# 应用统计学上机

经济管理学院, 北京航空航天大学

June 5, 2019

1、回归分析

2、聚类分析

3、主成分分析

4、判别分析

# 相关系数（The Correlation Coefficient）

**Pearson相关系数**适用于两个连续变量间呈线性相关；

cor(x, y, method = “pearson")

**Spearman秩相关系数**适合定序变量、非线性等情形；

cor(x, y, method = "spearman")

**Kendall 相关系数**适用于分析两个定序变量的协同性（例如两个专家对n个事物的评价，可能他们的评分松紧不同，但关键是评价排序是否一致）

cor(x, y, method = "kendall")

# 10.01 企业产量与生产费用

```
###读入数据
data <- read.csv("E:/20190605/10EX1.csv", header = T)
x <- data$产量.万台.
y <- data$生产费用.万元.
###计算pearson相关系数
cor(x, y, method = "pearson")
cor.test(x, y, method = "pearson")

###计算spearman相关系数
cor(x, y, method = "spearman")
cor.test(x, y, method = "spearman")

###计算kendall相关系数
cor(x, y, method = "kendall")
cor.test(x, y, method = "kendall")
```

回归分析是对客观事物数量依存关系的分析, 主要问题包括

- 确定$Y$与$X_1, \cdots, X_p$之间的定量关系表达式, 即回归方程;
(估计问题)
- 对求得的回归方程的可信度进行检验;
(模型显著性检验——F检验)
- 判断自变量$X_j$对$Y$有无影响;
(变量的显著性检验——t检验)
- 利用所求得的回归方程进行预测.
(预测)

# 线性回归 ( Linear Regression Model )

Model $y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_p X_{pi} + \varepsilon_i$, where

- $y = (y_1, \cdots, y_n)^T$ is $n \times 1$ response vector;
- $X = (1, X_1, \cdots, X_p)$ is $n \times (p+1)$ design matrix with the first column is $(1, \cdots, 1)^T$;
- $\beta = (\beta_0, \beta_1, \cdots, \beta_p)^T$ is $(p+1) \times 1$ regression parameter vector;
- $\varepsilon = (\varepsilon_1, \cdots, \varepsilon_n)^T$ is the random error term with $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$.

Eq. (1.1) matrix form

$$y = X\beta + \varepsilon.$$

###lm()函数、summary()函数

# 1、估计标准误（Standard Error of the Estimate）

估计标准误：残差的标准差

$$\bar{e} = \frac{1}{n}\sum_{i=1}^{n}\hat{e}_i = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) = \frac{1}{n}\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i) \overset{Normal\ Equation}{=} 0$$

$$s_e = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(\hat{e}_i - 0)^2} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} = \sqrt{\frac{SSE}{n-2}}$$

$$\sum_{i=1}^{n}\hat{e}_i = 0$$

$$\sum_{i=1}^{n}x_i\hat{e}_i = 0$$

自由度（Degree of Freedom ）= n-2

比较不同的模型时，估计标准误差越小的模型，精度越高。

# 2、拟合优度（Goodness of Fit）

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

$$SST = SSR + SSE$$

**自由度： $(n-1)$ $1$ $(n-2)$

**SSR越大**：用回归方程解释 $y_i$ 变异的部分越多

**SSE越小**：观测值 $y_i$ 绕回归线越紧密，拟合越好

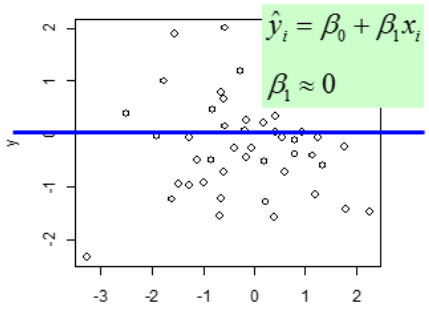
**测定系数** (Coefficient of Determination )

$$R^2=\frac{SSR}{SST}=1-\frac{SSE}{SST}$$

# 3、F 检验

(1) $H_0: \beta_1 = 0, \qquad H_1: \beta_1 \neq 0$

(2) 检验统计量




$$F = \frac{SSR/1}{SSE/n-2} \overset{H_0}{\underset{\beta_1=0}{\sim}} F(1, n-2)$$

$$SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2, \qquad SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

(3) $\alpha = 0.05 \Rightarrow P\{F > F_\alpha(1, n-2)\} = \alpha$

(4) 若 $F > F_\alpha$，拒绝$\mathbf{H_0}$ **(不能否定线性模型)**

若$F \leq F_\alpha$，不拒绝$\mathbf{H_0}$ (非线性模型或换变量)

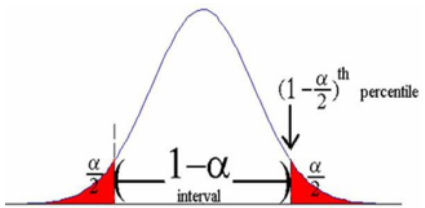
# 4、t 检验

(1) $H_0: \beta_1 = 0, \qquad H_1: \beta_1 \neq 0$

(2)检验统计量:

$$t = \frac{b_1}{s_{b_1}} \overset{H_0}{\underset{\beta_1=0}{\sim}} t(n-2)$$

$$s_{b_1} = \frac{s_e}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$




(3) $\alpha = 0.05 \Rightarrow P\{|t| > t_{\alpha/2}\} = \alpha$

(4) 若 $|t| > t_{\alpha/2}$, 拒绝 $H_0$ (**$X$ 有解释作用**)

若 $|t| \leq t_{\alpha/2}$, 不拒绝 $H_0$ ($X$ 没有解释作用)

# 5、置信区间

斜率的置信区间为：$(b_1 - t_{\alpha/2} s_{b_1},\ \ b_1 + t_{\alpha/2} s_{b_1})$

截距项的置信区间为：$(b_0 - t_{\alpha/2} s_{b_0},\ \ b_0 + t_{\alpha/2} s_{b_0})$

###confint()函数：提供模型参数的置信区间（默认95%）

# 6、预测

当X=xp时，预测对应的yp：

点估计（Point Estimation）：

$$\hat{y}_p = b_0 + b_1 x_p$$

区间估计（Interval Estimation）：

$$\hat{y}_p \pm t_{\alpha/2} s_e \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

### predict()函数

# 回归诊断（Regression diagnostics）

前面得到的结果，均是基于模型假设正确的基础之上，需要对回归结果进行诊断，包括检查模型、样本点和变量。

——残差分析：是否满足独立性、正态性、同方差性

——异常值检验：是否存在异常样本点

——多重共线性检验：自变量是否存在多重共线性

# 7、残差分析

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad H_0: \ \varepsilon_i \sim N(0, \sigma^2)$$

残差: $\hat{e}_i = (y_i - \hat{y}_i)$ 用 $\hat{e}_i$ 估计 $\varepsilon_i$

$$\bar{e} = \frac{1}{n}\sum_{i=1}^{n} \hat{e}_i = 0, \qquad \mathrm{Var}(\hat{e}_i) = \frac{1}{n-2}\sum_{i=1}^{n} (\hat{e}_i - 0)^2 = s_e^2$$

定义 "标准化残差": $e_i^* = \dfrac{\hat{e}_i - 0}{s_e} \underset{n\to\infty}{\sim} N(0,1)$

在正态假设下，标准化残差服从标准正态分布。

残差图（标准化残差图）：

以（标准化）残差为纵坐标，以自变量、拟合值或对应的观测样本序号i为横坐标的散点图。

# 7、残差分析

残差的独立性检验

—— 最好的方法是依据收集数据方式的先验知识

—— durbinWatsonTest检验函数（car package）也可以检测误差的独立性（序列相关性），p值不显著说明无序列相关性。

# 7、残差分析

残差的正态性检验

—— 标准化残差图中应该有95%的数据点落在区间[-2,2]，且不呈现出任何趋势。$P\left\{\left|e_i^*\right|<2\right\}=0.9545$

—— 绘制残差的直方图

—— 正态性检验shapiro.test()函数

（p值不显著表示样本数据符合正态分布）

—— Q-Q plot（分位数，在对角线附近）

# 7、残差分析

## 残差的方差齐性检验

当残差的绝对值随着观测值的增大有明显增加或减少，或先增加后减少的趋势时，表示关于模型中关于同方差的假设不成立。

——Scale-Location（位置尺度图）水平线周围的点随机分布

——ncvTest()函数：零假设为同方差，备择假设为误差方差随着拟合值水平的变化而变化

——powerTransform()：输出建议幂次变换（suggested power transformation）

含义是：经过p次幂变换，非恒定的方差将会平稳。

常用的方差稳定性变换有：开方变换、对数变换、倒数变换

解决异方差问题：分位数回归

# 8、异常值检验

离群点：对于给定的自变量值$x_i$来说，因变量值$y_i$异常的点

标准化残差绝对值大于3

高杠杆率点：自变量观测值$x_i$是异常的

通过帽子统计量（hat statistic）判断

强影响点：对模型估计值影响较大的点

Cook距离(Cook's D)，它综合反映了杠杆值和残差大小

删除异常点要谨慎。并非所有的异常点都意味着结果不好，有时候发现异常点可能会提示有更重要的信息。如果出现异常点，首先应检查数据是否录入错误，也可以选择其他相应模型来拟合，或者需要收集更多的数据来证实。

### influencePlot()函数（car package）

# 9、多重共线性检验

自变量彼此相关时，回归结果可能会出现：

——当增加或剔除一个变量，或改变一个观测值时，回归系数估计值会发生很大改变

——$R^2$很高，但一些重要的自变量没有通过显著性检验

——某些自变量回归系数的正负与定性结果分析不一致

……

VIF（variance inflation factor, 方差膨胀因子）：###vif()函数

——大于10有较强的共线性

相关矩阵的条件数kappa： ### kappa()函数

——小于100时多重共线程度很小；100到1000时中等或较强多重共线性；

大于1000严重的多重共线性。

# 线性回归一般步骤

step0：确定因变量和自变量

step1：对自变量进行多重共线性检验

step2：线性回归，找到最优回归方程

step3：回归诊断（残差分析、异常值检验）

step4：关于模型的假设正确时，方可基于结果进行解释和预测

# 逐步回归（Stepwise）

在实际问题中，影响因变量的因素有很多，但过多的自变量使用起来并不方便，回归方程中对因变量影响不大的自变量的存在，使得模型自由度减少，从而使得标准差的估计增大，从而影响预测精度。最优的模型应该包含尽可能少的变量，这就涉及到变量选择（Variable selection）。

向前选择法 （Forword）

向后筛选法 （Backward）

逐步回归法 （Stepwise）：边进边退

### step()函数

# Eg. 1-1

Table: 合金的强度与合金中碳含量数据表

| 碳含量(X) | 0.1 | 0.11 | 0.12 | 0.13 | 0.14 | 0.15 | 0.16 | 0.17 | 0.18 | 0.2 | 0.21 | 0.23 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 强度(Y) | 42.0 | 43.50 | 45.00 | 45.50 | 45.00 | 47.50 | 49.00 | 53.00 | 50.00 | 55.0 | 55.00 | 60.00 |

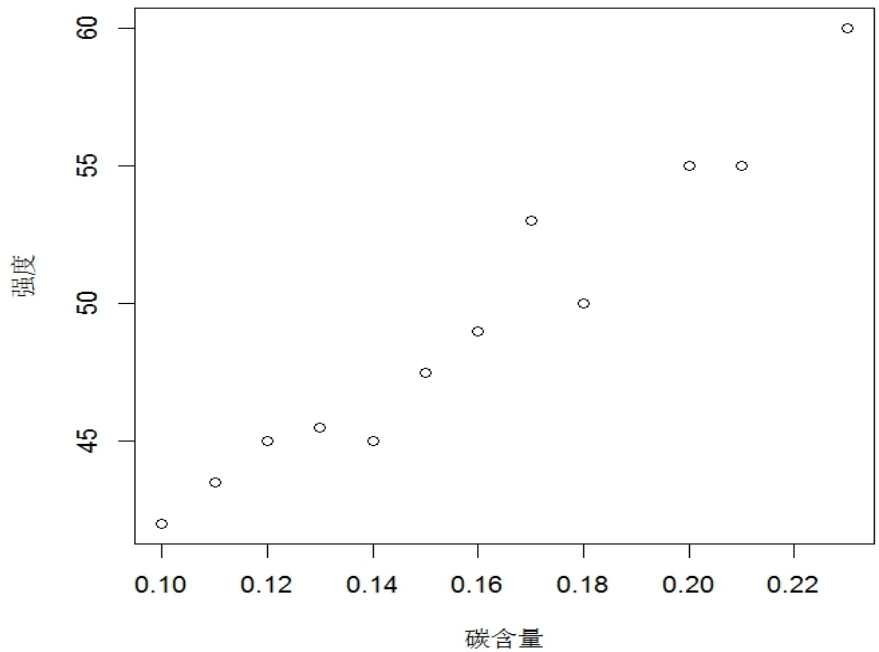



Figure: 数据的散点图

```
###input data
x<-c(0.10,0.11,0.12,0.13,0.14,0.15,0.16,0.17,0.18,
     0.20,0.21,0.23)
y<-c(42.0,43.5,45.0,45.5,45.0,47.5,49.0,53.0,50.0,
     55.0,55.0,60.0)
###Scatter plot
plot(x,y,xlab='碳含量', ylab='强度')###散点图
##12个点基本上在一条直线上，从而可以认为两者的关系基本上是线
性的
###Linear regression
lm.sol<-lm(y ~ 1+x)####y=beta0+beta1*x+error
summary(lm.sol)####提取模型的计算结果
###相应回归模型公式
Call:
lm(formula = y ~ 1 + x)
###列出残差的五数(最小, 25%,50%,75%,最大) lm.sol$resid
Residuals:
    Min      1Q  Median      3Q     Max
-2.0431 -0.7056  0.1694  0.6633  2.2653
```

```
###回归系数 lm.sol$coef
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   28.493      1.580   18.04 5.88e-09 ***
x            130.835      9.683   13.51 9.50e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
####残差的标准差, R-squared, F-test
Residual standard error: 1.319 on 10 degrees of freedom
Multiple R-squared:  0.9481,    Adjusted R-squared:  0.9429
F-statistic: 182.6 on 1 and 10 DF,  p-value: 9.505e-08
```

```
### 求线性模型系数的区间估计
confint(lm.sol)
### 系数的置信区间
                2.5 %       97.5%
(Intercept)  24.97279   32.01285
 x           109.25892  152.41074
### 预测
new<-data.frame(x=0.16)
### 必须采用数据框的形式
lm.pred<-predict(lm.sol,new,interval='prediction',level=0.95)

###predict(lm.sol)等价于lm.sol$fitted.values
###interval='prediction'表示给出相应的预测区间
###level表示相应的置信度
lm.pred
        fit       lwr      upr
1  49.42639  46.36621  52.48657
```

# Your turn (I)???

Table: 血压收缩压(Y)与体重($X_1$),年龄($X_2$) 数据

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $X_1$ | 76 | 91.5 | 85.5 | 82.5 | 79 | 80.5 | 74.5 | 79 | 85 | 76.5 | 82 | 95 | 92.5 |
| $X_2$ | 50 | 20.0 | 20.0 | 30.0 | 30 | 50.0 | 60.0 | 50 | 40 | 55.0 | 40 | 40 | 20.0 |
| Y | 120 | 141.0 | 124.0 | 126.0 | 117 | 125.0 | 123.0 | 125 | 132 | 123.0 | 132 | 155 | 147.0 |

- 建立$Y$与$X_1, X_2$之间的线性回归方程;并画出拟合曲线
- 画出残值与拟合值之间的散点图;
- 画出标准化残差与拟合值之间的散点图;
- 求参数$\beta$的置信区间($\alpha = 0.05$);
- 求$X = x_0 = (80, 40)^T$时相应$Y$的置信度为0.95的预测区间.

See code in Ex1.R.

# Eg 1-2: Forbes 数据

Table: 阿尔卑斯山及苏格兰的17个地点沸点及大气压的Forbes数据

| No | 沸点F | 气压h | ($\log_{10}$ 气压)log | ($100\times\log_{10}$ 气压)log100 |
|---|---|---|---|---|
| 1 | 194.5 | 20.79 | 1.3179 | 131.79 |
| 2 | 194.3 | 20.79 | 1.3179 | 131.79 |
| 3 | 197.9 | 22.40 | 1.3502 | 135.02 |
| 4 | 198.4 | 22.67 | 1.3555 | 135.55 |
| 5 | 199.4 | 23.15 | 1.3646 | 136.46 |
| 6 | 199.9 | 23.35 | 1.3683 | 136.83 |
| 7 | 200.9 | 23.89 | 1.3782 | 137.82 |
| 8 | 201.1 | 23.99 | 1.3800 | 138.00 |
| 9 | 201.4 | 24.02 | 1.3806 | 138.06 |
| 10 | 201.3 | 24.01 | 1.3805 | 138.05 |
| 11 | 203.6 | 25.14 | 1.4004 | 140.04 |
| 12 | 204.6 | 26.57 | 1.4244 | 142.44 |
| 13 | 209.5 | 28.49 | 1.4547 | 145.47 |
| 14 | 208.6 | 27.76 | 1.4434 | 144.34 |
| 15 | 210.7 | 29.04 | 1.4630 | 146.30 |
| 16 | 211.9 | 29.88 | 1.4754 | 147.54 |
| 17 | 212.2 | 30.06 | 1.4780 | 147.80 |

Remark: Forbes 的理论认为: 在观测值范围内, 沸点与气压值的对数成一条直线. (数据forbes.txt)

```
#Read data
read.table('forbes.txt',header=T)
##散点图
plot(forbes$F, forbes$log100)
##线性回归
lm.sol<-lm(log100~F, data=forbes)
summary(lm.sol)
abline(lm.sol)
##分析残差
y.res<-residuals(lm.sol);plot(y.res)
text(12,y.res[12], labels=12,adj=1.2)
##第12个样本点可能有问题（回归诊断）
##这里做简单的处理，去掉第12个样本点
lm12<-lm(log100~F,data=forbes,subset=-12)
summary(lm12)
```

```
##Forbes 数据所得到回归模型中的残差做正态性检验
forbes<-read.table('forbes.txt',header=T)
lm.sol<-lm(log100~F,data=forbes)
y.res<-resid(lm.sol)
shapiro.test(y.res)
        Shapiro-Wilk normality test

data:  y.res
W = 0.54654, p-value = 3.302e-06
##残差不满足正态性假设
shapiro.test(y.res)
lm12.sol<-lm(log100~F,data=forbes,subset=-12)
y12.res<-resid(lm12.sol)
shapiro.test(y12.res)
      Shapiro-Wilk normality test

data:  y12.res
W = 0.92215, p-value = 0.1827
##去掉第12个样本点后，残差通过正态性检验
```

# Your turn (II)???

Table: 某地区家庭人均收入与人均购买量数据

| X | Y | X | Y | X | Y |
|---|---|---|---|---|---|
| 679 | 0.79 | 745 | 0.77 | 770 | 1.74 |
| 292 | 0.44 | 435 | 1.39 | 724 | 4.10 |
| 1012 | 0.56 | 540 | 0.56 | 808 | 3.94 |
| 493 | 0.79 | 874 | 1.56 | 790 | 0.96 |
| 582 | 2.70 | 1543 | 5.28 | 783 | 3.29 |
| 1156 | 3.64 | 1029 | 0.64 | 406 | 0.44 |
| 997 | 4.73 | 710 | 4.00 | 1242 | 3.24 |
| 2189 | 9.50 | 1434 | 0.31 | 658 | 2.14 |
| 1097 | 5.34 | 837 | 4.20 | 1746 | 5.71 |
| 2078 | 6.85 | 1748 | 4.88 | 468 | 0.64 |
| 1818 | 5.84 | 1381 | 3.48 | 1114 | 1.90 |
| 1700 | 5.21 | 1428 | 7.58 | 413 | 0.51 |
| 747 | 3.25 | 1255 | 2.63 | 1787 | 8.33 |
| 2030 | 4.43 | 1777 | 4.99 | 3560 | 14.94 |
| 1643 | 3.16 | 370 | 0.59 | 1495 | 5.11 |
| 414 | 0.50 | 2316 | 8.19 | 2221 | 3.85 |
| 354 | 0.17 | 1130 | 4.79 | 1526 | 3.93 |
| 1276 | 1.88 | 463 | 0.51 | 0 | 0.00 |

See data in ex2.txt.

表8给出某地区家庭收入与人均购买量数据，试建立两者之间的关系

- Q1. 给出线性回归结果
- Q2. 画出标准化残差散点图，并给出标准化残差的正态性检验
- Q3. 画出标准化残差的QQ图
- Q4. 观察标准化残差图，并对线性回归方程中的响应变量做开方运算，更新模型；(非齐次方差的修正方法之一)
- Q5. 给出更新后回归模型的标准化残差图

See R code in Ex2.R.

# Eg 1-3

对下表中的数据进行多重共线性诊断，并找出哪些变量是多重共线性的。

Table: 原始数据

| Y | X1 | X2 | X3 | X4 | X5 | X6 |
|---|---|---|---|---|---|---|
| 10.006 | 8 | 1 | 1 | 1 | 0.541 | −0.099 |
| 9.737 | 8 | 1 | 1 | 0 | 0.130 | 0.070 |
| 15.087 | 8 | 1 | 1 | 0 | 2.116 | 0.115 |
| 8.422 | 0 | 0 | 9 | 1 | −2.397 | 0.252 |
| 8.625 | 0 | 0 | 9 | 1 | −0.046 | 0.017 |
| 16.289 | 0 | 0 | 9 | 1 | 0.365 | 1.504 |
| 5.958 | 2 | 7 | 0 | 1 | 1.996 | −0.865 |
| 9.313 | 2 | 7 | 0 | 1 | 0.228 | −0.055 |
| 12.960 | 2 | 7 | 0 | 1 | 1.380 | 0.502 |
| 5.541 | 0 | 0 | 0 | 10 | −0.798 | −0.399 |
| 8.756 | 0 | 0 | 0 | 10 | 0.257 | 0.101 |
| 10.937 | 0 | 0 | 0 | 10 | 0.440 | 0.432 |

```
collinear<-data.frame(
   Y=c(10.006, 9.737, 15.087, 8.422, 8.625, 16.289,
        5.958, 9.313, 12.960, 5.541, 8.756, 10.937),
   X1=rep(c(8, 0, 2, 0), c(3, 3, 3, 3)),X2=rep(c(1,
   0, 7, 0), c(3, 3, 3, 3)),X3=rep(c(1, 9, 0), c(3,
   3, 6)),
   X4=rep(c(1, 0, 1, 10), c(1, 2, 6, 3)),
   X5=c(0.541, 0.130, 2.116, -2.397, -0.046, 0.365,
           1.996, 0.228, 1.38, -0.798, 0.257, 0.440),
     X6=c(-0.099, 0.070, 0.115, 0.252, 0.017, 1.504,
           -0.865, -0.055, 0.502, -0.399, 0.101, 0.432))
  XX<-cor(collinear[2:7])

kappa(XX,exact=TRUE)
##2195.908
eigen(XX)
###min(eigenvalue)=0.001106
###eigen vector 0.4476,0.4211,0.5417,0.5734,0.006052,0.002167
###The last two nearly zero, i.e,
###X1,X2,X3,X4 存在多重共线性
```

# Eg. 1-4

下面通过一个例子(某种水泥凝固时所散发出的热量与四种化学成分之间的关系)来介绍R 软件中完成逐步回归的过程.

Table: 数据表

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $X_1$ | 7.0 | 1.0 | 11.0 | 11.0 | 7.0 | 11.0 | 3.0 | 1.0 | 2.0 | 21.0 | 1.0 | 11.0 | 10.0 |
| $X_2$ | 26.0 | 29.0 | 56.0 | 31.0 | 52.0 | 55.0 | 71.0 | 31.0 | 54.0 | 47.0 | 40.0 | 66.0 | 68.0 |
| $X_3$ | 6.0 | 15.0 | 8.0 | 8.0 | 6.0 | 9.0 | 17.0 | 22.0 | 18.0 | 4.0 | 23.0 | 9.0 | 8.0 |
| $X_4$ | 60.0 | 52.0 | 20.0 | 47.0 | 33.0 | 22.0 | 6.0 | 44.0 | 22.0 | 26.0 | 34.0 | 12.0 | 12.0 |
| 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 |

希望从中选出主要的变量，建立$y$关于他们的线性回归方程.

```
##输入数据
cement<-data.frame(
   X1=c( 7,  1, 11, 11,  7, 11,  3,  1,  2, 21,  1, 11, 10),
   X2=c(26, 29, 56, 31, 52, 55, 71, 31, 54, 47, 40, 66, 68),
   X3=c( 6, 15,  8,  8,  6,  9, 17, 22, 18,  4, 23,  9,  8),
   X4=c(60, 52, 20, 47, 33, 22,  6, 44, 22, 26, 34, 12, 12),
   Y =c(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))
####首先做多元线性回归方程
lm.sol<-lm(Y ~ X1+X2+X3+X4, data=cement)
summary(lm.sol)

Call:
lm(formula = Y ~ X1 + X2 + X3 + X4, data = cement)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1750 -1.6709  0.2508  1.3783  3.9254

Coefficients:
```

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  62.4054    70.0710   0.891   0.3991
X1            1.5511     0.7448   2.083   0.0708 .
X2            0.5102     0.7238   0.705   0.5009
X3            0.1019     0.7547   0.135   0.8959
X4           -0.1441     0.7091  -0.203   0.8441
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
Residual standard error: 2.446 on 8 degrees of freedom
Multiple R-squared:  0.9824,    Adjusted R-squared:  0.9736
F-statistic: 111.5 on 4 and 8 DF,  p-value: 4.756e-07
###如果所有变量做回归方程，系数没有通过检验；
###但是R-squared特别高（这是多重共线性表现现象之一；回归诊断
中会提及）
###下面通过step()函数做逐步回归
```

```
lm.step<-step9lm.sol)
lm.step
Start:  AIC=26.94####全部变量时的AIC值
Y ~ X1 + X2 + X3 + X4
       Df Sum of Sq    RSS    AIC
- X3    1    0.1091 47.973 24.974###去掉X3
- X4    1    0.2470 48.111 25.011###去掉X4
- X2    1    2.9725 50.836 25.728###去掉X2
<none>               47.864 26.944###全部去掉
- X1    1   25.9509 73.815 30.576###去掉X1

Step:  AIC=24.97
Y ~ X1 + X2 + X4
       Df Sum of Sq    RSS    AIC
<none>               47.97 24.974
- X4    1      9.93  57.90 25.420
- X2    1     26.79  74.76 28.742
- X1    1    820.91 868.88 60.629
###无论去掉哪一个变量，AIC取值均为增加，运算停止
```

```
summary(lm.step)
Call:
lm(formula = Y ~ X1 + X2 + X4, data = cement)

Residuals:
    Min      1Q  Median      3Q     Max
-3.0919 -1.8016  0.2562  1.2818  3.8982
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  71.6483    14.1424   5.066 0.000675 ***
X1            1.4519     0.1170  12.410 5.78e-07 ***
X2            0.4161     0.1856   2.242 0.051687 .
X4           -0.2365     0.1733  -1.365 0.205395
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
Residual standard error: 2.309 on 9 degrees of freedom
Multiple R-squared:  0.9823,    Adjusted R-squared:  0.9764
F-statistic: 166.8 on 3 and 9 DF,  p-value: 3.323e-08
###回归系数的显著性水平显著提高，但X2和X4系数的显著性仍然不理
想
```

```
##另外两个做逐步回归的函数 add1()和drop1()
drop1(lm.step)
Single term deletions

Model:
Y ~ X1 + X2 + X4
       Df Sum of Sq    RSS    AIC
<none>                47.97 24.974
X1      1    820.91 868.88 60.629
X2      1     26.79  74.76 28.742
X4      1      9.93  57.90 25.420
###去掉X4，AIC上升最少，RSS上升也最少，综合两个指标，下面去
掉X4
lm.opt<-lm(Y ~ X1+X2, data=cement); summary(lm.opt)
Call:
lm(formula = Y ~ X1 + X2, data = cement)
Residuals:
   Min     1Q Median     3Q    Max
-2.893 -1.574 -1.302  1.363  4.048
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 52.57735    2.28617   23.00 5.46e-10 ***
X1           1.46831    0.12130   12.11 2.69e-07 ***
X2           0.66225    0.04585   14.44 5.03e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
Residual standard error: 2.406 on 10 degrees of freedom
Multiple R-squared:  0.9787,    Adjusted R-squared:  0.9744
F-statistic: 229.5 on 2 and 10 DF,  p-value: 4.407e-09
```

对Eg1-4中的水泥数据进行多重共线性诊断；说明step函数中去掉的变量是否合理？

```
##cement data
cement<-read.table('cement.txt',header=T) XX<-
cor(cement[,1:4])
kappa(XX)
#2158.818
###存在严重多重共线性 #XX
```

中X1与X3；X2与X4的相关系数非常高

# 聚类分析(Cluster analysis)

聚类分析将数据所对应的研究对象进行分类。

- 事先不知道类别的个数与结构
- 进行分析的数据是对象之间的相似性或相异性的数据
- 聚类分析的共同思路：将这些相似（相异）性数据看成对象之间的“距离"远近的一种度量，将距离近的对象归为一类，不同类之间的对象距离较远。
- 根据分类对象的不同
  - Q型聚类分析–对样本进行聚类
  - R型聚类分析–对变量进行聚类分析

# 常用距离的定义

- euclidean: 欧几里得距离 $\sqrt{\{\sum_{k=1}^{p}(x_{ik}-x_{jk})^2\}}$
- maximum: 切比雪夫距离 $\max_{1\le k\le p}|x_{ik}-x_{jk}|$
- manhattan: 绝对值距离 $\sum_{k=1}^{p}|x_{ik}-x_{jk}|$
- canberra : Lance距离 $\sum_{k=1}^{p}\frac{|x_{ik}-x_{jk}|}{|x_{ik}+x_{jk}|}$
- minkowski: 闵可夫斯基距离 $\{\sum_{k=1}^{p}(x_{ik}-x_{jk})^q\}^{1/q}$ (参数p)
- binary:定性变量的距离 $\frac{m_2}{m_1+m_2}$ ($m_1$ 1-1配对的总数, $m_0$ 0-0配对的总数； $m_2$ 不配对的总数)

- R软件中给出计算各种距离的函数，dist(), i.e., dist(x, method='euclidean',diag=FALSE,upper=FALSE,p=2)
- 聚类分析前需要数据中心化或标准化scale(x,center=T, scale=T)
- 相似系数$c_{ij}$
  - 夹角余弦

$$\frac{\sum_{k=1}^{p} x_{ki}x_{kj}}{\sqrt{\sum_{k=1}^{p} x_{ki}^2}\sqrt{\sum_{k=1}^{p} x_{kj}^2}}$$

  - 相关系数cor(x)
- 变量之间利用相似系数来定义距离：　$d_{ij} = 1 - c_{ij}$

# 系统聚类法

- 聚类分析步骤: 设有$n$个样品，$p$个变量
  - 先将每个个体看成一类，共r类(Q型聚类，$r=n$; R型聚类，$r=p$);
  - 找出最相似的两类，合并成一个新类，得$r-1$类;
  - 在$r-1$类中，再找出最相似的两类合并，得$r-2$类;
  - 以此类推，将所有的样本合并成一大类.
- 最短距离法(single)，最长距离法(complete)，中间距离法(median)，相似法(mcquitty)，类平均法(average)，重心法(centroid)，离差平方和方法(ward)
- R软件计算：
  - hclust(d, method, members)系统聚类的计算
  - plot()画出谱系图
- Eg1-5: 设有5个样本，每一个样本只有一个指标，分别为1,2,6,8,11. 样本间的距离选用Euclid距离，试用最短距离法，最长距离法等进行聚类分析，并画出相应的谱系图。

# Eg.1-5

```
x<-c(1,2,6,8,11);    dim(x)<-c(5,1)
d<-dist(x)
hc1<-hclust(d,"single");        hc2<-hclust(d,"complete")
hc3<-hclust(d,"median");        hc4<-hclust(d,"average")

opar <- par(mfrow=c(2, 2))
plot(hc1,hang=-1);    plot(hc2,hang=-1)
plot(hc3,hang=-1);    plot(hc4,hang=-1)
par(opar)
```

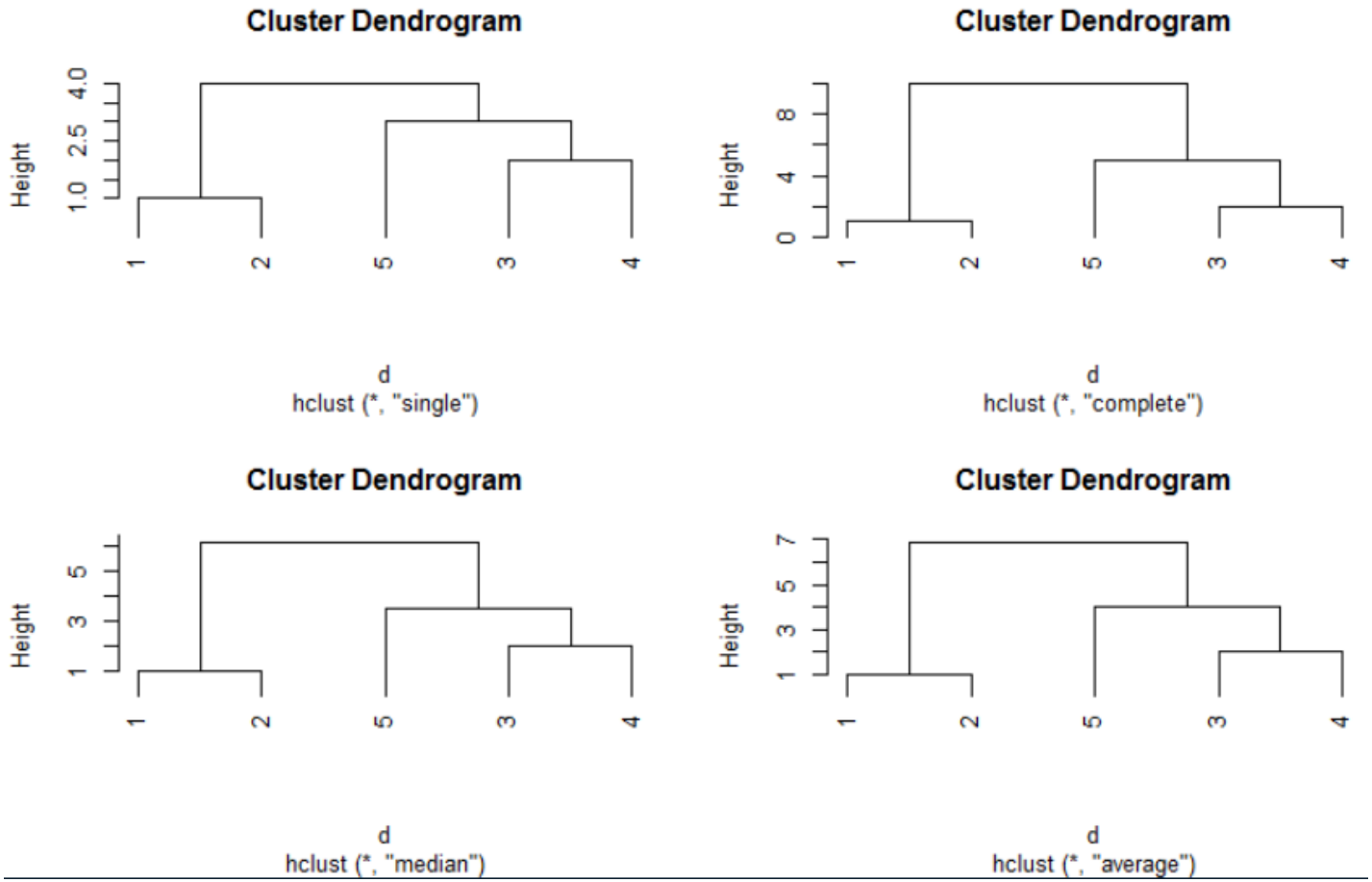



Figure: 四种不同距离下的谱系图

# cophenetic 函数

- cophenetic 函数是用来计算系统聚类的Cophenetic距离，用来计算Cophenetic距离和dist()函数的距离的相关系数，用来评价在众多聚类方法中每一个方法的好坏。通常认为相关系数越接近1，聚类方法越好。
- cophenetic(x);x为hclust()函数生产的对象
- 评价上述例子中的四种聚类方法。

```
####cophenetic()
method<-c('single','complete','median','average')
cc<-numeric(0)
for(m in method){
dc<-cophenetic(hclust(d,m))
cc[m]<-cor(d,dc)
}
cc
```

```
   single      complete    median      average
0.7744479   0.7847885   0.7859780   0.7865155
```

##四种聚类方法中，类平均法的相关系数最高，因此它相对来说是最好的。

# Eg 1-6

Table: 305名女中学生测量的8个体型指标之间的相关系数

| | 身高x1 | 手臂长x2 | 上肢长x3 | 下肢长x4 | 体重x5 | 颈围x6 | 胸围x7 | 胸宽x8 |
|---|---|---|---|---|---|---|---|---|
| 身高x1 | 1.000 | 0.846 | 0.805 | 0.859 | 0.473 | 0.398 | 0.301 | 0.382 |
| 手臂长x2 | 0.846 | 1.000 | 0.881 | 0.826 | 0.376 | 0.326 | 0.277 | 0.415 |
| 上肢长x3 | 0.805 | 0.881 | 1.000 | 0.801 | 0.380 | 0.319 | 0.237 | 0.345 |
| 下肢长x4 | 0.859 | 0.826 | 0.801 | 1.000 | 0.436 | 0.329 | 0.327 | 0.365 |
| 体重x5 | 0.473 | 0.376 | 0.380 | 0.436 | 1.000 | 0.762 | 0.730 | 0.629 |
| 颈围x6 | 0.398 | 0.326 | 0.319 | 0.329 | 0.762 | 1.000 | 0.583 | 0.577 |
| 胸围x7 | 0.301 | 0.277 | 0.237 | 0.327 | 0.730 | 0.583 | 1.000 | 0.539 |
| 胸宽x8 | 0.382 | 0.277 | 0.345 | 0.365 | 0.629 | 0.577 | 0.539 | 1.000 |

Eg 1-6：定义距离为$d_{ij} = 1 - r_{ij}$, 用最长距离法做系统聚类。

```
x<- c(1.000, 0.846, 0.805, 0.859, 0.473, 0.398, 0.301, 0.382,
      0.846, 1.000, 0.881, 0.826, 0.376, 0.326, 0.277, 0.277,
      0.805, 0.881, 1.000, 0.801, 0.380, 0.319, 0.237, 0.345,
      0.859, 0.826, 0.801, 1.000, 0.436, 0.329, 0.327, 0.365,
      0.473, 0.376, 0.380, 0.436, 1.000, 0.762, 0.730, 0.629,
      0.398, 0.326, 0.319, 0.329, 0.762, 1.000, 0.583, 0.577,
      0.301, 0.277, 0.237, 0.327, 0.730, 0.583, 1.000, 0.539,
      0.382, 0.415, 0.345, 0.365, 0.629, 0.577, 0.539, 1.000)
names<-c("身高 x1", "手臂长 x2", "上肢长 x3", "下肢长 x4",
"体重 x5", "颈围 x6", "胸围 x7", "胸宽 x8")
r<-matrix(x, nrow=8, dimnames=list(names, names))

## 作系统聚类分析,
## as.dist()的作用是将普通矩阵转化为聚类分析用的距离结构.
d<-as.dist(1-r); hc<-hclust(d); dend<-as.dendrogram(hc)
## 写一段小程序, 其目的是在绘图命令中调用它, 使谱系图更好看.
nP<-list(col=3:2, cex=c(2.0, 0.75), pch= 21:22,
         bg= c("light blue", "pink"),
         lab.cex = 1.0, lab.col = "tomato")
addE <- function(n){
```

```
    if(!is.leaf(n)) {
        attr(n, "edgePar") <- list(p.col="plum")
        attr(n, "edgetext") <- paste(attr(n,"members"),"members")
    }
    n
}
## 画出谱系图.
op<-par(mfrow=c(1,1), mar=c(4,3,0.5,0))
de <- dendrapply(dend, addE); plot(de, nodePar= nP)
par(op)
##类的确定
plot(hc,hang=-1); re<-rect.hclust(hc,k=3)
cutree(hc,k=3)
```

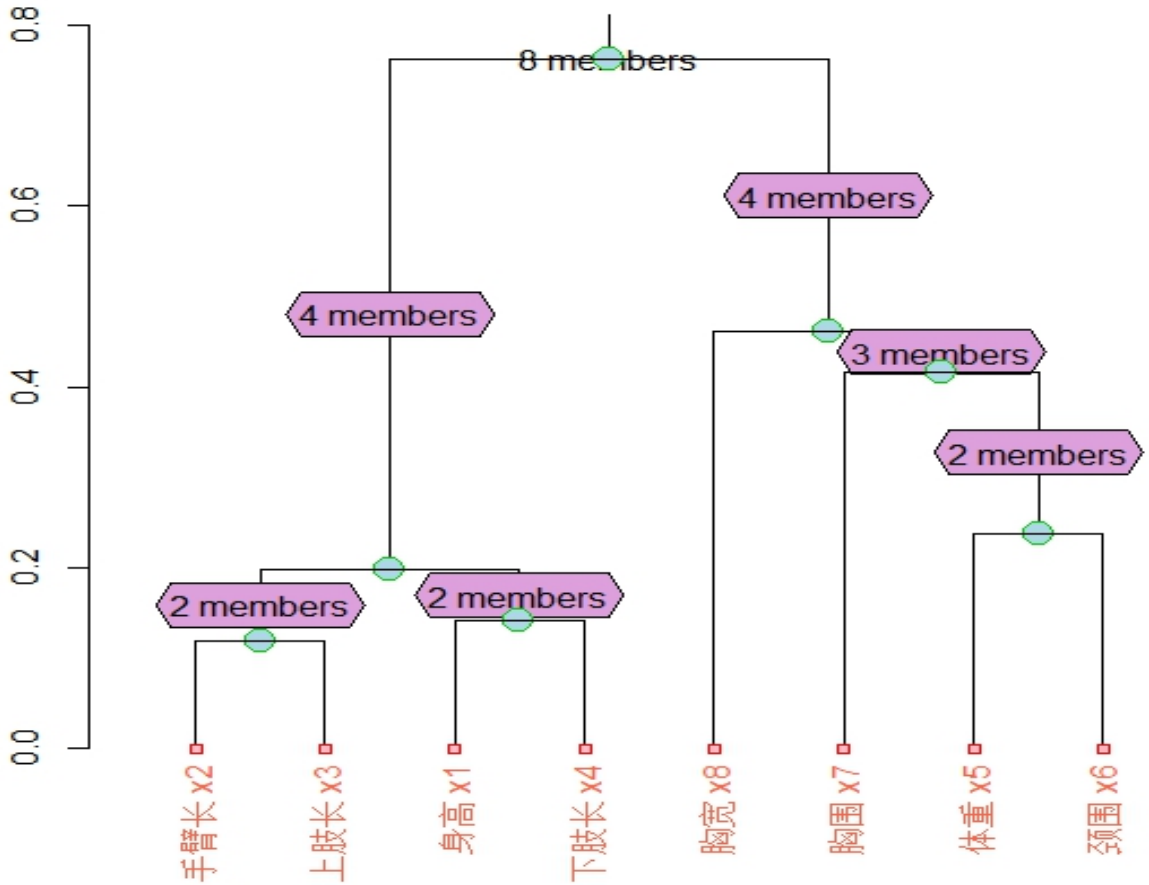



Figure: 8个体型指标的谱系图

# 类个数的确定

- 确定原则
  - 各类重心的距离必须很大
  - 确定的类中，各类所包含的元素都不要太多
  - 类的个数必须符合实用目的
  - 若采用几种不同的聚类方法处理，则在各自的聚类图中应发现相同的类。
- R 命令：
  - rect.hclust(tree, k=NULL,which=NULL,x=NULL,h=NULL,border=2,cluster=NULL)
  - cutree(tree,k=NULL,h=NULL)#k是类的个数；h为要求各类的距离大于h
- 在对8个体型指标的聚类分析中，将变量分为3类，plot(hc,hang=-1); re< −rect.hclust(hc,k=3)

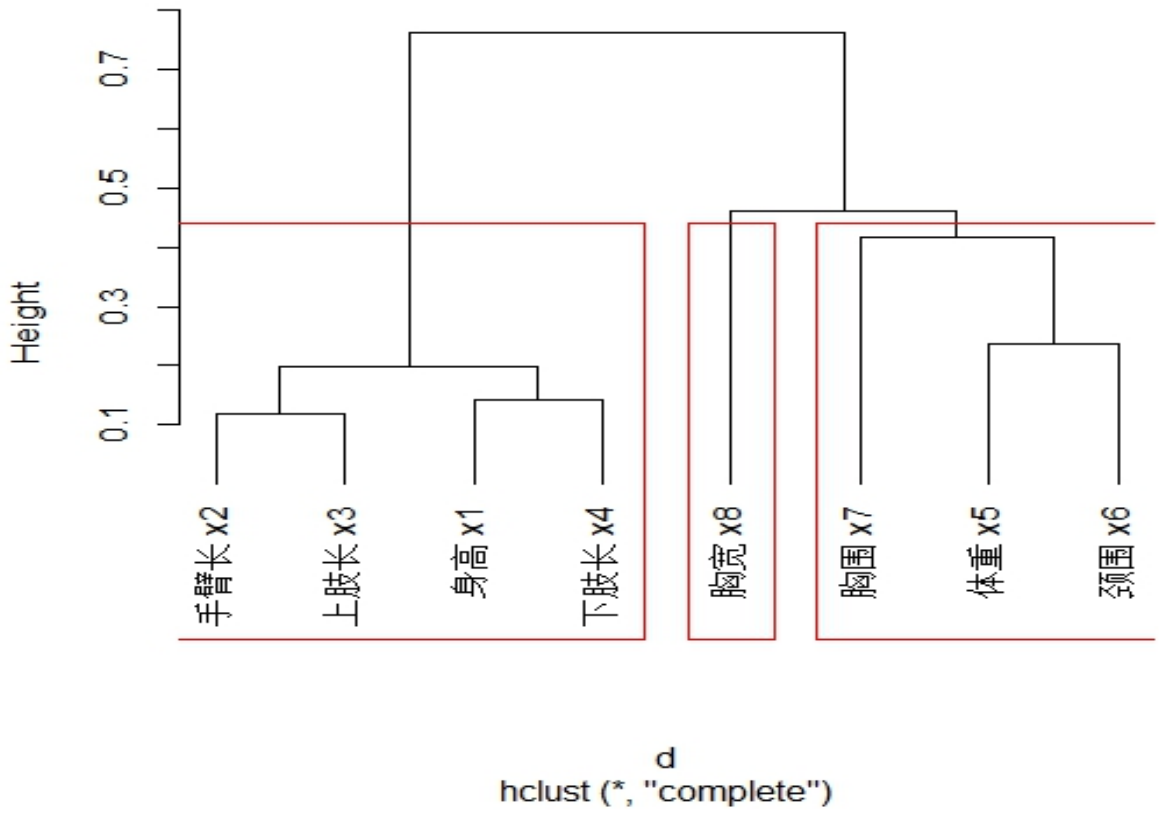



Figure: 8个体型指标的谱系图和聚类情况

# Eg 1-7

Table: 31个省、市、自治区消费性支出数据

| | x1 | x2 | x3 | x4 | x5 | x6 | x7 | x8 |
|---|---|---|---|---|---|---|---|---|
| 北京 | 2959.19 | 730.79 | 749.41 | 513.34 | 467.87 | 1141.82 | 478.42 | 457.64 |
| 天津 | 2459.77 | 495.47 | 697.33 | 302.87 | 284.19 | 735.97 | 570.84 | 305.08 |
| 河北 | 1495.63 | 515.90 | 362.37 | 285.32 | 272.95 | 540.58 | 364.91 | 188.63 |
| 山西 | 1046.33 | 477.77 | 290.15 | 208.57 | 201.50 | 414.72 | 281.84 | 212.10 |
| 内蒙古 | 1303.97 | 524.29 | 254.83 | 192.17 | 249.81 | 463.09 | 287.87 | 192.96 |
| 辽宁 | 1730.84 | 553.90 | 246.91 | 279.81 | 239.18 | 445.20 | 330.24 | 163.86 |
| 吉林 | 1561.86 | 492.42 | 200.49 | 218.36 | 220.69 | 459.62 | 360.48 | 147.76 |
| 黑龙江 | 1410.11 | 510.71 | 211.88 | 277.11 | 224.65 | 376.82 | 317.61 | 152.85 |
| 上海 | 3712.31 | 550.74 | 893.37 | 346.93 | 527.00 | 1034.98 | 720.33 | 462.03 |
| 江苏 | 2207.58 | 449.37 | 572.40 | 211.92 | 302.09 | 585.23 | 429.77 | 252.54 |
| 浙江 | 2629.16 | 557.32 | 689.73 | 435.69 | 514.66 | 795.87 | 575.76 | 323.36 |
| 安徽 | 1844.78 | 430.29 | 271.28 | 126.33 | 250.56 | 513.18 | 314.00 | 151.39 |
| 福建 | 2709.46 | 428.11 | 334.12 | 160.77 | 405.14 | 461.67 | 535.13 | 232.29 |
| 江西 | 1563.78 | 303.65 | 233.81 | 107.90 | 209.70 | 393.99 | 509.39 | 160.12 |
| 山东 | 1675.75 | 613.32 | 550.71 | 219.79 | 272.59 | 599.43 | 371.62 | 211.84 |
| 河南 | 1427.65 | 431.79 | 288.55 | 208.14 | 217.00 | 337.76 | 421.31 | 165.32 |
| 湖北 | 1783.43 | 511.88 | 282.84 | 201.01 | 237.60 | 617.74 | 523.52 | 182.52 |
| 湖南 | 1942.23 | 512.27 | 401.39 | 206.06 | 321.29 | 697.22 | 492.60 | 226.45 |
| 广东 | 3055.17 | 353.23 | 564.56 | 356.27 | 811.88 | 873.06 | 1082.82 | 420.81 |
| 广西 | 2033.87 | 300.82 | 338.65 | 157.78 | 329.06 | 621.74 | 587.02 | 218.27 |
| 海南 | 2057.86 | 186.44 | 202.72 | 171.79 | 329.65 | 477.17 | 312.93 | 279.19 |
| 重庆 | 2303.29 | 589.99 | 516.21 | 236.55 | 403.92 | 730.05 | 438.41 | 225.80 |

Table: 31个省、市、自治区消费性支出数据(续)

| | x1 | x2 | x3 | x4 | x5 | x6 | x7 | x8 |
|---|---|---|---|---|---|---|---|---|
| 四川 | 1974.28 | 507.76 | 344.79 | 203.21 | 240.24 | 575.10 | 430.36 | 223.46 |
| 贵州 | 1673.82 | 437.75 | 461.61 | 153.32 | 254.66 | 445.59 | 346.11 | 191.48 |
| 云南 | 2194.25 | 537.01 | 369.07 | 249.54 | 290.84 | 561.91 | 407.70 | 330.95 |
| 西藏 | 2646.61 | 839.70 | 204.44 | 209.11 | 379.30 | 371.04 | 269.59 | 389.33 |
| 陕西 | 1472.95 | 390.89 | 447.95 | 259.51 | 230.61 | 490.90 | 469.10 | 191.34 |
| 甘肃 | 1525.57 | 472.98 | 328.90 | 219.86 | 206.65 | 449.69 | 249.66 | 228.19 |
| 青海 | 1654.69 | 437.77 | 258.78 | 303.00 | 244.93 | 479.53 | 288.56 | 236.51 |
| 宁夏 | 1375.46 | 480.99 | 273.84 | 317.32 | 251.08 | 424.75 | 228.73 | 195.93 |
| 新疆 | 1608.82 | 536.05 | 432.46 | 235.82 | 250.28 | 541.30 | 344.85 | 214.40 |

表5列出了1999年全国31个省、市、自治区的城镇居民家庭平均每人全年消费性支出的8个主要指标数据，分别为: $x_1$ 食品, $x_1$ 衣着, $x_3$ 家庭设备用品及服务, $x_4$ 医疗保健, $x_5$ 交通与通信, $x_6$ 娱乐教育文化服务, $x_7$ 居住, $x_8$ 杂项食品和服务. 分别用最长距离法，类平均法，重心法和Wald方法对各地区做聚类分析。

```
Province<-dist(scale(X))##消除数据在数量级的影响，需要标准化
hc1<-hclust(Province, "complete")
hc2<-hclust(Province, "average")
hc3<-hclust(Province, "centroid")
hc4<-hclust(Province, "ward.D")
 #绘出谱系图和聚类情况（最长距离法和类平均法）
opar<-par(mfrow=c(2,1), mar=c(5.2,4,2,2))
plot(hc1,hang=-1)
re1<-rect.hclust(hc1,k=5,border="red")
plot(hc2,hang=-1)
re2<-rect.hclust(hc2,k=5,border="red")
par(opar)
#绘出谱系图和聚类情况（重心法和Ward法） opar<-
par(mfrow=c(2,1), mar=c(5.2,4,0,0))
plot(hc3,hang=-1)
re3<-rect.hclust(hc3,k=5,border="red")
plot(hc4,hang=-1)
re4<-rect.hclust(hc4,k=5,border="red")
par(opar)
```

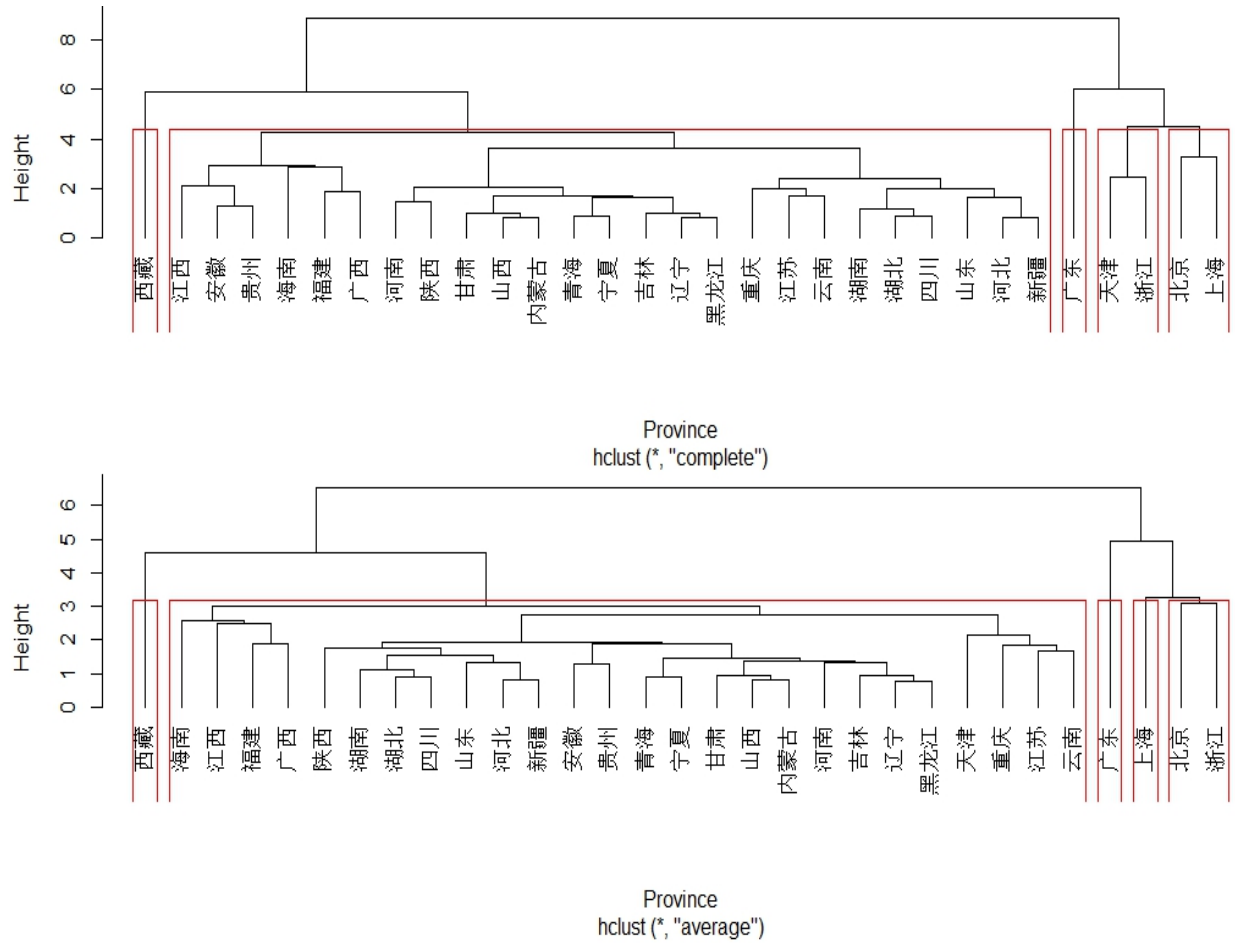



Figure: 消费性支出数据的谱系图和聚类结果(1)

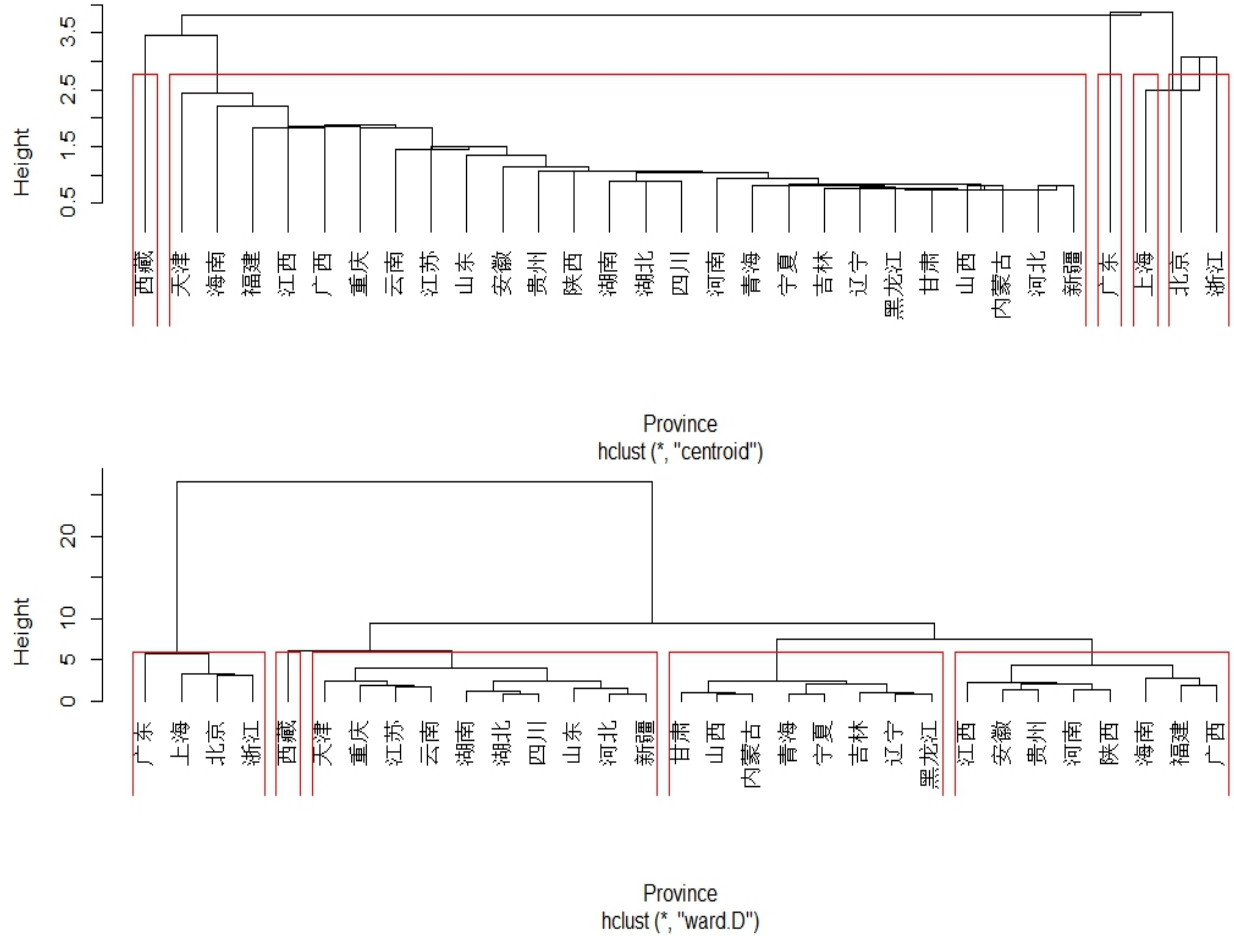



Figure: 消费性支出数据的谱系图和聚类结果(2)

```
##4中方法得到的类有的相同，有些不同，
##可以根据具体数据与背景再进一步确定哪一种聚类方法比较好
####cophenetic()
method<-c('complete','average','centroid','ward')
cc<-numeric(0)
for(m in method){
dc<-cophenetic(hclust(Province,m))
cc[m]<-cor(Province,dc)
}
cc
complete   average  centroid      ward
0.8514284 0.9163281 0.9023272 0.8520098
##类平均法相对聚类效果好
```

动态聚类又称为逐步聚类法

- 基本思想是：开始先给出一个大致的分类，然后按照某种最优原则修改不合理的分类，直至类比较合理位置
- R 函数：kmeans()函数
- 试利用动态聚类方法对31个省、市、自治区的消费水平进行聚类分析.
- km< −kmeans(scale(X),5,nstart=20);km
  sort(km$cluster)

- cluster_Ex1.xls数据中是2011年全国31个省、市、自治区消费性支出数据；
- cluster_Ex2.xls数据中是世界146个国际和地区人文发展情况数据；
- applicant.xls数据中是48名应聘者15项指标数据：FL(求职信的形式)，APP 外貌，专业能力AA，讨人喜欢LA，自信心SC，洞察力LC，诚实HON，推销能力SMS，经验EXP，驾驶水平DRV，事业心AMB，理解能力GSP，潜在能力POT，交际能力KJ和适应性SUIT。选择变量间的相关系数作为相似系数，并定义距离为$d_{ij} = 1 - c_{ij}$；
- 试利用系统聚类分析和动态聚类分析的方法对上述数据进行聚类分析，并给出合理性解释.

# 主成分分析(Principal Component Analysis)

将多指标转化为少数几个综合指标的一种统计方法。

- 本质是：降维，即将高位数据有效地转化为低维数据来处理
- 每一个主成分通常表示为原始变量的线性组合，能够反映原始数据的大部分信息
- 为消除各变量数值大小的差异，主成分分析一般可以从相关矩阵出发求解
- 相关的R函数
  - princomp()
  - summary();loadings()
  - predict();
  - screeplot()画出主成分的碎石图
  - biplot()画出数据关于主成分的散点图和原坐标在主成分下的方向

通过一个实例分析介绍主成分在R中的实现

# Eg 1-8

Table: 30名中学生身体4项指标数据

| 身高$X_1$ | 体重$X_2$ | 胸围$X_3$ | 坐高$X_4$ | 身高$X_1$ | 体重$X_2$ | 胸围$X_3$ | 坐高$X_4$ |
|---|---|---|---|---|---|---|---|
| 148 | 41 | 72 | 78 | 152 | 35 | 73 | 79 |
| 139 | 34 | 71 | 76 | 149 | 47 | 82 | 79 |
| 160 | 49 | 77 | 86 | 145 | 35 | 70 | 77 |
| 149 | 36 | 67 | 79 | 160 | 47 | 74 | 87 |
| 159 | 45 | 80 | 86 | 156 | 44 | 78 | 85 |
| 142 | 31 | 66 | 76 | 151 | 42 | 73 | 82 |
| 153 | 43 | 76 | 83 | 147 | 38 | 73 | 78 |
| 150 | 43 | 77 | 79 | 157 | 39 | 68 | 80 |
| 151 | 42 | 77 | 80 | 147 | 30 | 65 | 75 |
| 139 | 31 | 68 | 74 | 157 | 48 | 80 | 88 |
| 140 | 29 | 64 | 74 | 151 | 36 | 74 | 80 |
| 161 | 47 | 78 | 84 | 144 | 36 | 68 | 76 |
| 158 | 49 | 78 | 83 | 141 | 30 | 67 | 76 |
| 140 | 33 | 67 | 77 | 139 | 32 | 68 | 73 |
| 137 | 31 | 66 | 73 | 148 | 38 | 70 | 78 |

```
#### 用数据框形式输入数据
student<-data.frame(
X1=c(148, 139, 160, 149, 159, 142, 153, 150, 151, 139,
140, 161, 158, 140, 137, 152, 149, 145, 160, 156,
151, 147, 157, 147, 157, 151, 144, 141, 139, 148),
X2=c(41, 34, 49, 36, 45, 31, 43, 43, 42, 31,
29, 47, 49, 33, 31, 35, 47, 35, 47, 44,
42, 38, 39, 30, 48, 36, 36, 30, 32, 38),
X3=c(72, 71, 77, 67, 80, 66, 76, 77, 77, 68,
64, 78, 78, 67, 66, 73, 82, 70, 74, 78,
73, 73, 68, 65, 80, 74, 68, 67, 68, 70),
X4=c(78, 76, 86, 79, 86, 76, 83, 79, 80, 74,
74, 84, 83, 77, 73, 79, 79, 77, 87, 85,
82, 78, 80, 75, 88, 80, 76, 76, 73, 78))


#### 作主成分分析
student.pr<-princomp(student, cor=TRUE)
#### 并显示分析结果
summary(student.pr, loadings=TRUE)
```

```
Importance of components:
                          Comp.1     Comp.2     Comp.3      Comp.
Standard deviation     1.8817805 0.55980636 0.28179594 0.2571184
Proportion of Variance 0.8852745 0.07834579 0.01985224 0.0165274
Cumulative Proportion  0.8852745 0.96362029 0.98347253 1.0000000

Loadings:
   Comp.1 Comp.2 Comp.3 Comp.4
X1  0.497  0.543 -0.450  0.506
X2  0.515 -0.210 -0.462 -0.691
X3  0.481 -0.725  0.175  0.461
X4  0.507  0.368  0.744 -0.232
##第一主成分：大小因子；
##第二主成分：体型因子（高度与围度的差）
#### 作预测
predict(student.pr)
#### 画碎石图
screeplot(student.pr)
screeplot(student.pr,type="lines")
biplot(student.pr)
```

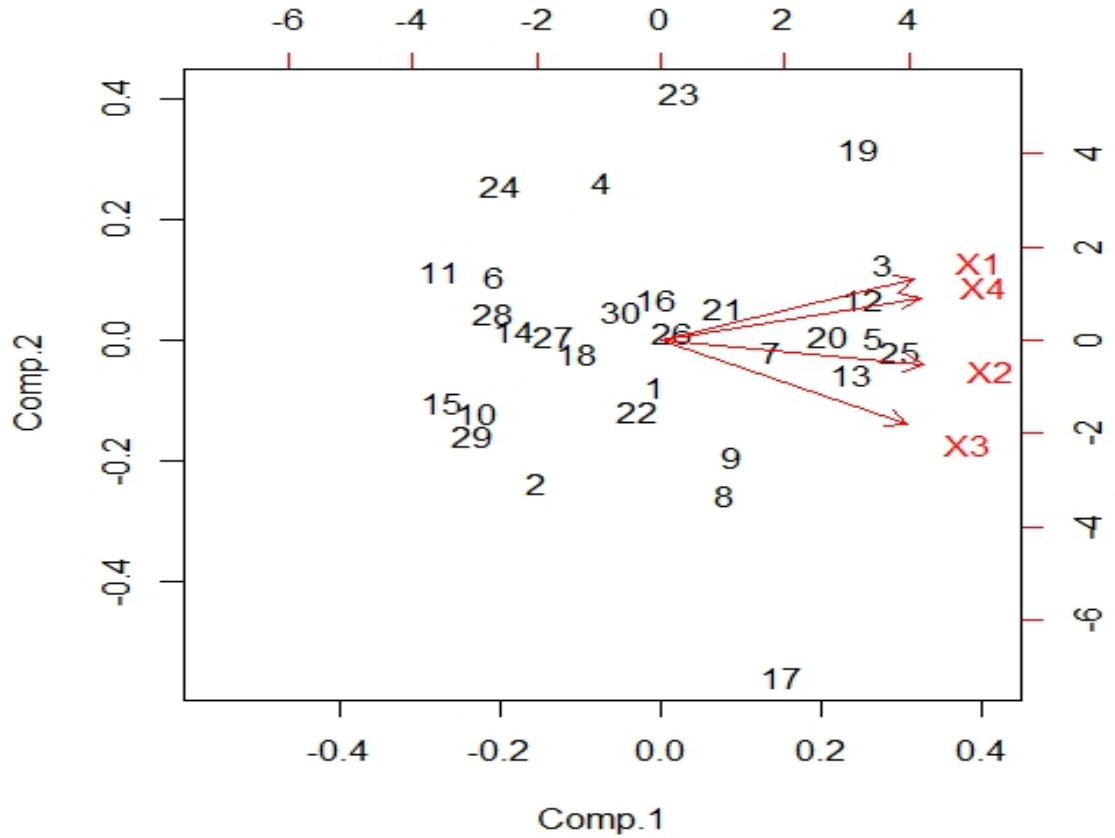



Figure: 30名中学生身体指标数据关于第1主成分和第2主成分的散点图

# 主成分的应用

- 主成分分类：从相关矩阵出发，对各变量进行分类

- Eg 1-9：对128个成年男子的身材进行测量，16个指标依次为：

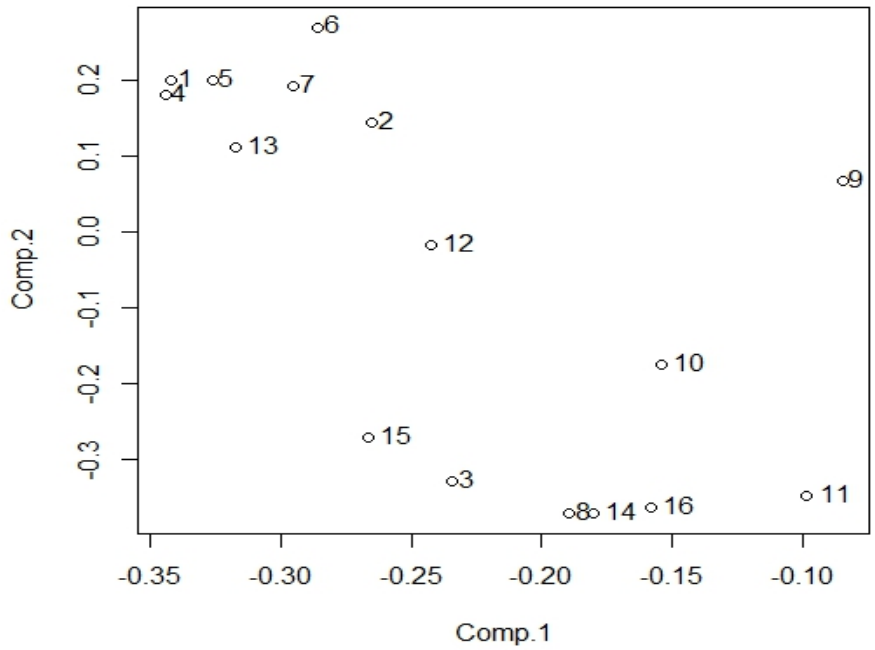
  身高、坐高、胸围、头高、裤长、下档、手长、领围、前胸、后背、肩厚、肩宽、袖长、肋围、腰围、腿肚

  对16项指标进行分类

# Eg 1-9

表 9.2: 16 项身体　标数据的相关矩

| | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ | $X_7$ | $X_8$ | $X_9$ | $X_{10}$ | $X_{11}$ | $X_{12}$ | $X_{13}$ | $X_{14}$ | $X_{15}$ | $X_{16}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $X_1$ | 1.00 | | | | | | | | | | | | | | | |
| $X_2$ | 0.79 | 1.00 | | | | | | | | | | | | | | |
| $X_3$ | 0.36 | 0.31 | 1.00 | | | | | | | | | | | | | |
| $X_4$ | 0.96 | 0.74 | 0.38 | 1.00 | | | | | | | | | | | | |
| $X_5$ | 0.89 | 0.58 | 0.31 | 0.90 | 1.00 | | | | | | | | | | | |
| $X_6$ | 0.79 | 0.58 | 0.30 | 0.78 | 0.79 | 1.00 | | | | | | | | | | |
| $X_7$ | 0.76 | 0.55 | 0.35 | 0.75 | 0.74 | 0.73 | 1.00 | | | | | | | | | |
| $X_8$ | 0.26 | 0.19 | 0.58 | 0.25 | 0.25 | 0.18 | 0.24 | 1.00 | | | | | | | | |
| $X_9$ | 0.21 | 0.07 | 0.28 | 0.20 | 0.18 | 0.18 | 0.29 | -0.04 | 1.00 | | | | | | | |
| $X_{10}$ | 0.26 | 0.16 | 0.33 | 0.22 | 0.23 | 0.23 | 0.25 | 0.49 | -0.34 | 1.00 | | | | | | |
| $X_{11}$ | 0.07 | 0.21 | 0.38 | 0.08 | -0.02 | 0.00 | 0.10 | 0.44 | -0.16 | 0.23 | 1.00 | | | | | |
| $X_{12}$ | 0.52 | 0.41 | 0.35 | 0.53 | 0.48 | 0.38 | 0.44 | 0.30 | -0.05 | 0.50 | 0.24 | 1.00 | | | | |
| $X_{13}$ | 0.77 | 0.47 | 0.41 | 0.79 | 0.79 | 0.69 | 0.67 | 0.32 | 0.23 | 0.31 | 0.10 | 0.62 | 1.00 | | | |
| $X_{14}$ | 0.25 | 0.17 | 0.64 | 0.27 | 0.27 | 0.14 | 0.16 | 0.51 | 0.21 | 0.15 | 0.31 | 0.17 | 0.26 | 1.00 | | |
| $X_{15}$ | 0.51 | 0.35 | 0.58 | 0.57 | 0.51 | 0.26 | 0.38 | 0.51 | 0.15 | 0.29 | 0.28 | 0.41 | 0.50 | 0.63 | 1.00 | |
| $X_{16}$ | 0.21 | 0.16 | 0.51 | 0.26 | 0.23 | 0.00 | 0.12 | 0.38 | 0.18 | 0.14 | 0.31 | 0.18 | 0.24 | 0.50 | 0.65 | 1.00 |




- 左上角的点可以为一类，“长"：身高、坐高、头高、裤长、下档、手长、袖长
- 右下角的点可以为一类，“围"：胸围、领围、肩厚、肋围、腿围、腿肚
- 中间的点为一类，“体型特征":前胸、后背、肩宽

- pca_Ex1.xls给出52名学生的数学、物理、化学、语文、历史、英语的成绩；
- pca_Ex2.xls给出某市工业部门13个产业8项重要经济指标数据，其中x1为年末固定资产净值；x2职工人数；x3工业总产值；x4 全员劳动生产率；x5百元固定资产原值实现产值；x6为资金利税率；x7为标准燃料消费量；x8为能源利用效果
- 对上述两个数据集分别进行主成分分析，并给出相应的解释.

# 判别分析（Discriminant）

- 是在已知样品所有可能分类的前提下，将给定的新样品按照某种分类准则判入其中某个类中的一种多元统计方法.
- 例如：
  - 根据患者的各项检查指标来判断该病人属于哪类病症；
  - 根据某地气象的记录资料来判别(预报)未来几天的天气状况；
  - 根据某地相关经济指标判断该地区属于哪一种经济类型地区；
  - 考古学中，对化石及文物年代的判断；
  - 地质学中，判断是有矿还是无矿；
  - 质量管理中，判断某种产品是合格品还是不合格品；
  - 植物学中，对于新发现的一种植物，判断其属于哪一科。

- 按照判别标准，常用的三种方法
  - 距离判别法
  - Bayes判别法
  - Fisher判别法
- 按照判别函数的形式
  - 线性判别法(Linear Discriminant Analysis, LDA) (Assume each observation comes from a multivariate Gaussian distribution with a class-specific mean vector and a covariance matrix that is common to all $K$ classes.)
  - 二次判别法(Quadratic Discriminant Analysis, QDA)(Assume each observation comes from a multivariate Gaussian distribution with a class-specific mean vector and a covariance matrix that is different for $K$ classes.)
- R中没有单独提供上述3种判别方法，而是将判别方法综合起来，分别给出线性判别函数lda()和二次判别函数qda() (需加载MASS程序包)
- R相关函数
  - lda();qda()
  - 预测或回带: predict() (返回值：class,posterior)

# 调用格式

- 公式形式lda(formula, data, ..., subset, na.action)
  - formula: groups$\sim x_1 + x_2 + \ldots$
  - predict(object,newdata): 新数据必须是数据框的形式
- 矩阵或数据框lda(x, grouping, prior = proportions, tol$= 1.0e-4$, method, CV = FALSE, nu, ...)
  - x: 矩阵或数据框
  - grouping: 指定样本属于哪一类的因子变量(factor(),gl())
  - prior: 各类的先验概率;
  - CV: 如果取值TRUR，返回值包含leave-one-off的交叉判别结果

# Eg 1-10

| 序号 | 春 旱 | | 无 春 旱 | |
|---|---|---|---|---|
| 1 | 24.8 | −2.0 | 22.1 | −0.7 |
| 2 | 24.1 | −2.4 | 21.6 | −1.4 |
| 3 | 26.6 | −3.0 | 22.0 | −0.8 |
| 4 | 23.5 | −1.9 | 22.8 | −1.6 |
| 5 | 25.5 | −2.1 | 22.7 | −1.5 |
| 6 | 27.4 | −3.1 | 21.5 | −1.0 |
| 7 | | | 22.1 | −1.2 |
| 8 | | | 21.4 | −1.3 |

Figure: 某气象站有无春旱的资料

表1中是某气象站监测前14年气象的实际资料，有两个综合预报因子，其中有春旱的是6个年份的资料，无春旱的是8个年份的资料。今年测到两个指标数据为(23.5,-1.6),试用lda()函数和qda()函数对数据做判别分析，并预报今年是否有春旱.

```
############按照矩阵和因子形式输入数据
TrnX1<-matrix(
   c(24.8, 24.1, 26.6, 23.5, 25.5, 27.4,
     -2.0, -2.4, -3.0, -1.9, -2.1, -3.1),
   ncol=2)
TrnX2<-matrix(
   c(22.1, 21.6, 22.0, 22.8, 22.7, 21.5, 22.1, 21.4,
     -0.7, -1.4, -0.8, -1.6, -1.5, -1.0, -1.2, -1.3),
   ncol=2)
Trn<-rbind(TrnX1,TrnX2)
spring<-factor(rep(1:2,c(dim(TrnX1)[1],dim(TrnX2)[1])),
                  labels=c('Have','No'))
##线性判别
lda.sol<-lda(Trn,spring)
Tst<-c(23.5,-1.6)##new data
predict(lda.sol,Tst)$class#无春旱
#[1] No
#Levels: Have No
table(spring,predict(lda.sol)$class)
spring Have No
```

```
  Have    5  1
  No      0  8
##原有的6个有春旱的年份，只判对了5个
##二次判别
qda.sol<-qda(Trn,spring)
Tst<-c(23.5,-1.6)##new data
predict(qda.sol,Tst)$class#有春旱
#[1] Have
#Levels: Have No

table(spring,predict(qda.sol)$class)
spring Have No
  Have    6  0
  No      0  8
###两次预测的结果不一致，但是从回带结果来看，可能有春旱更合理
一些

###############数据框+公式形式输入数据
eg1data<-data.frame(
X1=c(24.8, 24.1, 26.6, 23.5, 25.5, 27.4,
```

```
     22.1, 21.6, 22.0, 22.8, 22.7, 21.5, 22.1, 21.4),
X2=c(-2.0, -2.4, -3.0, -1.9, -2.1, -3.1,
     -0.7, -1.4, -0.8, -1.6, -1.5, -1.0, -1.2, -1.3),
spring=rep(c('Have','No'),c(6,8)))
new<-data.frame(X1=23.5,X2=-1.6)
##线性判别
lda.sol1<-lda(spring~X1+X2,data=eg1data)
predict(lda.sol1,new)$class
table(eg1data$spring,predict(lda.sol1)$class)
##二次判别
qda.sol1<-qda(spring~X1+X2,data=eg1data)
predict(qda.sol1,new)$class
table(eg1data$spring,predict(qda.sol1)$class)
#####选择参数CV=TRUR
lda.sol2<-lda(spring~X1+X2,data=eg1data,CV=TRUE);lda.sol2$class
[1] Have Have Have No Have Have No No No No No No No No
Levels: Have No
qda.sol2<-qda(spring~X1+X2,data=eg1data,CV=TRUE);qda.sol2$class
[1] Have Have Have Have Have Have No No No Have No No No No
Levels: Have No
```

# Eg 1-11: Fisher Iris 数据

调用R内在数据集iris(4个属性，萼片的长度，宽度，花瓣的长度和宽度；数据共150个样本，分为三类)，试用R软件中的两种判别函数对该数据进行判别分析。

- 在150个样本中随机选取100个样本作为训练样本，余下的50个作为测试样本，并假定先验概率各为1/3，并给出预测结果的准确性;
- 用全部样本，并采用leave-one-off的方式对每一个样本进行预测。

```
library('MASS')
data(iris)
head(iris)
##Q1
train<-sample(1:150,100)##抽样
##线性判别
lda.sol<-lda(Species~.,iris,prior=c(1,1,1)/3,subset=train)
class<-predict(lda.sol,iris[-train,])$class
table(iris[-train,]$Species,class)
##sum(class==iris$Species[-train])
##二次判别
qda.sol<-qda(Species~.,iris,prior=c(1,1,1)/3,subset=train)
class<-predict(qda.sol,iris[-train,])$class
table(iris[-train,]$Species,class)
###Q2
lda.cv<-lda(Species~.,iris,prior=c(1,1,1)/3,CV=TRUE)
table(iris$Species,lda.cv$class)
qda.cv<-qda(Species~.,iris,prior=c(1,1,1)/3,CV=TRUE)
table(iris$Species,qda.cv$class)
```

表discriminiant_Ex.excel中列出了1994年我国30个省市自治区影响各地区经济增长差异的制度变量数据,分为两组. 其中

- $x_1$为经济增长率(%);
- $x_2$为非国有化水平(%);
- $x_3$为开放度(%);
- $x_4$为市场化程度(%).
- 利用R已有线性判别和二次判别函数进行判别分析，并对江苏、安徽和陕西三个待判地区作出判定.