数学建模

数据处理与应用

背景意义

- 人类日常生产生活中，每天都面对大量的数据
- 如何从纷繁复杂的数据中挖掘出有用的信息?
- 大量的数据都是现实世界中某些变量的观测
- 现实世界中的变量大多存在随机性和不确定性
- 对观测数据的统计建模分析，是数据处理和应用的重要手段

常用数据统计分析建模方法

1. 描述性统计量和统计图
2. 参数估计
3. 假设检验
4. 回归分析
5. 聚类分析

描述性统计量和统计图

- 描述性统计量
  - 均值、方差、标准差、最大值、最小值、极差、中位数、分位数、众数、变异系数、中心矩、原点矩、偏度、峰度、协方差、相关系数
- 统计图
  - 箱线图、直方图、经验分布函数图、正态概率图、P-P图、Q-Q图

Matlab描述性统计量函数

函数名	说明	函数名	说明
max	最大值	corrcoef	线性相关系数
min	最小值	partialcorr	线性（或秩）偏相关系数
nanmax	忽略缺失值的样本最大值	moment	中心距
nanmin	忽略缺失值的样本最小值	kurtosis	峰度
sum	样本和	skewness	偏度
nansum	忽略缺失值的样本和	prctile	百分位数
mean	样本均值	quantile	分位数
nanmean	忽略缺失值的样本均值	iqr	内4分位极差
median	中位数	mode	众数
nanmedian	忽略缺失值的样本中位数	range	极差
std	样本标准差	geomean	几何平均值
nanstd	忽略缺失值的样本标准差	harmmean	调和平均值
var	样本方差	trimmean	截尾均值
nanvar	忽略缺失值的样本方差	mad	绝对偏差的均值
cov	协方差矩阵	tabulate	频率分布表
nancov	忽略缺失值的协方差矩阵	grpstats	分组统计量
corr	线性（或秩）相关系数	crosstab	列联表

例6.1 体测成绩案例

VarName1	VarName2	VarName3	VarName4	VarName5	VarName6	VarName7	VarName8	VarName9
Text	Text	Number	Number	Categorical	Number	Number	Number	Number
班级	学号	身高	体重	身高体重等级	肺活量	耐力类项目...	柔韧及力量...	速度及灵巧...
90401	9040101	169.8000	48.7000	营养不良	3327	69	72	60
90401	9040102	174	71.5000	超重	2805	84	94	75
90401	9040103	161.9000	52.1000	较低体重	3625	84	72	60
90401	9040104	178.3000	53.8000	营养不良	3678	60	100	50
90401	9040105	159.9000	55.2000	正常体重	3007	63	100	78
90401	9040106	162.1000	57.7000	正常体重	2800	60	87	78
90401	9040107	171.2000	72.2000	肥胖	1609	96	72	63
90401	9040108	162.1000	48.3000	较低体重	3059	75	100	60
90401	9040109	165.3000	62.7000	正常体重	4311	72	92	60
90401	9040110	180	58.3000	较低体重	3921		66	66
90401	9040111	181.8000	93.5000	肥胖	7359	63	60	63
90401	9040112	171.3000	61.6000	正常体重	5201	20	100	78
90401	9040113	180.4000	68	正常体重	6110	69	78	100
90401	9040114	161.4000	44.7000	营养不良	2961	63	100	60
90401	9040115	166	49.1000	较低体重	2583	75		75
90401	9040116	166.1000	46.8000	营养不良	3735		100	66
90401	9040117	158	51.3000	正常体重	3398	69	78	66
90401	9040118	173.2000	63.8000	正常体重	5064	20	78	60
90401	9040119	177.9000	56.6000	较低体重	3065	75	92	60

见附件：体测成绩.xls

1 分组统计

- **grpstats函数用来做分组统计**
  - **按班级分别计算身高、体重、肺活量的均值、标准差、最大最小值等。**

```
%% ex6.1-1
T = readtable('体测成绩.xls','ReadRowNames',false);
T.Properties.VariableNames =
{'Class','StudentId','Height',...
    'Weight','Rank','VC','Score1','Score2','Score3'};
whichstats = {'mean','std','min','max'};
T1 = T(:,{'Class','Height'});
statarray = grpstats(T1,'Class',whichstats)
……
```

其他详见代码

```
statarray =

  2×6 table

            Class       GroupCount    mean_Height    std_Height    min_Height    max_Height
    ________________    __________    ___________    __________    __________    __________

    090401    {'090401'}    19          169.51         7.7001         158           181.8
    090402    {'090402'}    18          164.94         9.3674        149.2          185.7
```

2. 计算变量间的相关系数矩阵

- **提取身高、体重、肺活量、耐力项分数、柔韧和力量分数、速度和灵巧分数，计算相关系数矩阵**

```
T4 =
T(:,{'Height','Weight','VC','Score1','Score2','Score3'});
T4 = table2array(T4);
id = any(isnan(T4),2);
T4(id,:) = [];
corrcoef(T4)
```

```
ans =

    1.0000    0.6572    0.4693   -0.2871   -0.4848   -0.2939
    0.6572    1.0000    0.6166   -0.1658   -0.5068   -0.0661
    0.4693    0.6166    1.0000   -0.4826   -0.3513   -0.0533
   -0.2871   -0.1658   -0.4826    1.0000    0.2930    0.3041
   -0.4848   -0.5068   -0.3513    0.2930    1.0000    0.3213
```

3. 频数和频率分布表

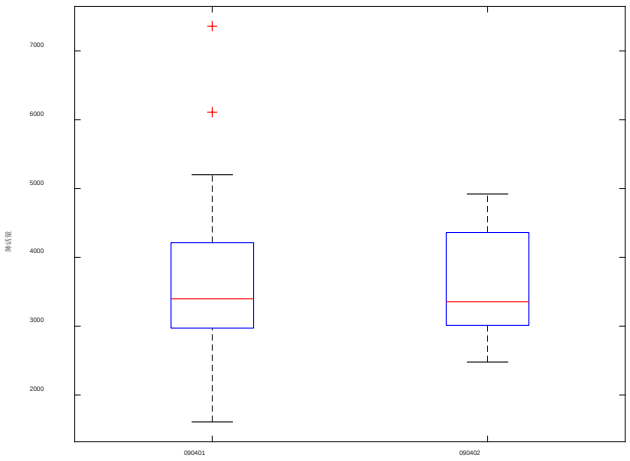
- **tabulate函数用来统计一个数组中各个数字（元素）出现的频数、频率。**

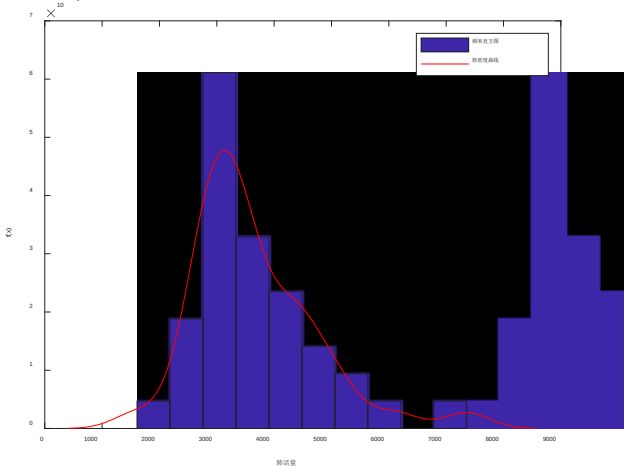
```
T5 = T.Rank; %提取身高体重等级数据
tabulate(T5); %生成数据的频数和频率分布表
```

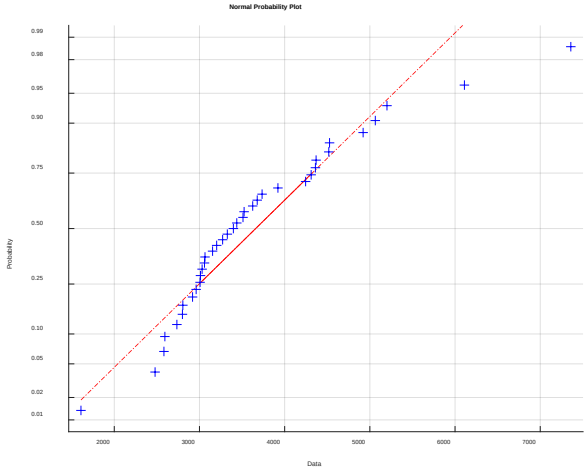
Value	Count	Percent
营养不良	5	13.51%
超重	1	2.70%
较低体重	14	37.84%
正常体重	14	37.84%
肥胖	3	8.11%

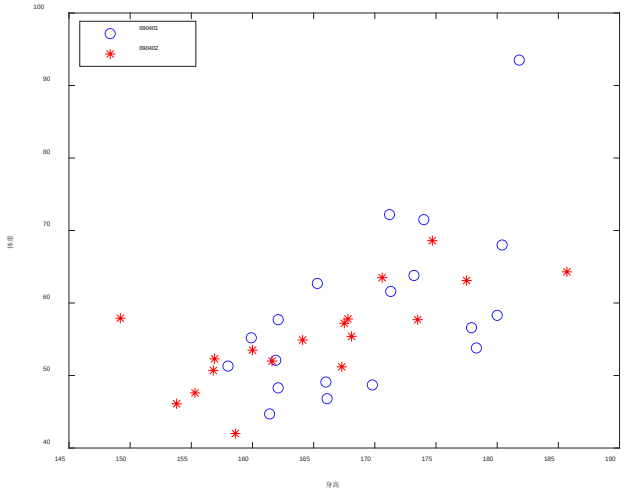
统计图

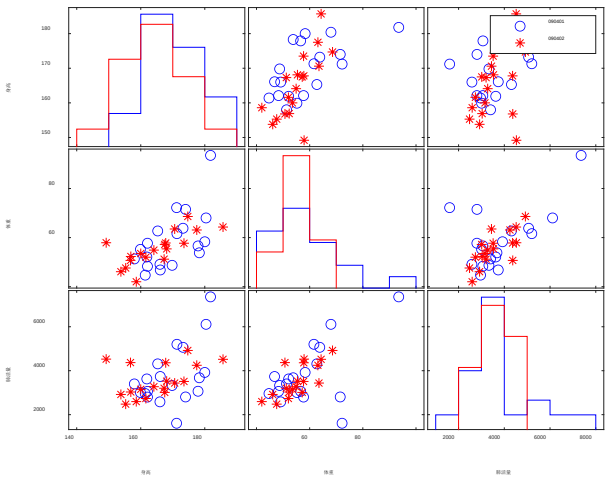
- 1. 箱线图
- 2. 频数直方图
- 3. 正态概率图
- 4. 分组散点图
- 5. 分组散点图矩阵
- 6. 饼图

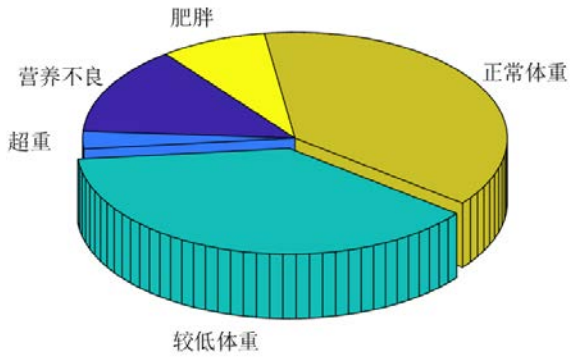












常用数据统计分析建模方法

1. 描述性统计量和统计图
2. 参数估计
3. 假设检验
4. 回归分析
5. 聚类分析

参数估计

- 无论总体 X 的分布函数 $F(x;\theta_1,\theta_2,\cdots,\theta_k)$ 的类型已知或未知，我们总是需要去估计某些未知参数或数字特征，这就是参数估计问题。即参数估计就是从样本 $(X_1,X_2,\cdots,X_n)$ 出发，构造一些统计量 $\hat{\theta}_i$，其中 $\hat{\theta}_i(X_1,X_2,\cdots,X_n)\,(i=1,2,\cdots,k)$ 去估计总体X中的某些参数（或数字特征）$\theta_i(i=1,2,\cdots,k)$ 。 这样的统计量称为估计量.

- 1. 点估计：构造 $(X_1,X_2,\cdots,X_n)$ 的函数 $\hat{\theta}_i(X_1,X_2,\cdots,X_n)$ 作为参数 θ_i 的点估计量，称统计量 $\hat{\theta}_i$ 为总体X参数 θ_i 的点估计量.

- 2.区间估计：构造两个函数 $\theta_{i1}(X_1,X_2,\cdots,X_n)$ 和 $\theta_{i2}(X_1,X_2,\cdots,X_n)$ 做成区间，把这 $(\theta_{i1},\theta_{i2})$ 作为参数 θ_i 的区间估计.

1点估计的求法

- 1.1 矩估计法

- 假设总体分布中共含有k个参数，它们往往是一些原点矩或一些原点矩的函数，例如，数学期望是一阶原点矩，方差是二阶原点矩与一阶原点矩平方之差等。因此，要想估计总体的某些参数$\theta_i(i=1,2,\cdots,k)$由于k个参数一定可以表为不超过k阶原点矩的函数，很自然就会想到用样本的r阶原点矩去估计总体相应的r阶原点矩，用样本的一些原点矩的函数去估计总体的相应的一些原点矩的函数，再将k个参数反解出来，从而求出各个参数的估计值.这就是矩估计法，它是最简单的一种参数估计法。

1.2. 极大似然估计法

- 极大似然法的想法是: 若抽样的结果得到样本观测值：$x1, x2, \cdots, x_n$
- 则应当这样选取参数θ_i的值,使这组样本观测值出现的可能性最大：
- 构造似然函数：

$$L(\theta_1, \theta_2, \cdots, \theta_k) = P(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n) = P(X_1 = x_1)P(X_2 = x_2) \cdots P(X_n = x_n)$$

$$= p(x_1, \theta_1, \cdots \theta_k)\, p(x_2, \theta_1, \cdots, \theta_k) \cdots p(x_n, \theta_1, \cdots \theta_k) = \prod_{j=1}^{n} p\,(x_i, \theta_1, \cdots \theta_k)$$

- 使 $L(\theta_1, \cdots, \theta_k)$ 达到最大，从而得到参数 θ_i的估计值 $\hat{\theta}_i$。此估计值叫极大似然估计值。函数 $L(\theta_1, \cdots, \theta_k)$ 称为**似然函数**。
- 求极大似然估计值的问题，就是求似然函数 $L(\theta_1, \cdots, \theta_k)$ 的最大值的问题，则

$$\frac{\partial L}{\partial \theta_i} = 0 \quad i = 1, 2, \cdots, k$$

2. 区间估计的求法

- 设总体X的分布中含有未知参数θ，若对于给定的概率$1-\alpha(0<\alpha<1)$
- 存在两个统计量 $\hat{\theta}_1(X_1,X_2,\cdots,X_n)$ 和 $\hat{\theta}_2(X_1,X_2,\cdots,X_n)$，使得：
- $$P\left(\hat{\theta}_1<\theta<\hat{\theta}_2\right)=1-\alpha$$
- 则称随机区间$\left(\hat{\theta}_1,\hat{\theta}_2\right)$ 为参数 θ 的置信水平为 $1-\alpha$ 的置信区间，称 $\hat{\theta}_1$为置信下限，称 $\hat{\theta}_2$为置信上限.

Matlab参数估计

- **例6.2-1 某工厂生产的滚珠中随机抽取10个，测得滚珠的直径如下：**
  - 15.14 14.81 15.11 15.26 15.08 15.17 15.12 14.95 15.05 14.87
  - **若直径服从正态分布，求 μ,σ 的最大似然估计和置信水平为90%的置信区间**
- **方法1：**

```
x = [15.14  14.81  15.11  15.26  15.08  15.17
15.12  14.95  15.05  14.87];
[muhat,sigmahat,muci,sigmaci] = normfit(x,0.1)
```

- **方法2：**

```
x = [15.14  14.81  15.11  15.26  15.08  15.17  15.12
14.95  15.05  14.87];
[mu_sigma,mu_sigma_ci] =
mle(x,'distribution','norm','alpha',0.1)
```

自定义分布的参数估计

- **例6.2-2 已知总体X的密度函数为：**

$$f(x;\theta)=\begin{cases}\theta x^{\theta-1} & 0<x<1\\ 0 & other\end{cases}$$

- **先从总体随机取样20个样本，观测值如下：**
  - **0.7917,0.8448,0.9802,0.8481,0.7627,**
    0.9013,0.9037,0.7399,0.7843,0.8424,
    0.9842,0.7134,0.9959,0.6444,0.8362,
    0.7651,0.9341,0.6515,0.7956,0.8733

- **根据观测值，求参数θ 的最大似然估计和置信水平为95%的置信区间。**

```
x = [0.7917,0.8448,0.9802,0.8481,0.7627
        0.9013,0.9037,0.7399,0.7843,0.8424
        0.9842,0.7134,0.9959,0.6444,0.8362
        0.7651,0.9341,0.6515,0.7956,0.8733];
x = x(:);
PdfFun = @(x,theta) theta*x.^(theta-1).*(x>0 & x<1);
[phat,pci] = mle(x,'pdf',PdfFun,'start',1)
```

多参数情形

- **例6.2-3 设总体X服从由正态分布和I型极小值分布（Gumbel分布）混合合成的混合分布，两种分布的比例为0.6和0.4.总体分布密度为：**

$$f(x)=\frac{0.6}{\sqrt{2\pi}\,\sigma_1}\exp\left(\frac{-(x-\mu)^2}{2\sigma_1^2}\right)+\frac{0.4}{\sigma_2}\exp\left(\frac{x-\mu_2}{\sigma_2}\right)\exp\left(-\exp\left(\frac{x-\mu_2}{\sigma_2}\right)\right)$$

- **试生成600个正态分布$(\mu_1=35,\sigma_1=25)$随机数和400个Gumbel分布$(\mu_2=20,\sigma_2=2)$随机数作为样本数据。求$\mu_1,\sigma_1,\mu_2,\sigma_2$的最大似然估计和置信水平为95%的置信区间。**

```
rand('seed',1);
randn('seed',1);
x = normrnd(35,5,600,1);
y = evrnd(20,2,400,1);
data = [x;y];
pdffun =
@(t,mu1,sig1,mu2,sig2)0.6*normpdf(t,mu1,sig1)+0.4*evpdf(t,mu2,sig2);
[phat,pci] = mle(data,'pdf',pdffun,'start',[10,10,10,10],...
    'lowerbound',[-inf,0,-inf,0],'upperbound',[inf,inf,inf,inf])
```

常用数据统计分析建模方法

1. 描述性统计量和统计图
2. 参数估计
3. 假设检验
4. 回归分析
5. 聚类分析

假设检验

- 对总体X的分布律或分布参数作某种假设，根据抽取的样本观察值，运用数理统计的分析方法，检验这种假设是否正确，从而决定接受假设或拒绝假设.

- 1.参数检验：如果观测的分布函数类型已知，这时构造出的统计量依赖于总体的分布函数，这种检验称为参数检验．参数检验的目的往往是对总体的参数及其有关性质作出明确的判断.

- 2.非参数检验：如果所检验的假设并非是对某个参数作出明确的判断，因而必须要求构造出的检验统计量的分布函数不依赖于观测值的分布函数类型，这种检验叫非参数检验．如要求判断总体分布类型的检验就是非参数检验.

假设检验的一般步骤

- 1．根据实际问题提出原假设H0与备择假设H1，即说明需要检验的假设的具体内容；
- 2．选择适当的统计量，并在原假设H0成立的条件下确定该统计量的分布；
- 3．按问题的具体要求，选取适当的显著性水平α，并根据统计量的分布查表，确定对应于α的临界值. 一般α取0.05,0.01或0.10
- 4．根据样本观测值计算统计量的观测值，并与临界值进行比较，从而在检验水平α条件下对拒绝或接受原假设H0作出判断.

单个正态总体均值检验

- 设取出一容量为n的样本，得到均值$\overline{X}$和标准差 s，现要对总体均值μ是否等于某给定值μ_0进行检验. 记:

$$H_0: \mu = \mu_0 \qquad\qquad H_1: \mu \neq \mu_0$$

- 称H0为原假设，H1为备择假设，两者择其一:
  - 接受H0;
  - 拒绝H0，即接受H1.

常用正态总体参数的假设检验

- 1.总体标准差已知时的单个正态总体均值检验
  - U检验
  - matlab函数：ztest
- 2. 总体标准差未知时的单个正态总体均值检验
  - t检验
  - matlab函数：ttest
- 3. 总体均值未知时的单个正态总体方差检验
  - χ^2检验
  - matlab函数：vartest
- 4. 总体标准差未知时的两个正态总体均值比较t检验
  - matlab函数：ttest2
- 5. 总体均值未知时的两个正态总体方差比较F检验
  - matlab函数：vartest2

例6.3-1 U检验

- 某切割正常工作时，切割的金属棒长度服从正态分布N(100,4).从该切割机切割的一批金属棒随机抽取15根，测得长度如下：
  - 97 102 105 112 99 103 102 94 100 95 105 98 102 100 103
- 假设总体方差不变，取显著性水平 $\alpha = 0.05$。检验该切割机工作是否正常。
- 假设：$H_0: \mu = \mu_0 = 100$ 正常　　$H_1: \mu \neq \mu_0$ 异常

```
x = [97  102  105  112  99  103  102  94  100  95  105
98  102  100  103];
mu0 = 100;
Sigma = 2;
Alpha = 0.05;
[h,p,muci,zval] = ztest(x,mu0,Sigma,Alpha)
```

例6.3-2 t检验

- **化肥厂用自动包装机包装化肥，测得某日9包化肥质量(单位kg)为：**
  - 49.4 50.5 50.7 51.7 49.8 47.9 49.2 51.4 48.9
- **设每包化肥质量服从正态分布，是否可以认为每包化肥平均质量为50kg.**
- **假设：** $H_0:\mu=\mu_0=50$ **正常** $H_1:\mu\neq\mu_0$ **异常**

```
x = [49.4  50.5  50.7  51.7  49.8  47.9  49.2  51.4  48.9];
mu0 = 50;
Alpha = 0.05;
[h,p,muci,stats] = ttest(x,mu0,Alpha)
```

例6.3-3 χ^2检验

- **6.3-2例中，每包化肥质量方差是否等于1.5？取显著性水平0.05.**
- **假设：** $H_0:\sigma^2=\sigma_0^2=1.5$ $H_1:\sigma^2\neq\sigma_0^2$

```
x = [49.4  50.5  50.7  51.7  49.8  47.9  49.2
51.4  48.9];
var0 = 1.5;
alpha = 0.05;
tail = 'both';
[h,p,varci,stats] = vartest(x,var0,alpha,tail)
```

例6.3-4 比较t检验

- **甲乙两台机床加工同一产品，从两台机床加工的产品随机抽取若干件，测得直径如下：**
  - $20.1,\ 20.0,\ 19.3,\ 20.6,\ 20.2,\ 19.9,\ 20.0,\ 19.9,\ 19.1,\ 19.9$
  - $18.6,\ 19.1,\ 20.0,\ 20.0,\ 20.0,\ 19.7,\ 19.9,\ 19.6,\ 20.2$
- **试比较甲乙两台机床加工的产品直径是否有显著差异。显著性水平取 0.05.**
- **假设** $H_0:\mu_1=\mu_2$ $H_1:\mu_1\neq\mu_2$

```
alpha = 0.05;
tail = 'both';
vartype = 'equal';
[h,p,muci,stats] =
ttest2(x,y,alpha,tail,vartype)
```

例6.3-5 比较F检验

- **6.3-4例中，两台机床加工产品直径方差是否相等？取显著性水平0.05.**
- **假设：** $H_0:\sigma_1^2=\sigma_2^2$ $H_1:\sigma_1^2\neq\sigma_2^2$

```
alpha = 0.05;
tail = 'both';
[h,p,varci,stats] =
vartest2(x,y,alpha,tail)
```

常用数据统计分析建模方法

1. 描述性统计量和统计图
2. 参数估计
3. 假设检验
4. 回归分析
5. 聚类分析

回归分析

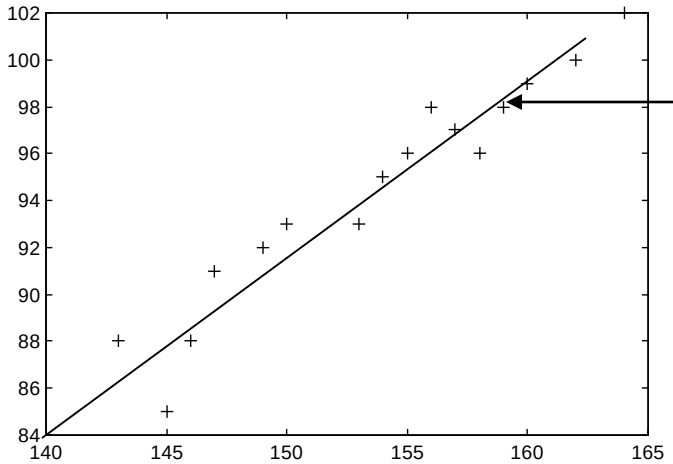
- 研究变量之间的关系
  - 确定性的函数关系
  - 不确定性的相关关系
- 回归分析是研究变量之间的相关关系的数学工具
- 本节讨论内容
  - 1. 一元线性回归
  - 2. 多元线性回归
  - 3. 常见非线性回归

一元线性回归问题引出

例：测16名成年女子的身高与腿长所得数据如下：

身高	143	145	146	147	149	150	153	154	155	156	157	158	159	160	162	164
腿长	88	85	88	91	92	93	93	95	96	98	97	96	98	99	100	102

以身高x为横坐标，以腿长y为纵坐标将这些数据点（x_i，y_i）在平面直角坐标系上标出.



散点图

$$y = \beta_0 + \beta_1 x + \varepsilon$$

一般地，称由 $y=\beta_0+\beta_1 x+\varepsilon$ 确定的模型为**一元线性回归模型**，记为

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

固定的未知参数 β_0 、 β_1 称为回归系数，自变量 x 也称为回归变量.

$Y=\beta_0+\beta_1 x$ ，称为 **y 对 x 的回归直线方程.**

一元线性回归分析的主要任务是:

1、用试验值（样本值）对 β_0 、 β_1 和 σ 作点估计；

2、对回归系数 β_0 、 β_1 作假设检验；

3、在 x= x_0 处对 y 作预测，对 y 作区间估计.

模型参数估计

- **回归系数的最小二乘估计**
  - **有n组独立观测值，（x1，y1），（x2，y2），…，（xn，yn）**
  - **设** $\begin{cases} y_i = \beta_0 + \beta x_1 + \varepsilon_i, i = 1, 2, \ldots, n \\ E\varepsilon_i = 0, D\varepsilon_i = \sigma^2 \quad \text{且 } \varepsilon_1 \varepsilon_2, \ldots, \varepsilon_n \text{ 相互独立} \end{cases}$
  - **记** $Q = Q(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$
- **最小二乘法就是选择** β_0 **和** β_1 **的估计** $\hat{\beta}_0$ **和** $\hat{\beta}_1$**，使得**

$$Q\left(\hat{\beta}_0, \hat{\beta}_1\right) = \min_{\beta_0, \beta_1} Q(\beta_0, \beta_1)$$

求解

- **解得：**

$$\begin{cases} \hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} \\ \hat{\beta}_1 = \dfrac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2} \end{cases} \quad \text{或} \quad \hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2}$$

- **其中：**

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2, \overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i$$

- **（经验）回归方程为：**

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = \overline{y} + \hat{\beta}_1 (x - \overline{x})$$

σ^2的无偏估计

- 记：

$$Q_e = Q\left(\hat{\beta}_0, \hat{\beta}_1\right) = \sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2 = \sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2$$

- 称 Q_e 为残差平方和或剩余平方和。
  - σ^2 的无偏估计为 $\hat{\sigma_e}^2 = \dfrac{Q_e}{n-2}$
- 称 $\hat{\sigma}_e^2$ 为剩余方差（残差的方差），$\hat{\sigma}_e^2$ 分别于 $\hat{\beta}_0, \hat{\beta}_1$ 独立。
- $\hat{\sigma}_e$ 称为剩余标准差

回归方程的显著性检验

- **对回归方程** $Y = \beta_0 + \beta_1 x$ **的显著性检验，归结为对假设:**
  - $H_0: \beta_1 = 0;\ H_1: \beta_1 \neq 0$
  - **进行检验.**
- **假设** $H_0: \beta_1 = 0$ **被拒绝，则回归显著，认为y与x存在线性关系，所求的线性回归方程有意义;**
- **否则回归不显著，y与x的关系不能用一元线性回归模型来描述，所得的回归方程也无意义.**
- **检验方法:**
  - **F检测、t检测，r检测**

回归系数的置信区间

- β_0, β_1 **置信水平为** $1-\alpha$ **的置信区间分别为**

$$\left[\hat{\beta}_0 - t_{1-\frac{\alpha}{2}}(n-2)\hat{\sigma}_e\sqrt{\frac{1}{n}+\frac{\overline{x}^2}{L_{xx}}}, \hat{\beta}_0 + t_{1-\frac{\alpha}{2}}(n-2)\hat{\sigma}_e\sqrt{\frac{1}{n}+\frac{\overline{x}^2}{L_{xx}}}\right]$$

和 $\left[\hat{\beta}_1 - t_{1-\frac{\alpha}{2}}(n-2)\hat{\sigma}_e/\sqrt{L_{xx}}, \hat{\beta}_1 + t_{1-\frac{\alpha}{2}}(n-2)\hat{\sigma}_e/\sqrt{L_{xx}}\right]$

- σ^2 **的置信水平为** $1-\alpha$ **的置信区间分别为**

$$\left[\frac{Q_e}{\chi^2_{1-\frac{\alpha}{2}}(n-2)}, \frac{Q_e}{\chi^2_{\frac{\alpha}{2}}(n-2)}\right]$$

预测

- **用y0的回归值** $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$ **作为y0的预测值.**
- **y0的置信水平为** $1-\alpha$ **的预期区间为**

$$[\hat{y}_0 - \delta(x_0), \hat{y}_0 + \delta(x_0)]$$

- **其中**

$$\delta(x_0) = \hat{\sigma}_e t_{1-\frac{\alpha}{2}}(n-2)\sqrt{1+\frac{1}{n}+\frac{(x_0-\overline{x})^2}{L_{xx}}}$$

- **特别，当n很大且x0在** $\overline{x}$ **附近取值是，y的置信水平为**$1-\alpha$ **的预测区间近似为：**

$$\left[\hat{y} - \hat{\sigma}_e u_{1-\frac{\alpha}{2}}, \hat{y} + \hat{\sigma}_e u_{1-\frac{\alpha}{2}}\right]$$

一元线性回归案例与Matlab实现

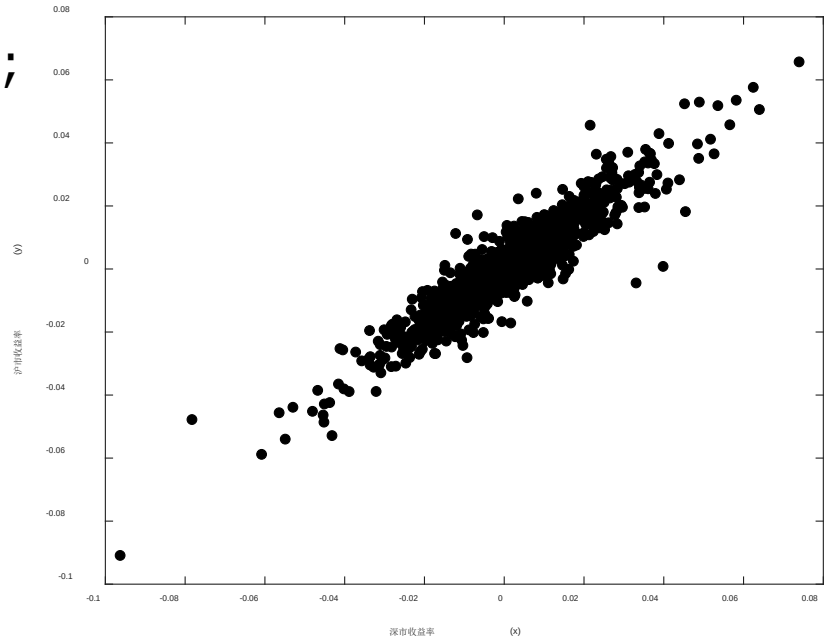
- 现有沪深股市同期日开盘价、最高价、最低价、收盘价、收益率等数据。据此研究沪市收益率y和深市收益率x的关系

Untitled

VarName1	VarName2	VarName3	VarName4	VarName5	VarName6	VarName7	VarName8	VarName9	VarName10	VarName11
Datetime	Number	Number	Number	Number	Number	Number	Number	Number	Number	Number
日期	深市					沪市				
	开盘价	最高价	最低价	收盘价	收益率	开盘价	最高价	最低价	收盘价	收益率
2000-01-04	3.3741e+03	3.5123e+03	3.3602e+03	3.4971e+03	0.0364	1.3687e+03	1.4075e+03	1.3612e+03	1.4064e+03	0.0275
2000-01-05	3.5001e+03	3.5892e+03	3.4687e+03	3.4863e+03	-0.0040	1.4078e+03	1.4338e+03	1.3983e+03	1.4097e+03	0.0013
2000-01-06	3.4755e+03	3.6632e+03	3.4544e+03	3.6552e+03	0.0517	1.4060e+03	1.4640e+03	1.4003e+03	1.4639e+03	0.0412
2000-01-07	3.7015e+03	3.8481e+03	3.7015e+03	3.8280e+03	0.0342	1.4772e+03	1.5228e+03	1.4772e+03	1.5166e+03	0.0267
2000-01-10	3.8818e+03	3.9291e+03	3.8322e+03	3.9215e+03	0.0102	1.5317e+03	1.5467e+03	1.5064e+03	1.5451e+03	0.0087
2000-01-11	3.9247e+03	3.9311e+03	3.6916e+03	3.7168e+03	-0.0530	1.5477e+03	1.5477e+03	1.4688e+03	1.4798e+03	-0.0439
2000-01-12	3.6993e+03	3.7681e+03	3.5855e+03	3.6058e+03	-0.0253	1.4738e+03	1.4893e+03	1435	1.4380e+03	-0.0243
2000-01-13	3.6043e+03	3.6377e+03	3.5521e+03	3.5810e+03	-0.0065	1.4375e+03	1.4441e+03	1.4188e+03	1.4244e+03	-0.0091
2000-01-14	3.5822e+03	3.6092e+03	3.5245e+03	3.5428e+03	-0.0110	1.4262e+03	1.4335e+03	1.4017e+03	1.4089e+03	-0.0122
2000-01-17	3.5355e+03	3.5978e+03	3496	3.5947e+03	0.0167	1.4090e+03	1.4334e+03	1.4027e+03	1.4333e+03	0.0173
2000-01-18	3.6019e+03	3.6331e+03	3.5589e+03	3.5720e+03	-0.0083	1.4369e+03	1.4436e+03	1.4216e+03	1.4266e+03	-0.0071
2000-01-19	3.5606e+03	3.6401e+03	3.5599e+03	3.6039e+03	0.0122	1.4259e+03	1.4437e+03	1.4251e+03	1.4407e+03	0.0104
2000-01-20	3.6082e+03	3.6702e+03	3.6040e+03	3.6688e+03	0.0168	1.4431e+03	1.4669e+03	1.4431e+03	1.4669e+03	0.0165
2000-01-21	3.6872e+03	3.7443e+03	3.6441e+03	3.7021e+03	0.0040	1.4719e+03	1.4765e+03	1.4589e+03	1.4651e+03	-0.0046
2000-01-24	3.7092e+03	3.7364e+03	3.6462e+03	3.7347e+03	0.0069	1.4659e+03	1.4774e+03	1.4495e+03	1.4773e+03	0.0078
2000-01-26	3.7181e+03	3.7541e+03	3.7007e+03	3.7380e+03	0.0053	1.4773e+03	1.4825e+03	1.4700e+03	1.4811e+03	0.0026
2000-01-27	3.7829e+03	3.8554e+03	3.7541e+03	3.8450e+03	0.0164	1.4905e+03	1.5069e+03	1.4852e+03	1.5068e+03	0.0109
2000-01-28	3.8891e+03	3.9524e+03	3.8694e+03	3.9524e+03	0.0163	1.5146e+03	1.5364e+03	1.5108e+03	1535	0.0135

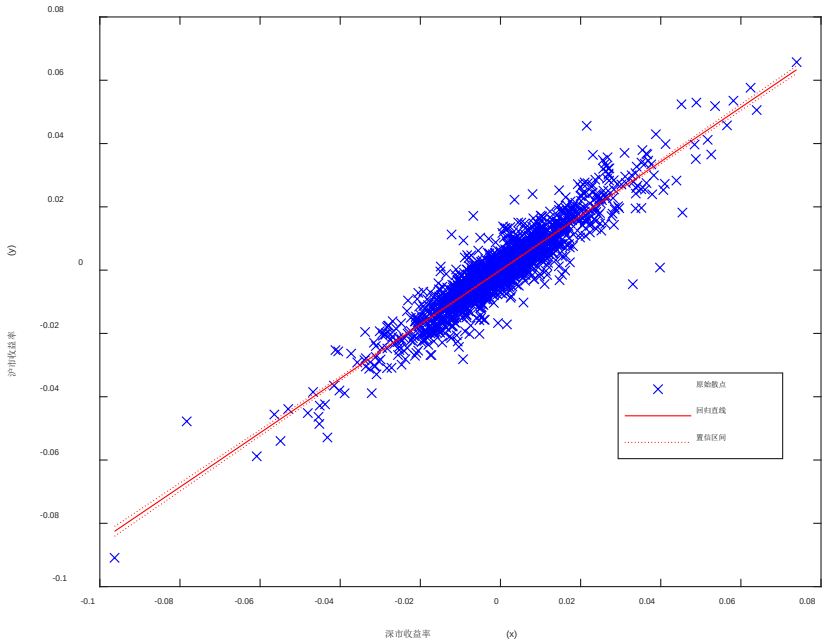
散点图观察

```
data = xlsread('沪深市收益率.xls');
x = data(:,5);
y = data(:,10);
plot(x, y, 'k.', 'Markersize',
15);
xlabel(' 深市收益率(x)');
ylabel('»沪市收益率(y)');
%计算xy线性相关系数
R = corrcoef(x, y)
```



模型建立

- 假定
  - 1. 线性假定
  - 2. 误差正态性假定
  - 3. 误差方差齐性假定
  - 4. 误差独立性假定
- 调用fitlm方法求解模型。
- 调用plot方法绘制拟合效果




```
mdl1 = fitlm(x,y)

figure;
mdl1.plot;
xlabel('深市收益率(x)');
ylabel('沪市收益率(y)');
title('');
legend('原始散点','回归直线','置信区间');
```

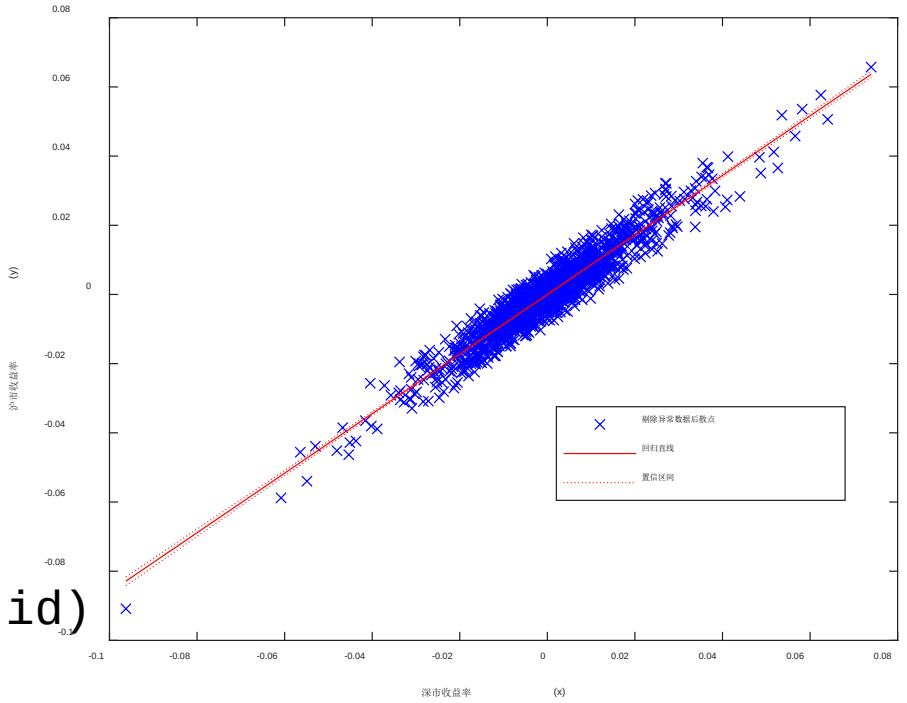
预测与模型改进

- **给定自变量x，调用predict方法计算因变量y**

```
xnew = [0.035,0.04]';
ynew = mdl1.predict(xnew)
```

- **模型改进：**
  - **剔除模型中的不显著项；**
  - **剔除数据集中的异常点。**

```
Res = mdl1.Residuals;
Res_Stu = Res.Studentized;
id = find(abs(Res_Stu)>2);
mdl2 = fitlm(x,y, 'Exclude',id)
```



多元线性回归

一般称 $\begin{cases} Y = X\beta + \varepsilon \\ E(\varepsilon) = 0, COV(\varepsilon, \varepsilon) = \sigma^2 I_n \end{cases}$

为高斯—马尔柯夫线性模型(**k 元线性回归模型**)，并简记为$(Y, X\beta, \sigma^2 I_n)$

$$Y = \begin{bmatrix} y_1 \\ \cdots \\ \cdots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \cdots \\ \beta_k \end{bmatrix}, \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdots \\ \varepsilon_n \end{bmatrix}$$

$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$ 称为**回归平面方程**.

线性模型$(Y, X\beta, \sigma^2 I_n)$考虑的主要问题是：

（1）用试验值（样本值）对未知参数 β 和 σ^2 作点估计和假设检验，从而建立 y 与 $x_1, x_2, \ldots, x_k$ 之间的数量关系；

（2）在 $x_1 = x_{01}, x_2 = x_{02}, \ldots, x_k = x_{0k}$，处对 y 的值作预测与控制，即对 y 作区间估计.

其他常见非线性回归

（1）双曲线 $\frac{1}{y}=a+\frac{b}{x}$

（2）幂函数曲线 $y=a\,x^{b}$，其中 x>0,a>0

（3）指数曲线 $y=a\,e^{bx}$ 其中参数 a>0.

（4）倒指数曲线 $y=a\,e^{b/x}$ 其中 a>0，

（5）对数曲线 y=a+blogx,x>0

（6）S 型曲线 $y=\frac{1}{a+be^{-x}}$ （logistic 曲线函数）

建模实例：软件开发人员的薪金

建立模型研究薪金与资历、管理责任、教育程度的关系.

分析人事策略的合理性，作为新聘用人员薪金的参考.

46名软件开发人员的档案资料

编号	薪金	资历	管理	教育
01	13876	1	1	1
02	11608	1	0	3
03	18701	1	1	3
04	11283	1	0	2
…	…	…	…	…

编号	薪金	资历	管理	教育
42	27837	16	1	2
43	18838	16	0	2
44	17483	16	0	1
45	19207	17	0	2
46	19346	20	0	1

资历～从事专业工作的年数；管理～1=管理人员, 0=非管理人员；教育～ 1=中学，2=大学，3=更高程度.

分析与假设

y~薪金，x_1 ~资历（年）

$x_2=1$~管理人员，$x_2=0$~非管理人员

教育

1=中学

2=大学

3=更高

$$x_3=\begin{cases}1, & \text{中学}\\ 0, & \text{其他}\end{cases}$$

$$x_4=\begin{cases}1, & \text{大学}\\ 0, & \text{其他}\end{cases}$$

⇨

中学：$x_3=1, x_4=0$；

大学：$x_3=0, x_4=1$；

更高：$x_3=0, x_4=0$

假设资历每加一年薪金的增长是常数；

且管理、教育、资历之间无交互作用.

线性回归模型

$$y=a_0+a_1x_1+a_2x_2+a_3x_3+a_4x_4+\varepsilon$$

$a_0, a_1, \ldots, a_4$是待估计的回归系数，ε是随机误差

模型求解

$$y = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_3 + a_4 x_4 + \varepsilon$$

参数	参数估计值	置信区间
a_0	11033	[10258 11807]
a_1	546	[484 608]
a_2	6883	[6248 7517]
a_3	-2994	[-3826 -2162]
a_4	148	[-636 931]
R^2=0.9567 F=226 p<0.0001 s^2=10^6		

$R^2, F, p \rightarrow$ 模型整体上可用

x_1 ~资历（年）

x_2 = 1 ~ 管理，
x_2 = 0 ~ 非管理

中学：x_3=1, x_4=0;
大学：x_3=0, x_4=1;
更高：x_3=0, x_4=0.

资历增加1年
薪金增长546

管理人员薪金
多6883

中学程度薪金比
更高的少2994

大学程度薪金比
更高的多148

a_4置信区间包含零点，解释不可靠！

结果分析　残差分析方法

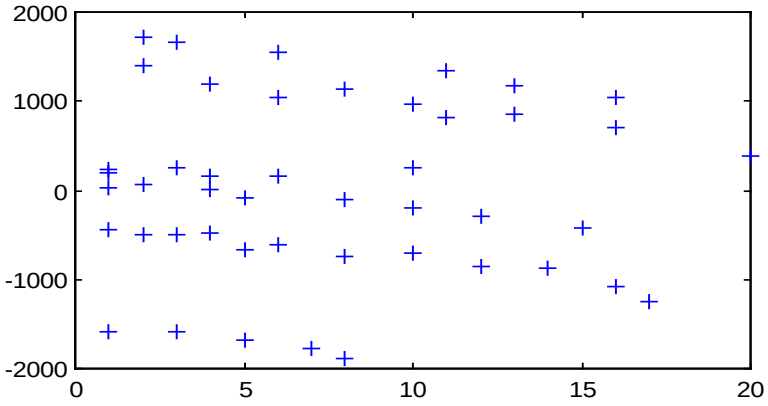
$$\hat{y} = \hat{a}_0 + \hat{a}_1 x_1 + \hat{a}_2 x_2 + \hat{a}_3 x_3 + \hat{a}_4 x_4$$

残差 $e = y - \hat{y}$

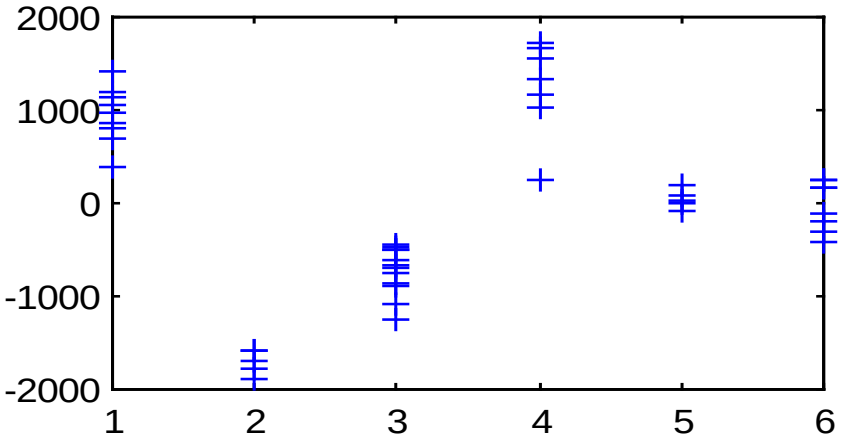
管理与教育的组合

组合	1	2	3	4	5	6
管理	0	1	0	1	0	1
教育	1	1	2	2	3	3

e 与资历 x_1 的关系



e 与管理—教育组合的关系



残差大概分成3个水平，6种管理—教育组合混在一起，未正确反映.

残差全为正，或全为负，管理—教育组合处理不当.

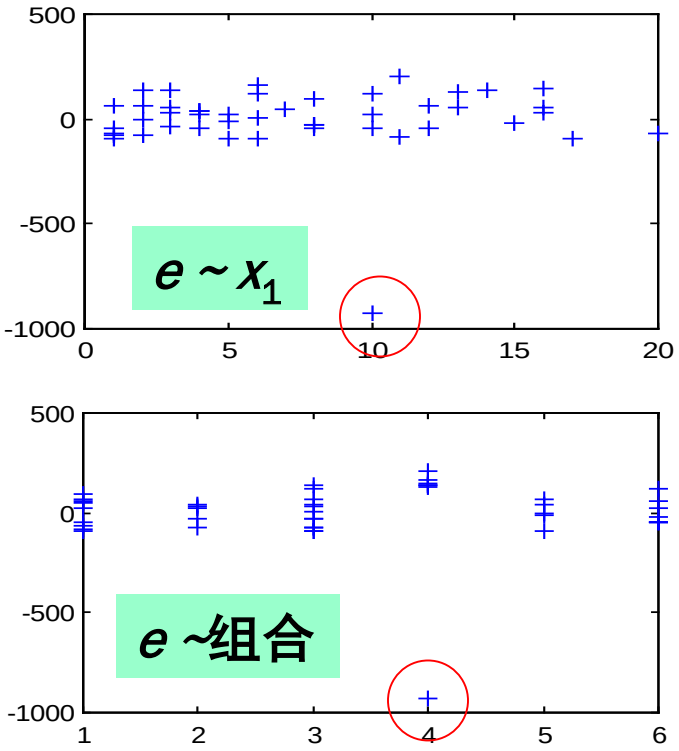
应在模型中增加管理 x_2 与教育 x_3, x_4 的交互项 .

进一步的模型

增加管理x_2与教育x_3, x_4的交互项

$$y = a_0 + a_1x_1 + a_2x_2 + a_3x_3 + a_4x_4 + a_5x_2x_3 + a_6x_2x_4 + \varepsilon$$

参数	参数估计值	置信区间
a_0	11204	[11044 11363]
a_1	497	[486 508]
a_2	7048	[6841 7255]
a_3	-1727	[-1939 -1514]
a_4	-348	[-545 –152]
a_5	-3071	[-3372 -2769]
a_6	1836	[1571 2101]
R^2=0.9988 F=554 p<0.0001 s^2=3×10^4		



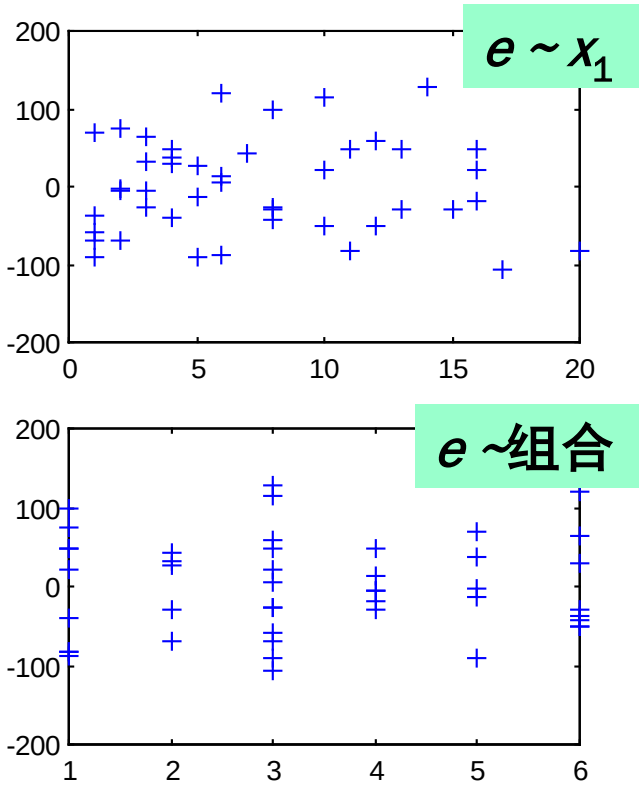
R^2, F有改进, 所有回归系数置信区间不含零点, 模型完全可用

消除了不正常现象

异常数据(33号)应去掉!

去掉异常数据后的结果

参数	参数估计值	置信区间
a_0	11200	[11139 11261]
a_1	498	[494 503]
a_2	7041	[6962 7120]
a_3	-1737	[-1818 -1656]
a_4	-356	[-431 -281]
a_5	-3056	[-3171 -2942]
a_6	1997	[1894 2100]
R^2= 0.9998 F=36701 p<0.0001 s^2=4×10³		



R^2: 0.9567
→0.9988→0.9998

F: 226 → 554 → 36701

s^2: 10^4 → 3×10^4 → 4×10^3

置信区间长度更短

残差图十分正常

最终模型的结果可以应用

模型应用

$$\hat{y}=\hat{a}_0+\hat{a}_1x_1+\hat{a}_2x_2+\hat{a}_3x_3+\hat{a}_4x_4+\hat{a}_5x_2x_3+\hat{a}_6x_2x_4$$

制订6种管理—教育组合人员的“基础”薪金(资历为0)

$x_1=0$； $x_2=1$~管理，$x_2=0$~非管理

中学：$x_3=1, x_4=0$ ； 大学：$x_3=0, x_4=1$； 更高：$x_3=0, x_4=0$

组合	管理	教育	系数	“基础”薪金
1	0	1	a_0+a_3	9463
2	1	1	$a_0+a_2+a_3+a_5$	13448
3	0	2	a_0+a_4	10844
4	1	2	$a_0+a_2+a_4+a_6$	19882
5	0	3	a_0	11200
6	1	3	a_0+a_2	18241

大学程度管理人员比更高程度管理人员的薪金高.

大学程度非管理人员比更高程度非管理人员的薪金略低.

软件开发人员的薪金

对定性因素(如管理、教育)，可以引入0–1变量处理，0–1变量的个数可比定性因素的水平少1.

残差分析方法可以发现模型的缺陷，引入交互作用项常常能够改善模型.

剔除异常数据，有助于得到更好的结果.

注：可以直接对6种管理—教育组合引入5个0–1变量.

常用数据统计分析建模方法

1. 描述性统计量和统计图
2. 参数估计
3. 假设检验
4. 回归分析
5. 聚类分析

聚类分析

- 物以类聚，人以群分
- 聚类分析是研究分类问题的一种多元统计方法。
- 把分类对象按一定规则分成若干类
- 这些类不是事先给定的。
- 在同一类里的对象在某种意义上倾向于彼此类似。
- 本节简单介绍最常见的Kmeans方法。

应用案例：有损压缩

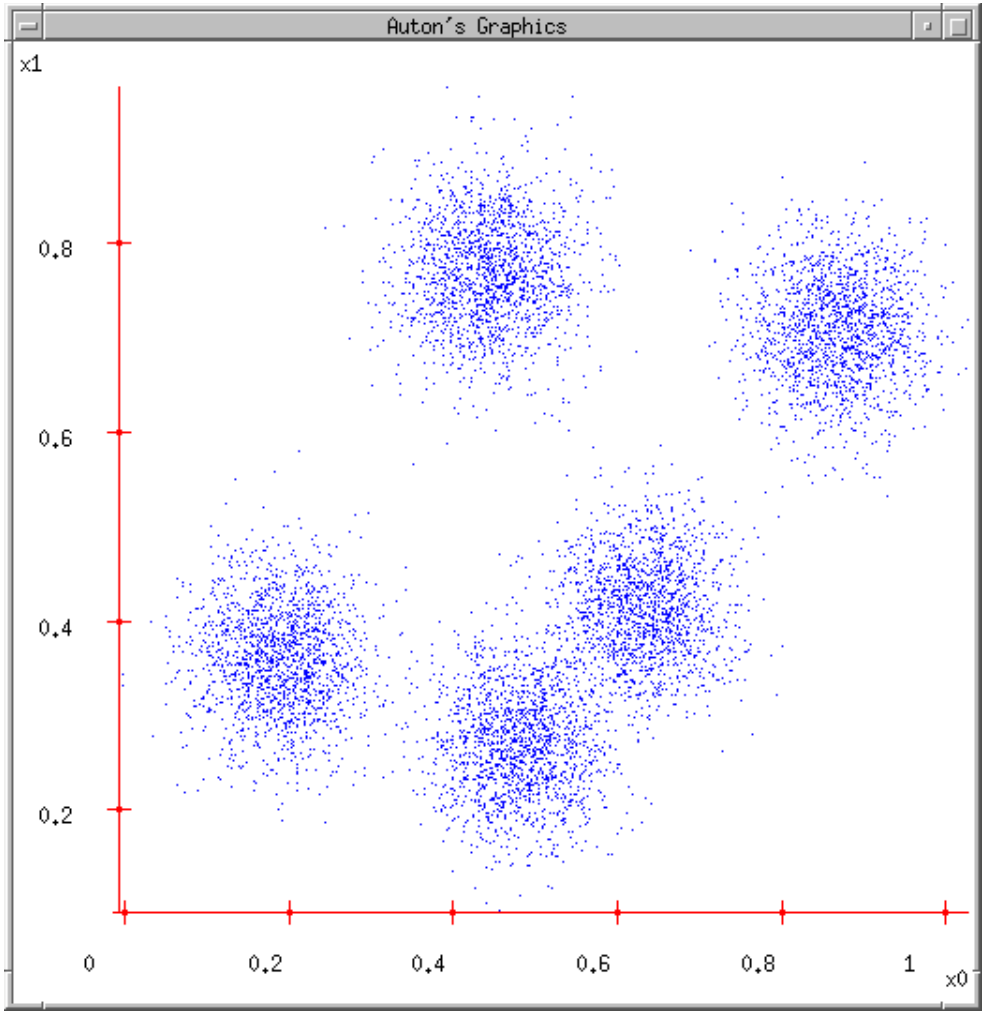
假设一些数据样本在二维空间中分布如图

假设需要用少量的数据传输大量二维空间中的点信息。

类似数字通信中的星座图

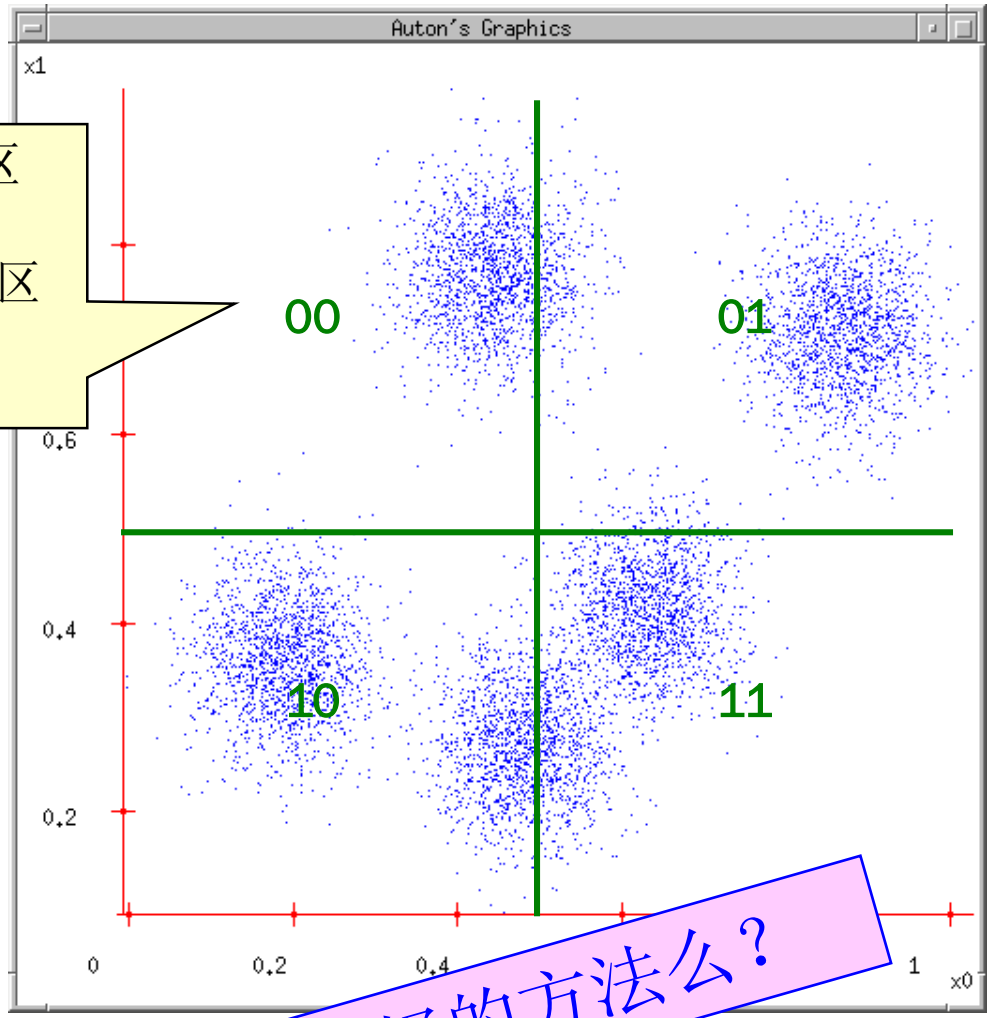
信息损失 =编码和解码坐标距离之和.

如何设计编码/译码方法使信息损失最小?



直观想法

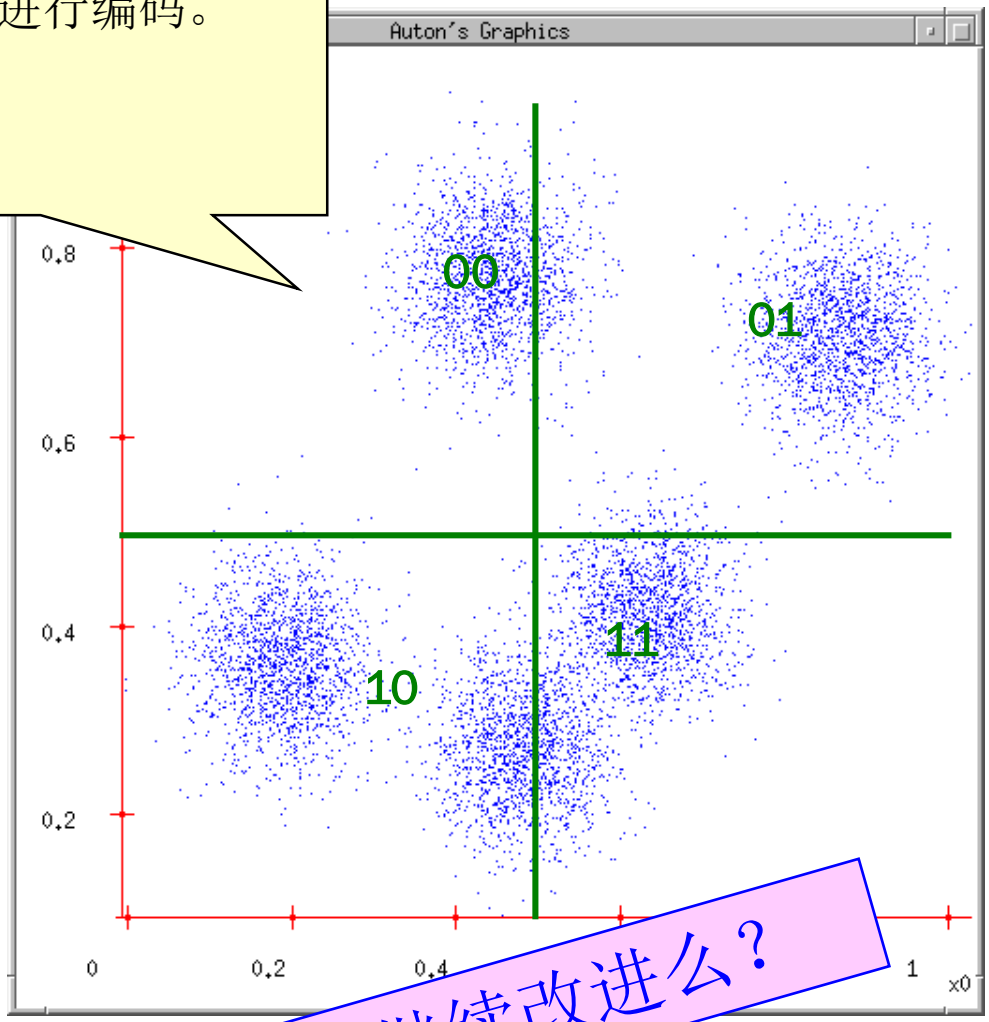
将二维空间划分4个区域；
各个区域内点分别用区域中心坐标代替；
使用



有更好的方法么？

改进

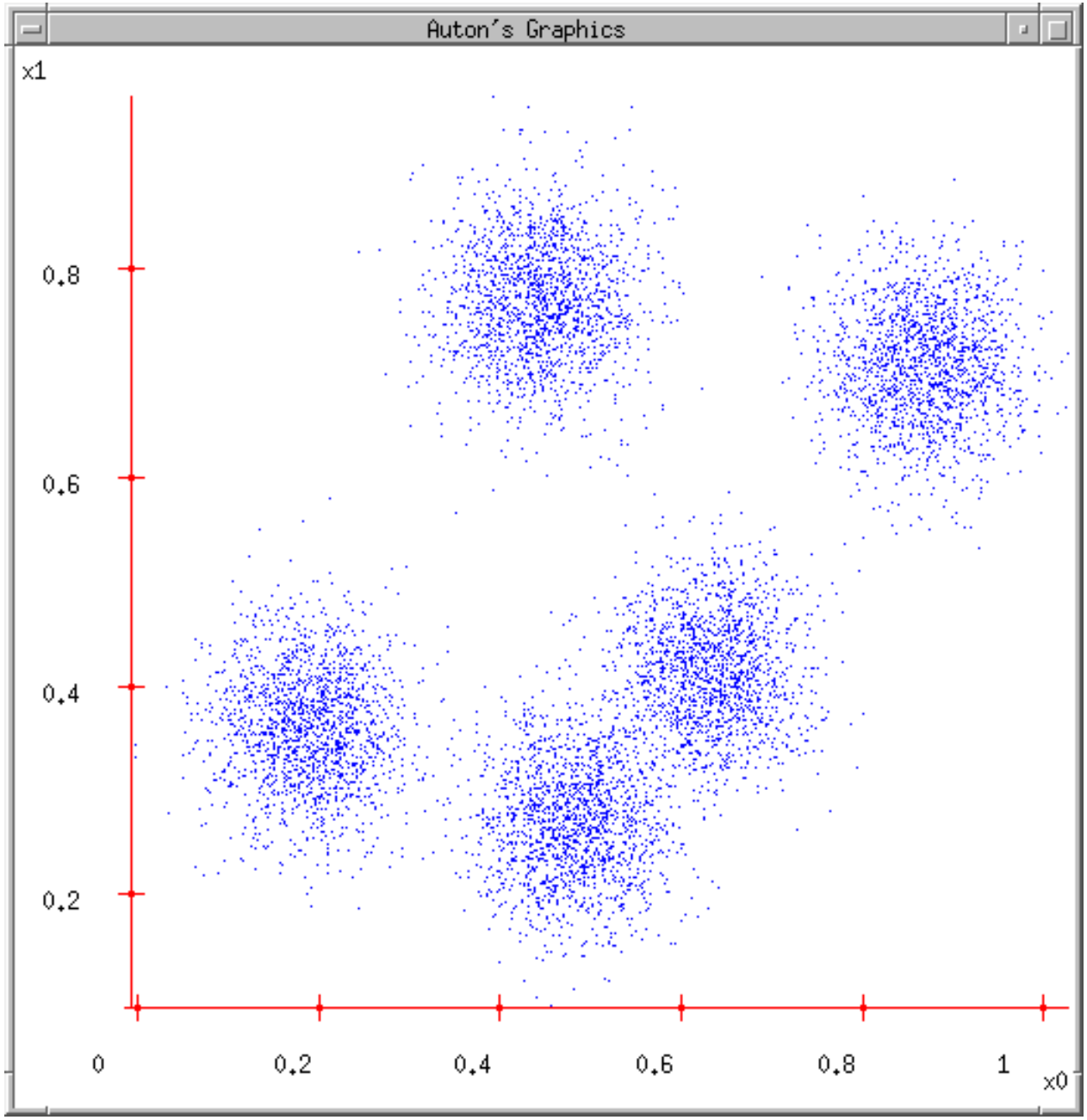
将4个区域中心改进为4个区域内样本点的质心进行编码。



还能继续改进么？

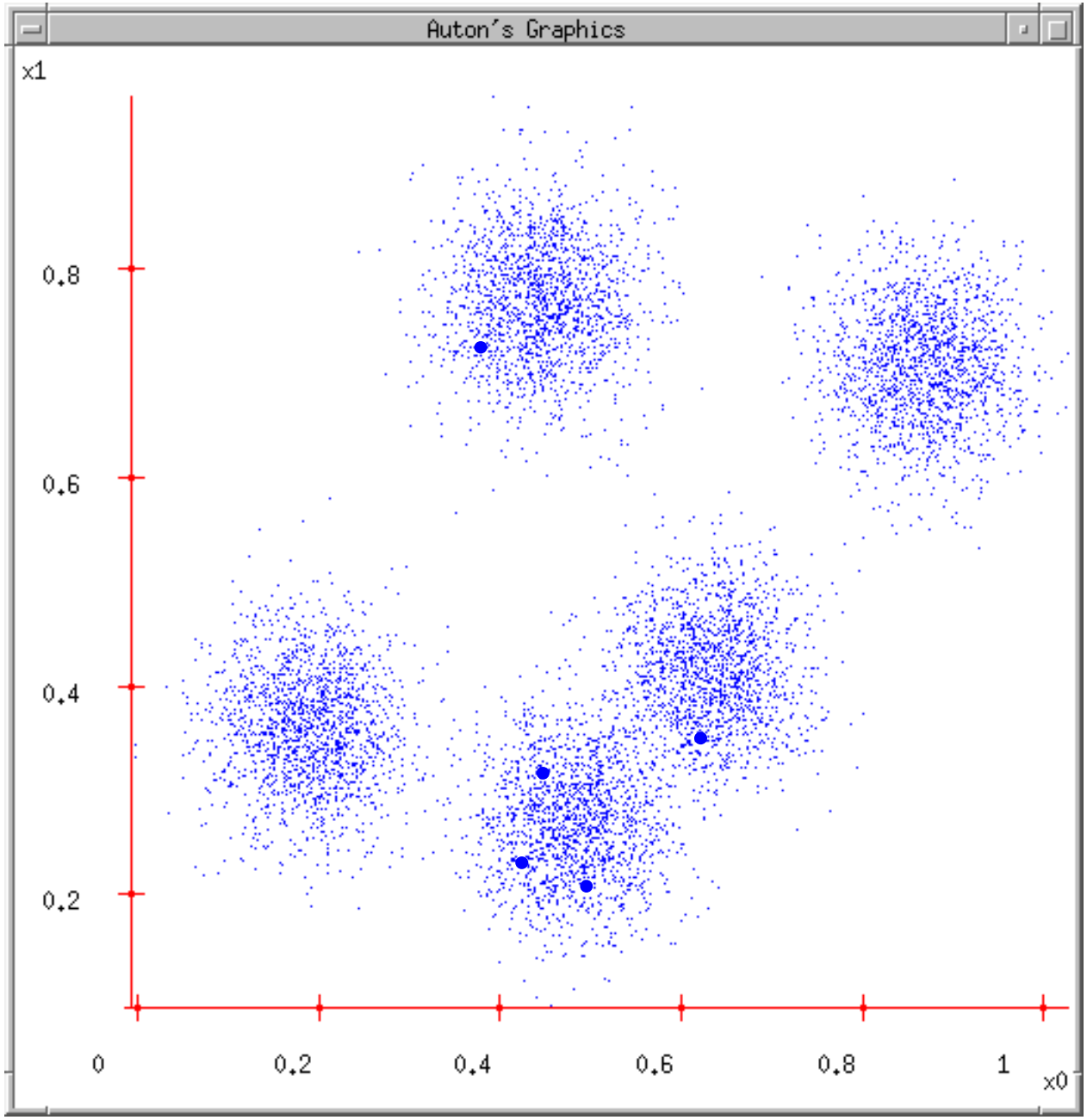
K-means

1. 确定聚类的个数*(e.g. k=5)*



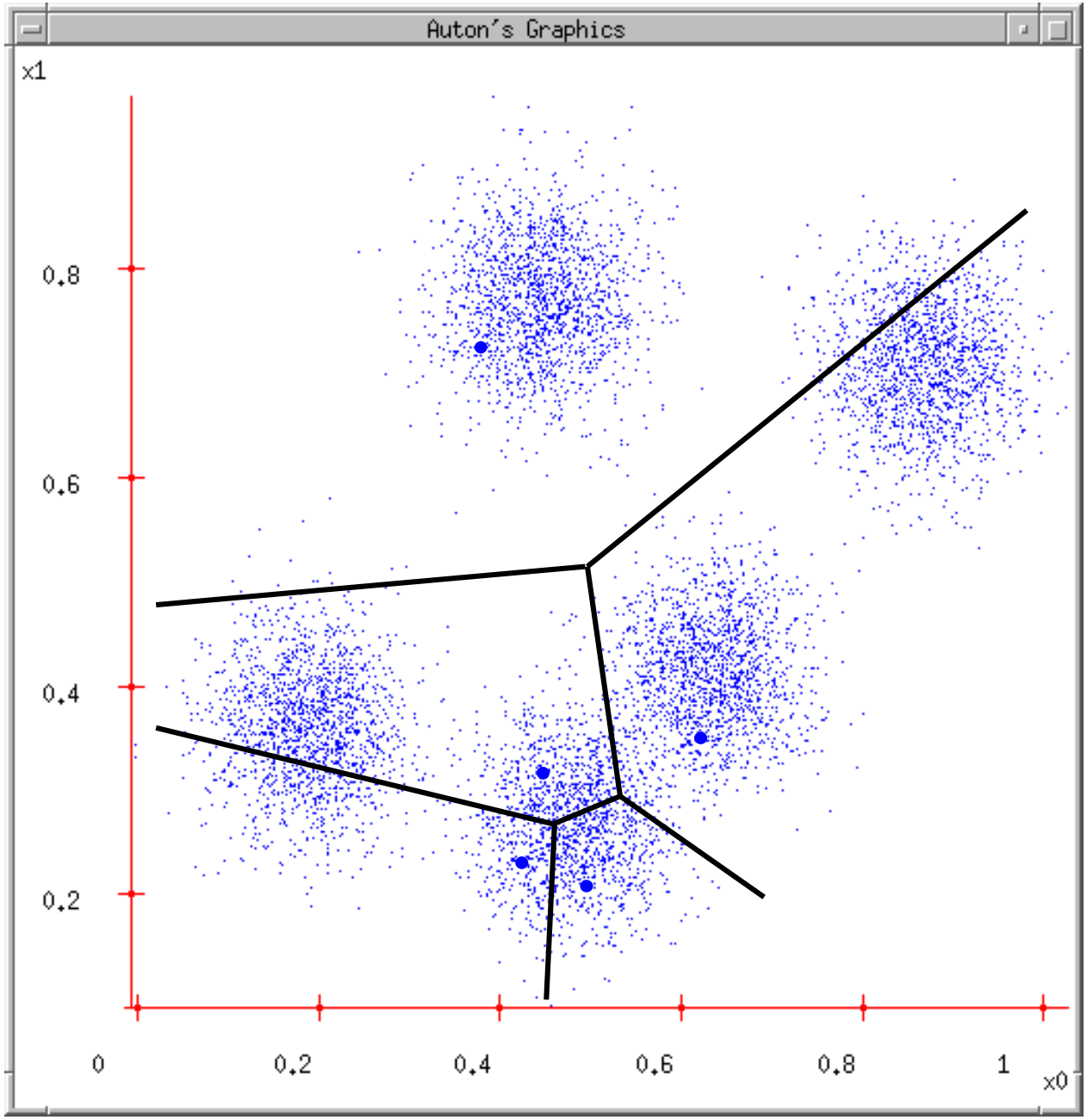
K-means

1. 确定聚类的个数*(e.g. k=5)*
2. 随机猜测每类的质心



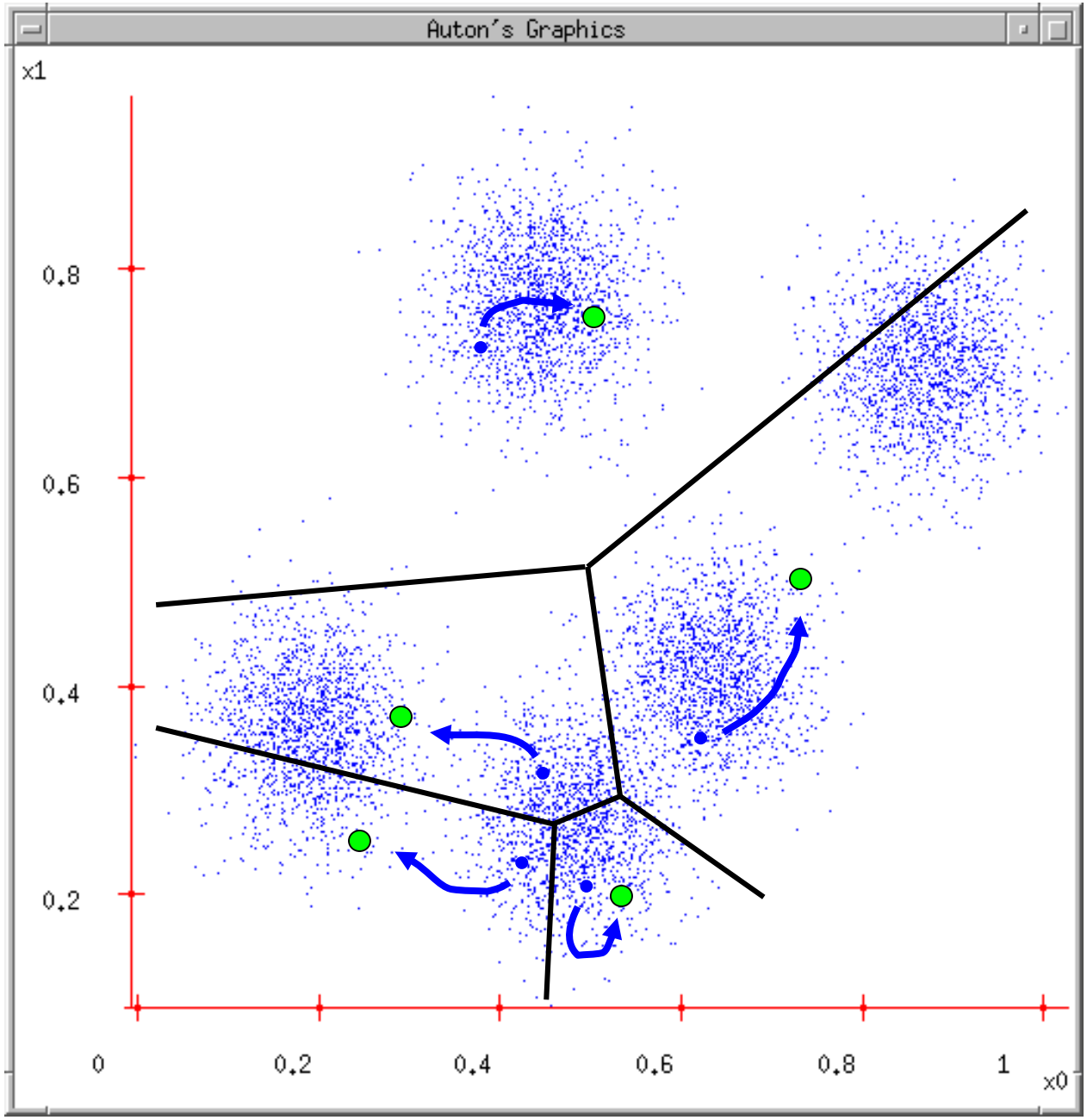
K-means

1. 确定聚类的个数*(e.g. k=5)*
2. 随机猜测每类的质心
3. 每个样本点找到与自己最近的质心，将自己归为该质心所在的类。



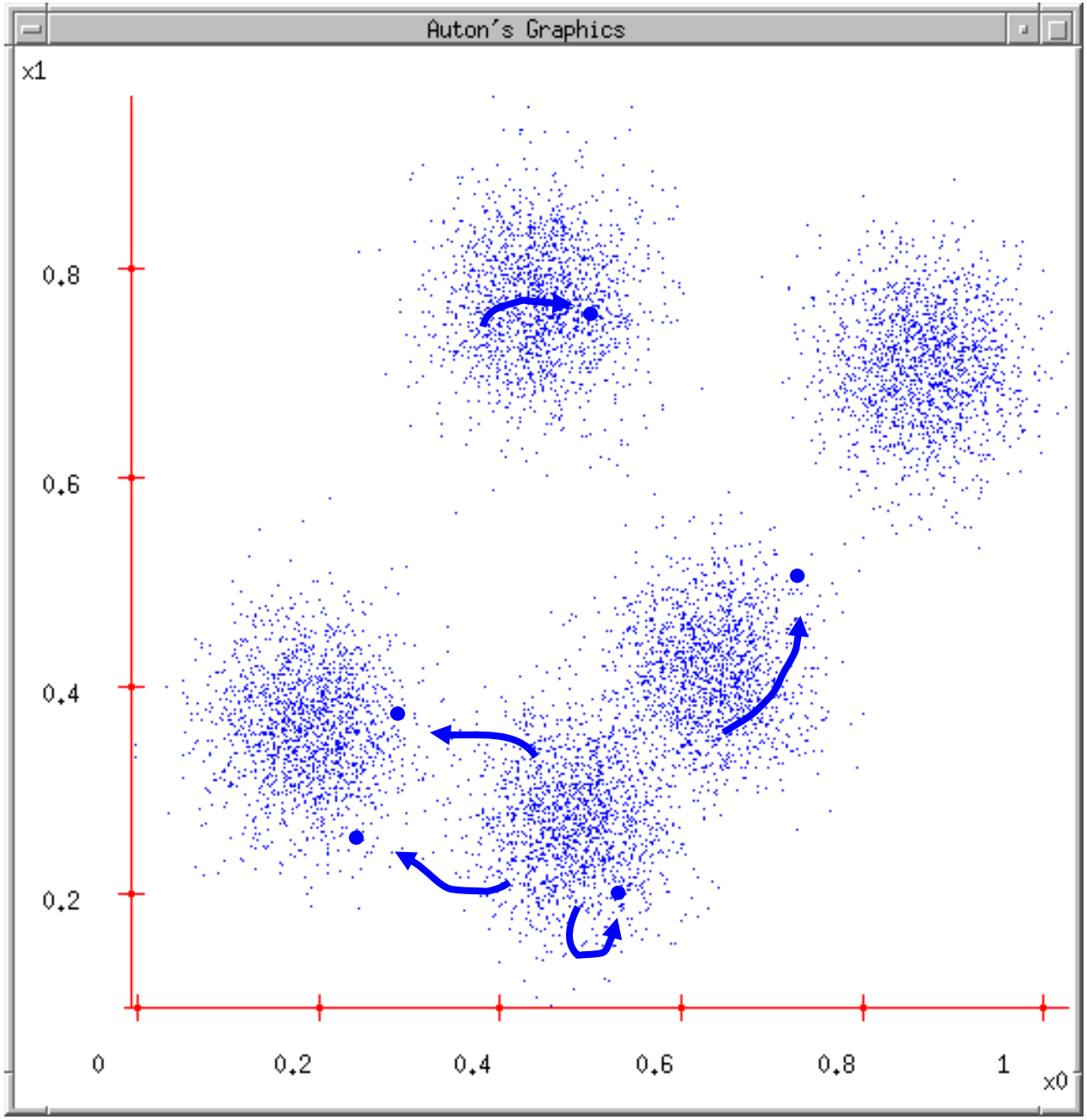
K-means

1. 确定聚类的个数*(e.g. k=5)*
2. 随机猜测每类的质心
3. 每个样本点找到与自己最近的质心，将自己归为该质心所在的类。
4. 归类后，所有的分类重新计算新的质心。



K-means

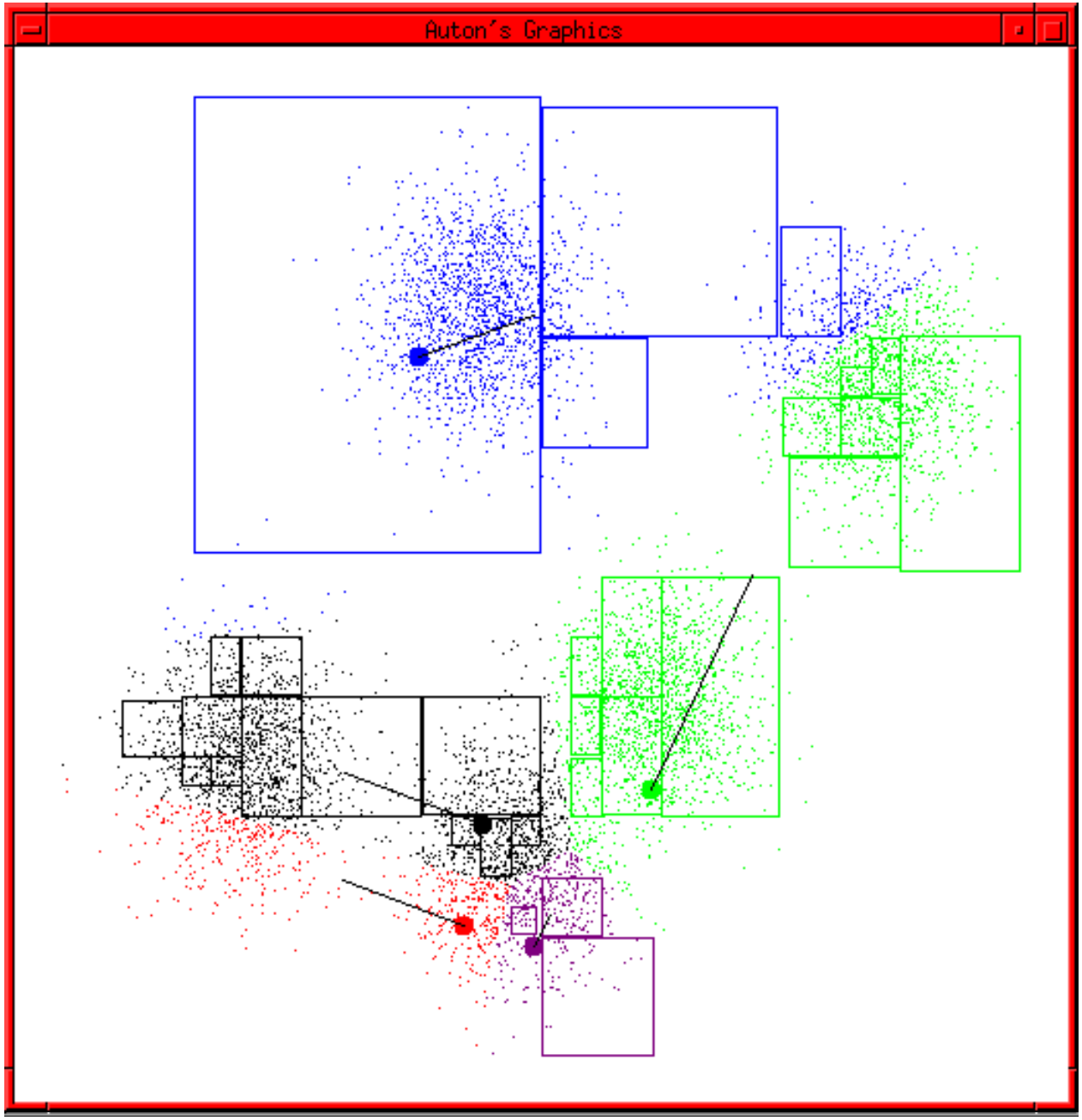
1. 确定聚类的个数*(e.g. k=5)*
2. 随机猜测每类的质心
3. 每个样本点找到与自己最近的质心，将自己归为该质心所在的类。
4. 归类后，所有的分类重新计算新的质心。
5. 质心位置更新...
6. 重复3-5步，直至质心位置不再变化，终止。



K-means 可视化示意

Example generated by Dan Pelleg's super-duper fast K-means system:

Dan Pelleg and Andrew Moore. Accelerating Exact k-means Algorithms with Geometric Reasoning. Proc. Conference on Knowledge Discovery in Databases 1999, (KDD99) (available on www.autonlab.org/pap.html*)*



聚类分析案例

- 2013年-2016年我国31省市自治区人均消费支出数据。
- 根据这些观测数据，对各地区进行聚类分析。

A	B	C	D	E
		Untitled		
VarName1	VarName2	VarName3	VarName4	VarName5
Text	Number	Number	Number	Number
分地区居民...				
单位：元				
城　市	2013年	2014年		
全　国	1.3220e+04	14491.4	1.5712e+04	1.7111e+04
北　京	2.9176e+04	3.1103e+04	3.3803e+04	3.5416e+04
天　津	2.0419e+04	22343	2.4163e+04	2.6129e+04
河　北	1.0872e+04	1.1932e+04	1.3031e+04	1.4248e+04
山　西	1.0118e+04	1.0864e+04	1.1729e+04	1.2683e+04
内蒙古	1.4878e+04	1.6258e+04	1.7179e+04	1.8072e+04
辽　宁	1.4950e+04	16068	1.7200e+04	1.9853e+04
吉　林	1.2054e+04	13026	1.3764e+04	1.4773e+04
黑龙江	1.2037e+04	1.2769e+04	1.3403e+04	1.4446e+04
上　海	3.0400e+04	3.3065e+04	3.4784e+04	3.7458e+04
江　苏	1.7926e+04	1.9164e+04	2.0556e+04	2.2130e+04
浙　江	2.0610e+04	22552	2.4117e+04	2.5527e+04
安　徽	1.0544e+04	11727	1.2840e+04	1.4712e+04
福　建	1.6177e+04	1.7645e+04	1.8850e+04	2.0168e+04
江　西	1.0053e+04	1.1089e+04	1.2403e+04	1.3259e+04
山　东	1.1897e+04	1.3329e+04	1.4578e+04	1.5926e+04
河　南	1.0003e+04	1.1000e+04	1.1835e+04	1.2712e+04
湖　北	1.1761e+04	1.2928e+04	1.4317e+04	1.5889e+04
湖　南	1.1946e+04	1.3289e+04	1.4267e+04	1.5751e+04

1.4491e+04 Converted To [Type: Nu

Kmeans聚类Matlab调用

```
[X,textdata] = xlsread('分地区居民人均消费支出.xls');
obslabel = textdata(4:end,1);
X = zscore(X);

%********************************* Kmeans聚类
*********************************
startdata = X(1:3,:);
id2 = kmeans(X,3,'Start',startdata);
obslabel(id2 == 1)
obslabel(id2 == 2)
obslabel(id2 == 3)
```

实验

实验1

- 某校60名学生的一次考试成绩如下:
  - 93 75 83 93 91 85 84 82 77 76 77 95 94 89 91 88 86 83 96 81 79 97 78 75 67 69 68 84 83 81 75 66 85 70 94 84 83 82 80 78 74 73 76 70 86 76 90 89 71 66 86 73 80 94 79 78 77 63 53 55
- 1)计算均值、标准差、极差、偏度、峰度，画出直方图;
- 2)检验分布的正态性;
- 3)若检验符合正态分布，估计正态分布的参数并检验参数.

实验2

- 据说某地汽油的价格是每加仑115美分，为了验证这种说法，一位学者开车随机选择了一些加油站，得到某年一月和二月的数据如下：
  - 一月：119 117 115 116 112 121 115 122 116 118 109 112 119 112 117 113 114 109 109 118
  - 二月：118 119 115 122 118 121 120 122 128 116 120 123 121 119 117 119 128 126 118 125
- 1）分别用两个月的数据验证这种说法的可靠性；
- 2）分别给出1月和2月汽油价格的置信区间；
- 3）给出1月和2月汽油价格差的置信区间.

实验3

- **考察温度x对某农产品产量y的影响，测得下列10组数据：**

温度（℃）	20	25	30	35	40	45	50	55	60	65
产量（kg）	13.2	15.1	16.4	17.1	17.9	18.7	19.6	21.2	22.5	24.3

- **求y关于x的线性回归方程，检验回归效果是否显著，并预测x=42℃时产量的估值及预测区间（置信度95%）.**

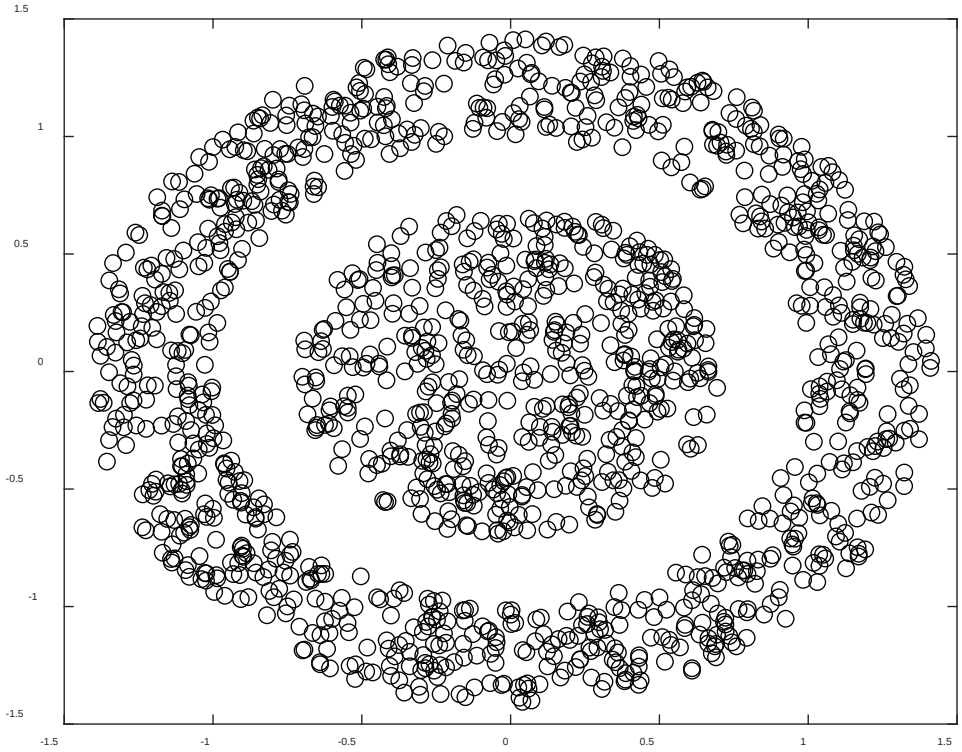
实验4

- **混凝土的抗压强度随养护时间的延长而增加，现将一批混凝土作成12个试块，记录了养护日期x（日）及抗压强度y（kg/cm2）的数据：**

养护时间 x	2	3	4	5	7	9	12	14	17	21	28	56
抗压强度 y	35	42	47	53	59	65	68	73	76	82	86	99

- **试求** $\hat{y} = a + b\ln x$ **型回归方程。**

实验5
思考题



如何实现上图样本的Kmeans聚类？

数据文件：L6data.mat