	俄罗斯	174.90	5.00	5.12	24.00	1101.00	0.48
14	波 兰	169.00	6.50	3.68	40.00	1796.00	1.45
15	匈牙利	262.20	49.40	2.66	68.00	3067.00	3.09
16	马来西亚	195.50	88.40	4.19	53.00	2734.00	1.25
17	泰 国	78.60	27.80	4.95	22.00	1662.00	0.11
18	印 度	13.60	0.30	6.28	2.00	101.00	0.01
19	法 国	559.10	42.90	1.27	201.00	11702.00	4.76
20	英 国	521.10	122.50	0.98	248.00	14461.00	11.91

描述信息基础设施的变量主要有六个：(1)call——每千人拥有电话线数；(2)movecall——每千房居民蜂窝移动电话数；(3)fee——高峰时期每三分钟国际电话的成本；(4)computer——每千人拥有的计算机数；|(5)mips——每千人中计算机功率(每秒百万指令)；(6)net——每千人互联网络户主数。（数据摘自 1997 年《世界竞争力报告》）

解：采用欧氏距离和重心距离法进行分类，利用 Matlab 编程，可得聚类图如图 9.3 所示。

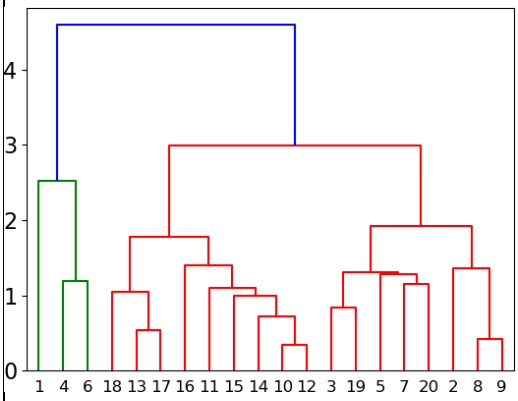


图9.3 聚类图

从聚类图看，结合实际情况分析采用重心距离法把 20 个国家和地区分为三类：

第I类：美国、瑞典、丹麦；

第II类：德国、法国、瑞士、新加坡、英国、日本、中国台湾、韩国；

第III类：印度、俄罗斯、泰国、马来西亚、智利、匈牙利、波兰、巴西、墨西哥。

	其中第I类为欧美发达国家，信息基础设施发展非常成熟，为最好，可以单独分成一类。第II类中的国家是德国、法国、日本等发达国家，新加坡、韩国等新兴工业化国家和地区中国台湾， 这几十年来发展迅速，努力赶超发达国家，在信息基础设施的发展上已非常接近发达国家。第III类中的国家为转型国家和亚洲、拉美等发展中国家，这些国家经济较不发达，基础设施薄弱，属于信息基础设施比较落后的国家。 **（四）课堂总结（5 分钟）** 总结本次课的教学内容： 1）：常见的聚类分析的方法； 2）：聚类分析的 Matlab 软件实现。
课后作业	课下巩固学习 Matlab 软件求解聚类分析问题。

教学内容	第三节 主成分分析
教学目标	**知识目标：** 理解主成分分析法的基本理论；掌握用 Matlab 软件进行主成分分析的方法。 **能力目标：**通过实例演示、问题驱动等方式激发学生学习的积极性，通过观察对比、学生交流、师生交流、小组探究等多种形式， 培养学生的逻辑思维能力、分析判断能力、和解决问题能力。 **素质目标：** 培养学生敢于质疑、善于分析、勇于创新的精神；培养学生的自主学习意识和团队协作精神；培养学生脚踏实地、不畏艰辛、锲而不舍的精神。
授课学时	2 学时
教学重点	主成分分析的基本理论。
教学难点	用 Matlab 软件进行主成分分析。
教学方法	采用问题驱动法、案例演示法、启发式讲授法及自主学习法相结合，以教师的讲解为主，学生的课堂报告、分组讨论为辅，充分调动学生学习的主动性和思考问题的积极性。
教学手段	以课堂讲授为主，主要是多媒体课件和板书相结合的形式。同时借助在线资源，如慕课、雨课堂等平台。
教学过程	**（一）采用讲授法，介绍主成分分析的基本理论（约 45 分钟）** 知识点 1：主成分分析法的思想。 将原来 p 个指标作线性组合，作为新的综合指标，但是这种线性组合，如果不加限制，则可以有很多。为了让这种综合指标反映足够多的原来的信息，要求综合指标的方差要大，即若Var(F_1) 越大，表示 F_1包含的信息越多，因此在所有线性组合中选取的 F_1 应该是方差最大的，故称 F_1为第一主成分。如果第一主成分不足以代表原来 p 个指标的信息，再考虑选取第二个

组合F_2，称F_2为第二主成分。为了有效地反映原来的信息，F_1中已有的信息就不需要出现在F_2中。数学表达式就是要求 Cov (F_1,F_2)=0。依此类推，可以构造出第三，第四，……，第p个主成分，这些主成分之间不仅不相关，而且它们的方差是依次递减的。在实际工作中，通常挑选前几个最大主成分，虽然可能会失去一小部分信息，但抓住了主要矛盾。

知识点 2：主成分分析法的原理与步骤。

设有p项指标$X_1,X_2,\cdots,X_p$，每个指标有n个观测数据，得到原始数据资料矩阵

$$X=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \triangleq (X_1,X_2,\cdots,X_p)$$

其中

$$X_i=\begin{bmatrix} x_{1i} \\ x_{2i} \\ \vdots \\ x_{ni} \end{bmatrix},\ i=1,2,\cdots,p$$

用矩阵x的p个向量$X_1,X_2,\cdots,X_p$作线性组合为：

$$\begin{cases} F_1=a_{11}X_1+a_{21}X_2+\cdots+a_{p1}X_p \\ F_2=a_{12}X_1+a_{22}X_2+\cdots+a_{p2}X_p \\ \cdots\cdots \\ F_p=a_{1p}X_1+a_{2p}X_2+\cdots+a_{pp}X_p \end{cases}$$

简写成

$$F_i=a_{1i}X_1+a_{2i}X_2+\cdots+a_{pi}X_p \quad (i=1,2,\cdots,p)$$

为了不使F_i的方差为无穷大，对上述方程组的系数要求 $a_{1i}^2+a_{2i}^2+\cdots+a_{pi}^2=1$（$i=1,2,\cdots,p$），且系数$a_{ij}$由下列原则决定：

（1）F_i与F_j（$i\neq j$）不相关；

（2）F_1是X_1，…，X_p的一切线性组合（系数满足上述方程组）中方差最大的一个，F_2是F_1不相关的X_1，…，X_p一切线

性组合中方差最大的一个，…，F_p 是与F_1，F_2，…，F_{p-1} 都不相关的 X_1，…，X_p 的一切线性组合中方差最大的一个。

定理　在上述条件下，a_{1i}，a_{2i}，…，a_{pi}（i=1,2, ⋯, p）是X的协方差矩阵的特征值对应的特征向量

设 $F_1 = a_1 X_1 + a_2 X_2 + \cdots + a_p X_p \triangleq a^T X$，其中，$a = (a_1, a_2, \ldots, a_p)^T$，$a_i = (a_{i1}, a_{i2}, \ldots, a_{ip})^T$，$X = (X_1, X_2, \ldots, X_p)^T$，记$\Sigma$为$X$的协方差矩阵，$F = (F_1, F_2, \ldots, F_p)^T$。

该定理表明 X_1，…，X_p 的主成分是以Σ的特征向量为系数的线性组合，它们互不相关且其方差为 Σ 的特征根，由于 Σ 的特征根 $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p > 0$，所以$\mathrm{Var}(F_1) \geq \mathrm{Var}(F_2) \geq \cdots \geq \mathrm{Var}(F_p) > 0$。

在解决实际问题时，一般不全取 p 个主成分，而是根据累计贡献率的大小取前k个。

定义　称 $\lambda_i \Big/ \sum_{i=1}^{p} \lambda_i$ 为第 i 个主成分的贡献率，称 $\sum_{i=1}^{k} \lambda_i \Big/ \sum_{i=1}^{p} \lambda_i$ 为前 k 个主成分的累计贡献率。

显然，贡献率越大，表明该成分综合的信息越多。

通过上述主成分分析的基本原理，归纳主成分分析计算步骤如下：

①　对原来的 p 个指标进行标准化，以消除变量在量纲上的影响。

②　根据标准化后的数据矩阵求出相关系数矩阵 $R_{p\times p} = (r_{ij})$，

$$R = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22} & \cdots & r_{2p} \\ \vdots & \vdots & & \vdots \\ r_{p1} & r_{p2} & \cdots & r_{pp} \end{bmatrix}$$

其中 r_{ij}（i, j=1,2, ⋯, p）为原变量的x_i与x_j之间的相关系数，其计

算公式为

$$r_{ij}=\frac{\sum_{k=1}^{n}\left(x_{ki}-\bar{x}_i\right)\left(x_{kj}-\bar{x}_j\right)}{\sqrt{\sum_{i=1}^{n}\left(x_{ki}-\bar{x}_i\right)\sum_{k=1}^{n}\left(x_{kj}-\bar{x}_j\right)}}$$

(3) 求出相关系数矩阵的特征根 λ_i（$i=1, 2, \ldots, p$）和对应的特征向量 e_i（$i=1,2,\cdots,p$），其中 e_{ij} 表示向量 e_i 的第 j 个分量。

(4) 计算主成分贡献率及累计贡献率，主成分 z_i 的贡献率为

$$\frac{\lambda_i}{\sum_{k=1}^{p}\lambda_k},\ (i=1,2,\cdots,p)$$

累计贡献率为

$$\frac{\sum_{k=1}^{i}\lambda_k}{\sum_{k=1}^{p}\lambda_k},\ (i=1,2,\cdots,p)$$

一般取累计贡献率达85%~95%的特征根 λ_1，λ_2，… λ_k 所对应的第一，第二，…，第 $k\ (k\le p)$ 个主成分。

(5) 计算主成分载荷，其计算公式为

$$a_{ij}=p\left(z_i,x_j\right)=\sqrt{\lambda_i}e_{ij}\ (i,j=1,2,\cdots,p)$$

得到各主成分的载荷矩阵 $A=\left(a_{ij}\right)$。

(6) 对主成分载荷归一化，对 $a_i=\left(a_{i1},a_{i2},\cdots,a_{ip}\right)$，$a_{i1}^2+a_{i2}^2+\cdots+a_{ip}^2\neq 1$，归一化得

$$a_{ik}^{*}=\frac{a_{ik}}{\sqrt{\sum_{k=1}^{p}a_{ik}^2}},\ i=1,2,\cdots.$$

(7) 写出主成分的表达式。

（二）采用讲授法，介绍 Matlab 软件的相关命令。（约 15 分钟）

（三）采用案例启发法，根据我国多年的宏观投资数据，进行主成分分析（约 25 分钟）

表9.4　　1984 年—2000 年宏观投资效益主要指标

年份	投资效果系数（无时滞）	投资效果系数（时滞一年）	全社会固定资产交付使用率	建设项目投产率	基建房屋竣工率
1984	0.71	0.49	0.41	0.51	0.46
1985	0.40	0.49	0.44	0.57	0.50
1986	0.55	0.56	0.48	0.53	0.49
1987	0.62	0.93	0.38	0.53	0.47
1988	0.45	0.42	0.41	0.54	0.47
1989	0.36	0.37	0.46	0.54	0.48
1990	0.55	0.68	0.42	0.54	0.46
1991	0.62	0.90	0.38	0.56	0.46
1992	0.61	0.99	0.33	0.57	0.43
1993	0.71	0.93	0.35	0.66	0.44
1994	0.59	0.69	0.36	0.57	0.48
1995	0.41	0.47	0.40	0.54	0.48
1996	0.26	0.29	0.43	0.57	0.48
1997	0.14	0.16	0.43	0.55	0.47
1998	0.12	0.13	0.45	0.59	0.54
1999	0.22	0.25	0.44	0.58	0.52
2000	0.71	0.49	0.41	0.51	0.46

解：用 $x_1, x_2, \cdots, x_5$ 分别表示投资效果系数（无时滞），投资效果系数（时滞一年），全社会固定资产交付使用率，建设项目投产率，基建房屋竣工率。用 $i = 1,2,\cdots,17$ 分别表示 1984 年，1985 年，…，2000 年，第 i 年第 j 个指标变量 x_j 的取值记作 a_{ij}，构造矩阵 $A = (a_{ij})_{17\times 5}$。

基于主成分分析法的评价和排序步骤如下。

（1）对原始数据进行标准化处理

将各指标值 a_{ij} 转换成标准化指标 $\tilde{a}_{ij}$，

$$\tilde{a}_{ij} = \frac{a_{ij} - \mu_j}{s_j}, \quad (i = 1,2,\cdots,17;\ j = 1,2,\cdots,5)$$

其中 $\mu_j = \frac{1}{17}\sum_{i=1}^{17} a_{ij}$，$s_j = \sqrt{\frac{1}{17}\sum_{i=1}^{17}(a_{ij} - \mu_j)^2}$，$(j = 1,2,\cdots,5)$，即 μ_j, s_j 为第 j 个指标的样本均值和样本标准差。对应地，称

$$\tilde{x}_j = \frac{x_j - \mu_j}{s_j}, \quad (j = 1,2,\cdots,5)$$

为标准化指标变量。

（2）计算相关系数矩阵R

相关系数矩阵 $R = (r_{ij})_{5\times 5}$

$$r_{ij} = \frac{\sum_{k=1}^{17} \tilde{a}_{ki} \cdot \tilde{a}_{kj}}{17-1}, \quad (i,j = 1,2,..,5)$$

式中 $r_{ii} = 1$，$r_{ij} = r_{ji}$，r_{ij} 是第 i 个指标与第 j 个指标的相关系数。

（3）计算特征值和特征向量

计算相关系数矩阵 R 的特征值 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_5 \geq 0$，及对应的标准化特征向量 $u_1, u_2, \ldots, u_5$，其中 $u_j = (u_{1j}, u_{2j}, .., u_{5j})^T$，由特征向量组成5个新的指标变量

$$\begin{cases} y_1 = u_{11}x_1 + u_{21}x_2 + \cdots + u_{51}x_5 \\ y_2 = u_{12}\tilde{x}_1 + u_{22}\tilde{x}_2 + \cdots + u_{52}\tilde{x}_5 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ y_5 = u_{15}\tilde{x}_1 + u_{25}\tilde{x}_2 + \cdots + u_{55}\tilde{x}_5 \end{cases}$$

式中 y_1 是第1主成分，y_2 是第2主成分，…，y_5 是第5主成分。

（4）选择 p（$p \leq 5$）个主成分，计算综合评价值

i）计算特征值 $\lambda_j (j = 1,2,\ldots,5)$ 的信息贡献率和累积贡献率。

称

$$b_j = \frac{\lambda_j}{\sum_{k=1}^{5} \lambda_k} \quad (j = 1,2;\cdots,5)$$

为主成分 y_j 的信息贡献率；

$$\alpha_p = \frac{\sum_{k=1}^{p} \lambda_k}{\sum_{k=1}^{5} \lambda}$$

为主成分 $y_1, y_2, \ldots, y_p$ 的累积贡献率，当 α_p 接近于1（$\alpha_p = 0.85, 0.90, 0.95$）时，则选择前 p 个指标变量 $y_1, y_2, .., y_p$ 作为 p 个主成分，代替原来5个指标变量，从而可对 p 个主成分进行综合分析。

ii）计算综合得分

$$Z = \sum_{j=1}^{p} b_j y_j$$

其中 b_j 为第 j 个主成分的信息贡献率，根据综合得分值就可进行评价。

利用Matlab软件求得相关系数矩阵的前5个特征根及其贡献率如表9.3。

表9.3　　主成分分析结果

序号	特征根	贡献率	累计贡献率
1	3.1343	62.6866	62.6866
2	1.1683	23.3670	86.0536
3	0.3502	7.0036	93.0572
4	0.2258	4.5162	97.5734
5	0.1213	2.4266	100.0000

可以看出，前三个特征根的累计贡献率就达到 93%以上，主成分分析效果很好。下面选取前三个主成分进行综合评价。前三个特征根对应的特征向量见表 9.4。

表9.4　标准化变量的前4个主成分对应的特征向量

	$\tilde{x}_1$	$\tilde{x}_2$	$\tilde{x}_3$	$\tilde{x}_4$	$\tilde{x}_5$
第1特征向量	0.490542	0.525351	-0.48706	0.067054	-0.49158
第2特征向量	-0.29344	0.048988	-0.2812	0.898117	0.160648
第3特征向量	0.510897	0.43366	0.371351	0.147658	0.625475

由此可得三个主成分分别为

$$y_1 = 0.491\tilde{x}_1 + 0.525\tilde{x}_2 - 0.487\tilde{x}_3 + 0.067\tilde{x}_5 - 0.492\tilde{x}_5$$

$$y_2 = -0.293\tilde{x}_1 + 0.049\tilde{x}_2 - 0.281\tilde{x}_3 + 0.898\tilde{x}_4 + 0.161\tilde{x}_5$$

$$y_3 = 0.511\tilde{x}_1 + 0.434\tilde{x}_2 + 0.371\tilde{x}_3 + 0.148\tilde{x}_4 + 0.625\tilde{x}_5$$

分别以三个主成分的贡献率为权重，构建主成分综合评价模型：

$$Z = 0.6269y_1 + 0.2337y_2 + 0.076y_3$$

把各年度的三个主成分值代入上式，可以得到各年度的综合评价值以及排序结果

（四）课堂总结（5 分钟）

总结本次课的教学内容：

（1）主成分分析的原理与步骤；

	（2）主成分分析的 Matlab 软件实现。
课后作业	课下巩固学习 Matlab 软件求解主成分分析问题。