第2讲 数学建模中的概率统计模型

- 数据的统计描述和分析
- 回归分析
- 概率模型
- 统计回归模型
- 马尔科夫链模型

一、数据的统计描述和分析

1.数据的录入、保存和调用

2.基本统计量

3.频 数 直 方 图 的 描 绘

4.参数估计

5.假设检验

1.数据的录入、保存和调用

例1.1 上海市区社会商品零售总额和全民所有制职工工资总额的数据如下

年份	78	79	80	81	82	82	84	85	86	87
职工工资总额（亿元）	23.8	27.6	31.6	32.4	33.7	34.9	43.2	52.8	63.8	73.4
商品零售总额（亿元）	41.4	51.8	61.7	67.9	68.7	77.5	95.9	137.4	155.0	175.0

方法1

1、年份数据以1为增量，用产生向量的方法输入。命令格式： **x=a:h:b**

t=78:87;

2、分别以x和y代表变量职工工资总额和商品零售总额。

x=[23.8,27.6,31.6,32.4,33.7,34.9,43.2,52.8,63.8,73.4];

y=[41.4,51.8,61.7,67.9,68.7,77.5,95.9,137.4,155.0,175.0];

3、将变量t、x、y的数据保存在文件data中。

save data t x y

4、进行统计分析时，调用数据文件data中的数据。

load data

方法2

1、输入矩阵：

```
data=[78,79,80,81,82,83,84,85,86,87;
 23.8,27.6,31.6,32.4,33.7,34.9,43.2,52.8,63.8,73.4;
 41.4,51.8,61.7,67.9,68.7,77.5,95.9,137.4,155.0,175.0]
```

2、将矩阵data的数据保存在文件data1中：

save data1 data

3、进行统计分析时，先用命令： load data1

调用数据文件data1中的数据，再用以下命令分别将矩阵data的各行的数据赋给变量t、x、y：

t=data(1,:)

x=data(2,:)

y=data(3,:)

若要调用矩阵data的第j列的数据，可用命令：

data(:,j)

2.基本统计量

对随机变量x，计算其基本统计量的命令如下：

- 均值：mean(x)
- 中位数：median(x)
- 标准差：std(x)
- 方差：var(x)
- 偏度：skewness(x)
- 峰度：kurtosis(x)

对例1.1中的职工工资总额x，可计算上述基本统计量

```
load data1
x=data(2,:)
Mean=mean(x)
Median=median(x)
Std=std(x)
Var=var(x)
Skewness=skewness(x)
Kurtosis=kurtosis(x)
```

3.频 数 直 方 图 的 描 绘

1、给出数据data的频数表的命令为：

[N,X]=hist(data,k)

此命令将区间[min(data),max(data)]分为k个小区间（缺省为10），返回数据data落在每一个小区间的频数N和每一个小区间的中点X.

2、描绘数组data的频数直方图的命令为：

hist(data,k)

4. 参数估计

1、正态总体的参数估计

设总体服从正态分布，则其点估计和区间估计可同时由以下命令获得：

[muhat,sigmahat,muci,sigmaci]

= normfit(X,alpha)

此命令在显著性水平alpha下估计数据X的参数（alpha缺省时设定为0.05），返回值muhat是X的均值的点估计值，sigmahat是标准差的点估计值, muci是均值的区间估计,sigmaci是标准差的区间估计.

2、其它分布的参数估计

有两种处理办法:

一.取容量充分大的样本（n>50），按中心极限定理，它近似地服从正态分布。

二.使用Matlab工具箱中具有特定分布总体的估计命令：

（1）[muhat, muci] = expfit(X,alpha)----- 在显著性水平alpha下，求指数分布的数据X的均值的点估计及其区间估计.

（2）[lambdahat, lambdaci] = poissfit(X,alpha)----- 在显著性水平alpha下，求泊松分布的数据X 的参数的点估计及其区间估计.

（3）[phat, pci] = weibfit(X,alpha)----- 在显著性水平alpha下，求Weibull分布的数据X 的参数的点估计及其区间估计.

5.假设检验

在总体服从正态分布的情况下，可用以下命令进行假设检验.

1、总体方差**sigma2**已知时，总体均值的检验使用 **z-检验**

[h,sig,ci] = ztest(x,mu,sigma,alpha,tail)

- tail = 0，检验假设“x 的均值等于 mu ”
- tail = 1，检验假设“x 的均值大于 mu ”
- tail =-1，检验假设“x 的均值小于 mu ”
- tail的缺省值为 0， alpha的缺省值为 0.05.

返回值 h 为一个布尔值，h=1 表示可以拒绝假设，h=0 表示不可以拒绝假设，sig 为假设成立的概率，ci 为均值的 1-alpha 置信区间.

2、总体方差**sigma2**未知时，总体均值的检验使用**t-**检验

[h,sig,ci] = ttest(x,m,alpha,tail)

3、两总体均值的假设检验使用 **t-**检验

[h,sig,ci] = ttest2(x,y,alpha,tail)

二、回归分析

回归分析是研究变量间相关关系的一种统计分析

1、回归分析的基本理论

2、用数学软件求解回归分析问题

1、回归分析的基本理论

一元线性回归模型： 正态误差下的一元线性回归模型是:

$$\begin{cases} Y = a + bx + \varepsilon; \\ \varepsilon \sim N(0, \sigma^2) \end{cases} \quad \text{或} \quad Y \sim N(a + bx, \sigma^2)$$

需要解决的问题：

1） 在回归模型中如何估计参数a、b和σ²?

2） 模型的假设是否正确？需要检验。

3）利用回归方程对试验指标y进行预测或控制?

估计量$\hat{y}_0 = \hat{a} + \hat{b}x_0$, 区间估计$(\hat{y}_0 - d, \hat{y}_0 + d)$

参数估计

设观测值为(x_i, y_i)（i=1,2,…,n), 代入模型中，$y_i = a + bx_i + \varepsilon_i$

最小二乘法:

$$\min Q(a,b) = \sum_{i=1}^{n}[y_i - (a + bx_i)]^2$$

解出的参数记为 $\hat{a}, \hat{b}$

则回归方程 $\hat{y} = \hat{a} + \hat{b}x$

$\hat{y}_i = \hat{a} + \hat{b}x_i$ $y_i - \hat{y}_i$残差值

回归模型的假设检验

模型： $Y = a + bx + \varepsilon$

提出问题： $H_0 : b = 0; H_1 : b \neq 0$

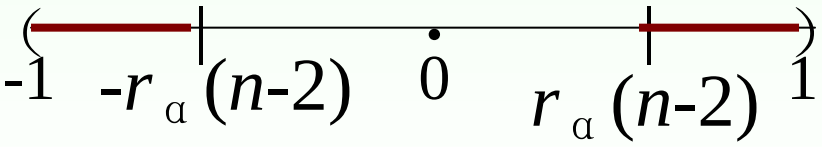
1、相关系数检验法

$$r = \frac{\mathrm{cov}(X,Y)}{\sqrt{DX \cdot DY}} \Leftarrow \hat{r} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

$|r| \leqslant 1$

$|r| \to 1$，线性相关

$|r| \to 0$，非线性相关



H_0的拒绝域为

$$\chi_0 = \{|\hat{r}| > r_\alpha(n-2)\}$$

2、F-检验法

平方和分解公式：

实测值

估计值

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

残差值，剩余平方和,越小越好

记为 $L_{yy}=Q+U$

$$F=\frac{U}{Q/(n-2)}\sim F(1,n-2)$$

拒绝域 $\chi_0=\{F>F_{1-\alpha}(1,n-2)\}$

认为线性回归效果好

预测与控制

给定的自变量x_0，给出$E(y_0)$的点估计量:

$$\hat{y}_0 = \hat{a} + \hat{b}x_0$$

y_0的置信度为$(1-\alpha)\%$的预测区间为:

$$(\hat{y}_0 - d_n, \hat{y}_0 + d_n)$$

$$d_n = t_{\frac{\alpha}{2}}(n-2)\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{L_{xx}}} \qquad \hat{\sigma}^2 = \frac{Q}{n-2}$$

设y在某个区间(y_1, y_2)取值时，应如何控制x的取值范围，这样的问题称为控制问题。

可线性化的一元非线性回归

需要配曲线，配曲线的一般方法是:

- 先对两个变量x和y 作n次试验观察得画出散点图。
- 根据散点图确定须配曲线的类型。
- 由n对试验数据确定每一类曲线的未知参数a和b采用的方法是通过变量代换把非线性回归化成线性回归，即采用非线性回归线性化的方法。

通常选择的六类拟合曲线如下：

（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 $y=\frac{1}{a+be^{-x}}$

多元线性回归模型

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.

2、用数学软件求解回归分析问题

例2.1 为了研究钢材消费量与国民收入之间的关系，在统计年鉴上查得一组历史数据。

年 份	1964	1965	1966	.....	1978	1979	1980
消费(吨)	698	872	988	.....	1446	2736	2825
收入(亿)	1097	1284	1502	.....	2948	3155	3372

试分析预测若1981年到1985年我国国民收入以4.5%的速度递增，钢材消费量将达到什么样的水平？

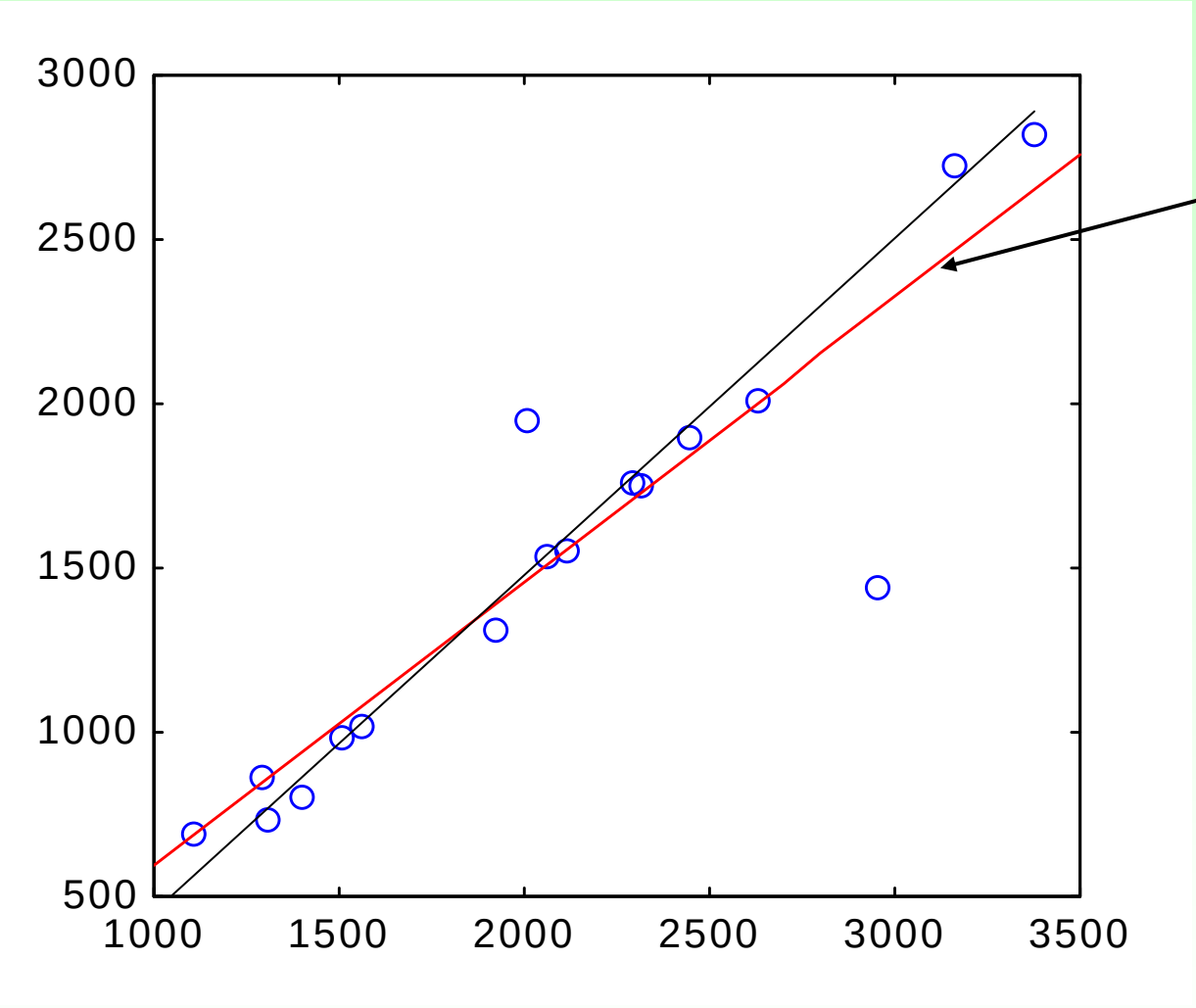
问题分析：

钢材消费量--------试验指标(因变量)Y;

国民收入----------自变量 x;

建立数据拟合函数 $y = \mathrm{E}(Y \mid x) = f(x)$;

作拟合曲线图形分析。



钢材消费量y与国民收入x的散点图

MATLAB软件实现

使用命令regress实现一元线性回归模型的计算

b = regress (Y, X)　或

[b, bint, r, rint, stats] = regress(Y, X, alpha)

回　相关系数R^2，F-统计量和与α_0对应的概率p。

$$\begin{bmatrix} 1 & x_n \end{bmatrix} \qquad \begin{bmatrix} y_n \end{bmatrix}$$

残差及其置信区间可以用rcoplot(r,rint)画图。

求解

输入: (eg2_1.m)

```
x=[1097 1284 1502 1394 1303 1555 1917 2051 2111
   2286 2311 2003 2435 2625 2948 3155  3372];
y=[698 872 988 807 738 1025 1316 1539 1561
   1765 1762 1960 1902 2013 2446 2736 2825];
X=[ones(size(x')),x'],pause
[c,cint,r,rint,stats]=regress(y',X,0.05),pause
rcoplot(r,rint)
```

输出：

$$\hat{y} = \hat{a} + \hat{b}x$$

c = -460.5282 (参数a) 0.9840 (参数b)

cint = -691.8478 -229.2085 (a的置信区间)

0.8779 1.0900 (b的置信区间)

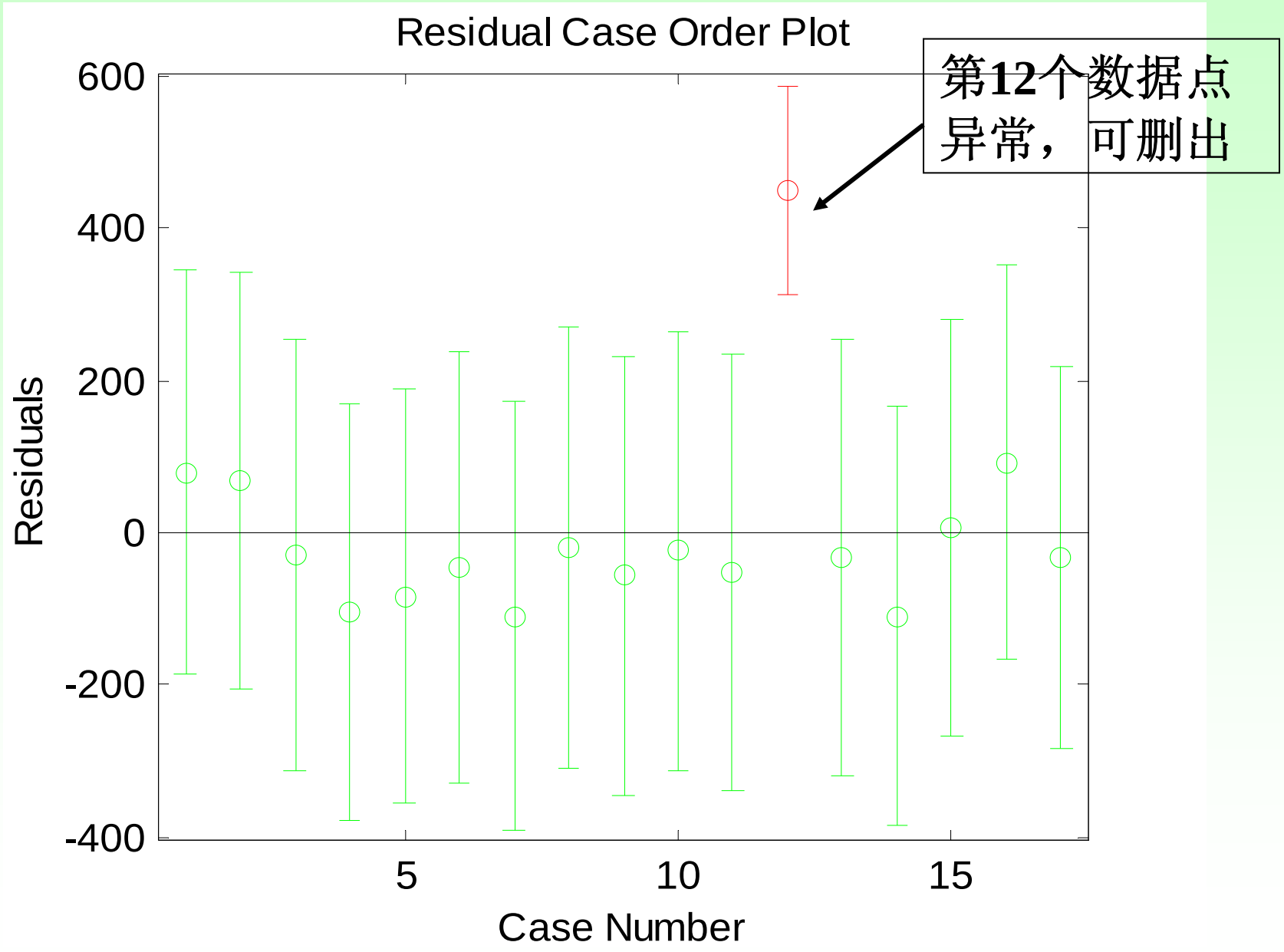
r = [79.1248 69.1244 -29.3788 -104.1112 -83.5709 -44.5286

-109.7219 -18.5724 -55.6100 -23.8029 -51.4019 449.6576

-33.4128 -109.3651 5.8160 92.1364 -32.3827]’(残差向量)

rint=（略）（参见下页残差分析图）

stats = 0.9631(R^2) 391.2713(F) 0.0000 (P{ χ_0})


Residual Case Order Plot
第12个数据点
异常，可删出
600
400
200
0
-200
-400
Residuals
5
10
15
Case Number

预测

```
x1(1)=3372;(hgy1.m)
for  i=1:5
    x1(i+1)=1.045*x1(i);%未来五年国民收入以4.5%的
                                速度递增
    y1(i+1)=-460.5282+0.9840*x1(i+1);%钢材的预
                                测值
end
x1, y1
```

结果

```
x1 = 3372.0   3523.7   3682.3   3848.0  4021.2   4202.1

y1 =  3006.8    3162.9    3325.9    3496.3    3674.4
```

三、概率模型

3.1 报童的诀窍

3.2 随机存贮策略

3.1 报童的诀窍

问题

报童售报： a (零售价) > b(购进价) > c(退回价)

售出一份赚 $a-b$；退回一份赔 $b-c$

每天购进多少份可使收入最大？

分析

购进太多→卖不完退回→赔钱

购进太少→不够销售→赚钱少

存在一个合适的购进量

应根据需求确定购进量

每天需求量是随机的 → 每天收入是随机的

优化问题的目标函数应是长期的日平均收入

等于每天收入的期望

准备

调查需求量的随机规律——每天需求量为 r 的概率 $f(r)$，$r=0,1,2\ldots$

建模

- 设每天购进 n 份，日平均收入为 $G(n)$
- 已知售出一份赚 $a-b$；退回一份赔 $b-c$

$$r \le n \Rightarrow \text{售出 } r \Rightarrow \text{赚}(a-b)r$$

$$\Rightarrow \text{退回 } n-r \Rightarrow \text{赔}(b-c)(n-r)$$

$$r > n \Rightarrow \text{售出} n \Rightarrow \text{赚}(a-b)n$$

$$G(n)=\sum_{r=0}^{n}[(a-b)r-(b-c)(n-r)]f(r)+\sum_{r=n+1}^{\infty}(a-b)nf(r)$$

求 n 使 $G(n)$ 最大

求解　将*r*视为连续变量　$f(r) \Rightarrow p(r)$（概率密度）

$$G(n)=\int_0^n[(a-b)r-(b-c)(n-r)]p(r)dr+\int_n^{\infty}(a-b)np(r)dr$$

$$\frac{dG}{dn}=(a-b)np(n)-\int_0^n(b-c)\,p(r)dr$$

$$-(a-b)np(n)+\int_n^{\infty}(a-b)\,p(r)dr$$

$$=-(b-c)\int_0^n p(r)dr+(a-b)\int_n^{\infty}p(r)dr$$

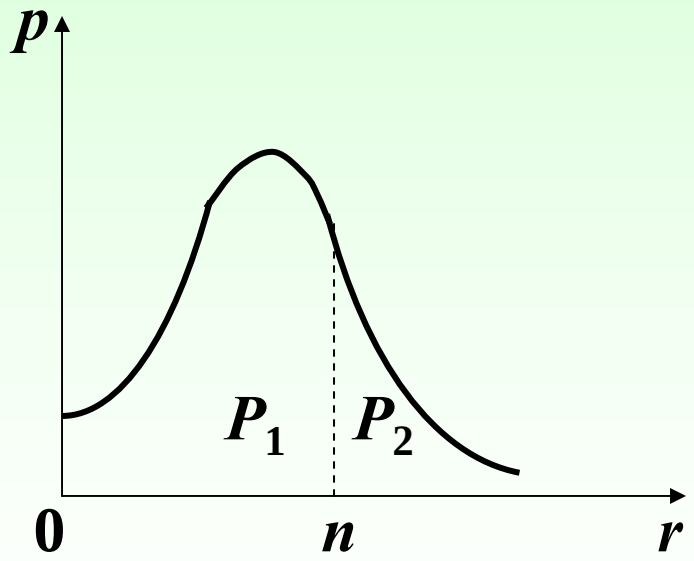
$$\frac{dG}{dn}=0 \quad \Rightarrow \quad \frac{\int_0^n p(r)\,dr}{\int_n^{\infty} p(r)\,dr}=\frac{a-b}{b-c}$$

结果解释

$$\frac{\int_0^n p(r)\,dr}{\int_n^\infty p(r)\,dr} = \frac{a-b}{b-c}$$

$$\int_0^n p(r)dr = P_1, \int_n^\infty p(r)dr = P_2$$

取***n***使 $\dfrac{P_1}{P_2} = \dfrac{a-b}{b-c}$



a-b ~售出一份赚的钱

b-c ~退回一份赔的钱

$$(a-b)\uparrow \Rightarrow n\uparrow, \quad (b-c)\uparrow \Rightarrow n\downarrow$$

3.2 随机存贮策略

问题 以周为时间单位；一周的商品销售量为随机；
周末根据库存决定是否订货，供下周销售。

（*s*, *S*) 存贮策略

制订下界*s*, 上界*S*，当周末库存小于*s* 时订货，使下周初的库存达到*S*；否则，不订货。

考虑订货费、存贮费、缺货费、购进费，制订（*s*, *S*) 存贮策略, 使(平均意义下) 总费用最小

模型假设

- 每次订货费c_0, 每件商品购进价c_1,每件商品一周贮存费c_2,每件商品缺货损失费c_3 ($c_1<c_3$)

- 每周销售量 r 随机、连续，概率密度 $p(r)$

- 周末库存量x, 订货量 u, 周初库存量 $x+u$

- 每周贮存量按 $x+u-r$ 计

建模与求解

（*s*, *S*) 存贮策略

$$x \geq s \Rightarrow u = 0 \qquad x < s \Rightarrow u > 0,\ x + u = S$$

确定(*s*, *S*), 使目标函数——每周总费用的平均值最小

订货费c_0, 购进价c_1, 贮存费c_2, 缺货费c_3, 销售量 r

平均费用

$$J(u) = \begin{cases} c_0 + c_1 u + L(x+u), & u > 0 \\ L(x) & u = 0 \end{cases}$$

$$L(x) = c_2 \int_0^x (x - r)\, p(r)\, dr + c_3 \int_x^{\infty} (r - x)\, p(r)\, dr$$

建模与求解

$$J(u)=\begin{cases} c_0+c_1u+L(x+u), & u>0 \\ L(x) & u=0 \end{cases}$$

$$L(x)=c_2\int_0^x(x-r)p(r)dr+c_3\int_x^\infty(r-x)p(r)dr$$

1）设 $x<s$, 求 u 使 $J(u)$ 最小，确定S

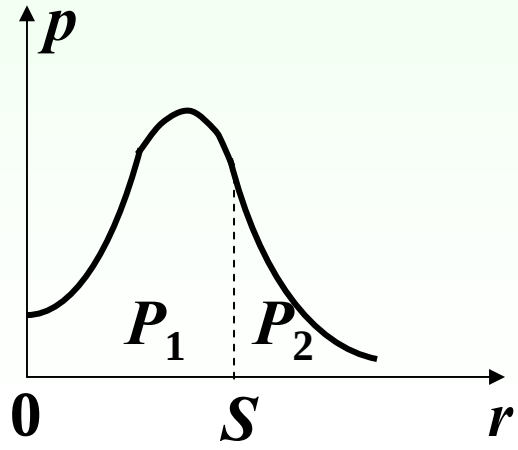
$$\frac{dJ}{du}=c_1+c_2\int_0^{x+u}p(r)dr-c_3\int_{x+u}^{\infty}p(r)dr$$

$x+u=S$

$\int_0^\infty p(r)dr=1$ ⇨

$$=(c_1+c_2)\int_0^S p(r)dr-(c_3-c_1)\int_S^\infty p(r)dr$$

$$\frac{dJ}{du}=0 \Rightarrow \frac{\int_0^S p(r)dr}{\int_S^\infty p(r)dr}=\frac{c_3-c_1}{c_2+c_1}=\frac{P_1}{P_2}$$



$$c_3\uparrow \Rightarrow S\uparrow,\quad c_2\uparrow \Rightarrow S\downarrow$$

建模与求解

$$J(u)=\begin{cases} c_0+c_1u+L(x+u), & u>0 \\ L(x) & u=0 \end{cases}$$

$$L(x)=c_2\int_0^x(x-r)p(r)dr+c_3\int_x^\infty(r-x)p(r)dr$$

2）对库存 x，确定s

若订货u, $u+x=S$, 总费用为 $J_1=c_0+c_1(S-x)+L(S)$

若不订货, $u=0$, 总费用为 $J_2=L(x)$

$$J_2\le J_1 \Leftrightarrow L(x)\le c_0+c_1(S-x)+L(S)$$

$$\Downarrow$$

$$c_1x+L(x)\le c_0+c_1S+L(S)$$

$$\Downarrow$$

记 $c_1x+L(x)=I(x)$

$$I(x)\le c_0+I(S)$$

s 是 $I(x)=c_0+I(S)$ 的最小正根

建模与求解

$I(x)=c_0+I(S)$ 最小正根的图解法

$$J(u)=\begin{cases} c_0+c_1u+L(x+u), & u>0 \\ L(x) & u=0 \end{cases}$$

$$I(x)=c_1x+L(x)$$

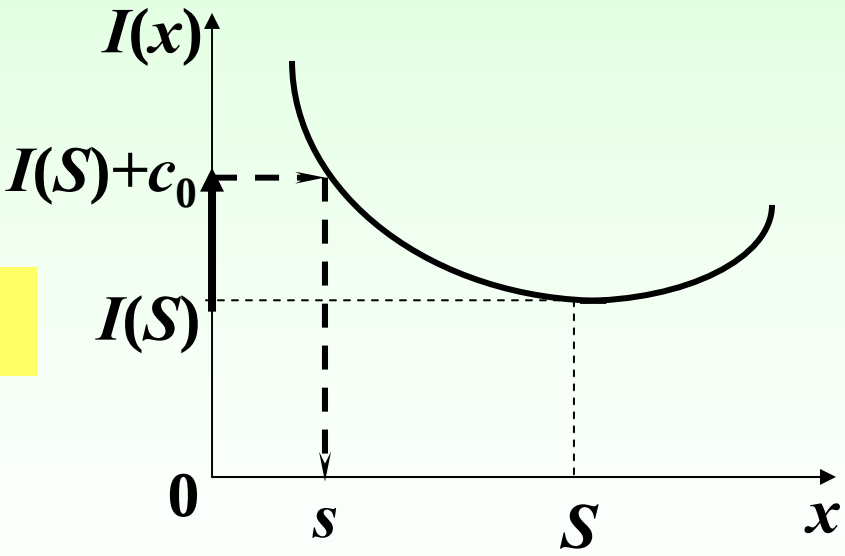
$$L(x)=c_2\int_0^x(x-r)p(r)dr+c_3\int_x^\infty(r-x)p(r)dr$$

***J(u)*在*u*+*x*=*S*处达到最小**

***I(x)* 与*J(u)*相似** ⇩

I(x)*在*x*=*S*处达到最小值*I(S)

I(x)*图形** ⇨ ***I(S)



⇨ $I(x)=c_0+I(S)$ 的最小正根 *s*

四　统计回归模型

4.1　牙膏的销售量

（回归模型）

4.2　投资额与国民生产总值和物价指数

（时间序列模型）

4.1 牙膏的销售量

问题 建立牙膏销售量与价格、广告投入之间的模型，预测在不同价格和广告费用下的牙膏销售量 。

收集了30个销售周期内本公司牙膏销售量、价格、广告费用及同期其它厂家同类牙膏的平均售价

销售周期	本公司价格(元)	其它厂家价格(元)	广告费用(百万元)	价格差（元）	销售量(百万支)
1	3.85	3.80	5.50	-0.05	7.38
2	3.75	4.00	6.75	0.25	8.51
…	…	…	…	…	…
29	3.80	3.85	5.80	0.05	7.93
30	3.70	4.25	6.80	0.55	9.26

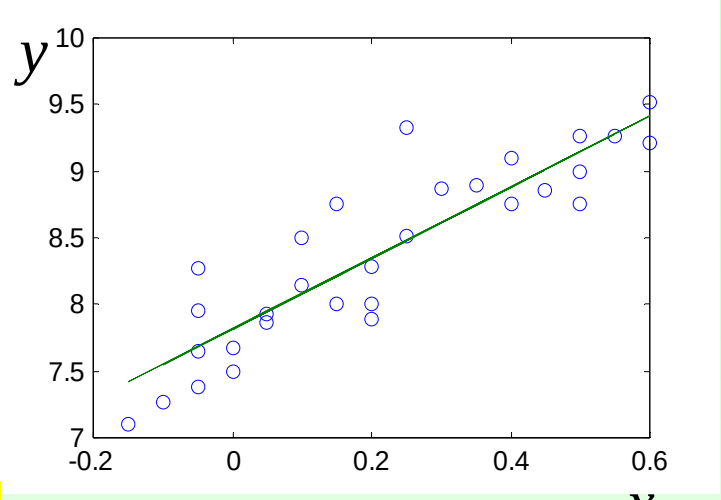
基本模型

y ~公司牙膏销售量

x_1~其它厂家与本公司价格差

x_2~公司广告费用

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_2^2 + \varepsilon$$



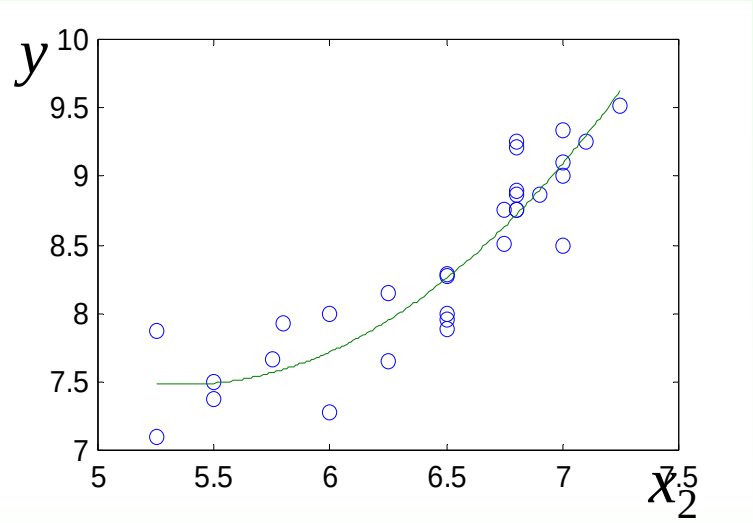
$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

y~被解释变量（因变量）

x_1, x_2~解释变量(回归变量, 自变量)

$\beta_0, \beta_1, \beta_2, \beta_3$ ~回归系数

ε~随机误差（均值为零的正态分布随机变量）



$$y = \beta_0 + \beta_1 x_2 + \beta_2 x_2^2 + \varepsilon$$

模型求解

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_2^2 + \varepsilon$　由数据 y, x_1, x_2 估计 β

[b,bint,r,rint,stats]=regress(y,x,alpha)

输入　**y~n维数据向量**

x= $[1 \quad x_1 \quad x_2 \quad x_2^2]$ **~n×4数据矩阵, 第1列为全1向量**

输出　**b~β的估计值**

bint~b的置信区间

r ~残差向量y-xb

rint~r的置信区间

参数	参数估计值	置信区间
β_0	17.3244	[5.7282　28.9206]
β_1	1.3070	[0.6829　1.9311]
β_2	-3.6956	[-7.4989　0.1077]
β_3	0.3486	[0.0379　0.6594]
R^2=0.9054　F=82.9409　p=0.0000		

Stats~

检验统计量

R^2, F, p

结果分析

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_2^2 + \varepsilon$$

参数	参数估计值	置信区间
β_0	17.3244	[5.7282 28.9206]
β_1	1.3070	[0.6829 1.9311]
β_2	-3.6956	[-7.4989 0.1077]
β_3	0.3486	[0.0379 0.6594]
R^2=0.9054	F=82.9409	p=0.0000

y的90.54%可由模型确定

F远超过F检验的临界值

p远小于α=0.05

模型从整体上看成立

β_2的置信区间包含零点
(右端点距零点很近)

x_2对因变量y 的影响不太显著

x_2^2项显著

可将x_2保留在模型中

销售量预测

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_2^2$$

价格差x_1=其它厂家价格x_3-本公司价格x_4

估计x_3 调整x_4 ⇨ 控制x_1 ⇨ 通过x_1, x_2预测y

控制价格差x_1=0.2元，投入广告费x_2=650万元

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_2^2 = 8.2933 \text{ (百万支)}$$

销售量预测区间为 [7.8230，8.7636]（置信度95%）

上限用作库存管理的目标值　下限用来把握公司的现金流

若估计x_3=3.9，设定x_4=3.7，则可以95%的把握知道销售额在 7.8320×3.7≈ 29（百万元）以上

模型改进

x_1和x_2对y的影响独立

x_1和x_2对y的影响有交互作用

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_2^2 + \varepsilon$$

参数	参数估计值	置信区间
β_0	17.3244	[5.7282 28.9206]
β_1	1.3070	[0.6829 1.9311]
β_2	-3.6956	[-7.4989 0.1077]
β_3	0.3486	[0.0379 0.6594]
R^2=0.9054	F=82.9409	p=0.0000

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_2^2 + \beta_4 x_1 x_2 + \varepsilon$$

参数	参数估计值	置信区间
β_0	29.1133	[13.7013 44.5252]
β_1	11.1342	[1.9778 20.2906]
β_2	-7.6080	[-12.6932 -2.5228]
β_3	0.6712	[0.2538 1.0887]
β_4	-1.4777	[-2.8518 -0.1037]
R^2=0.9209	F=72.7771	p=0.0000

两模型销售量预测比较

控制价格差x_1=0.2元，投入广告费x_2=6.5百万元

$$\hat{y}=\hat{\beta}_0+\hat{\beta}_1x_1+\hat{\beta}_2x_2+\hat{\beta}_3x_2^2$$

$\hat{y}=8.2933$ (百万支)

区间 [7.8230，8.7636]

$$\hat{y}=\beta_0+\hat{\beta}_1x_1+\hat{\beta}_2x_2+\hat{\beta}_3x_2^2+\hat{\beta}_4x_1x_2$$

$\hat{y}=8.3272$ (百万支)

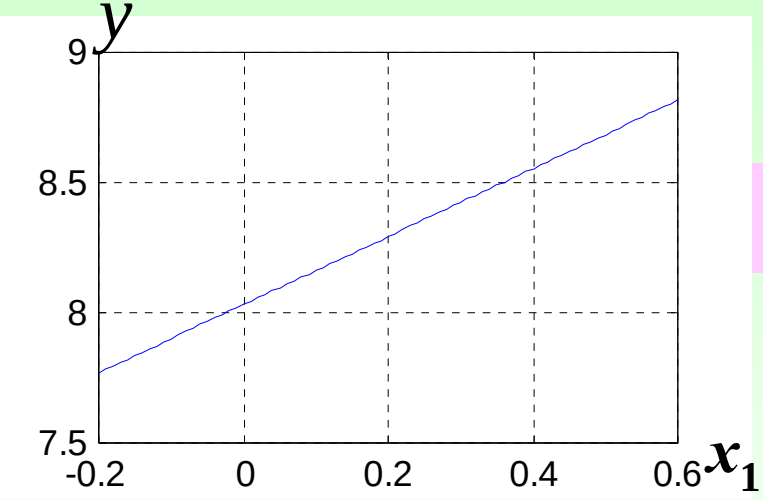
区间 [7.8953，8.7592]

$\hat{y}$ 略有增加

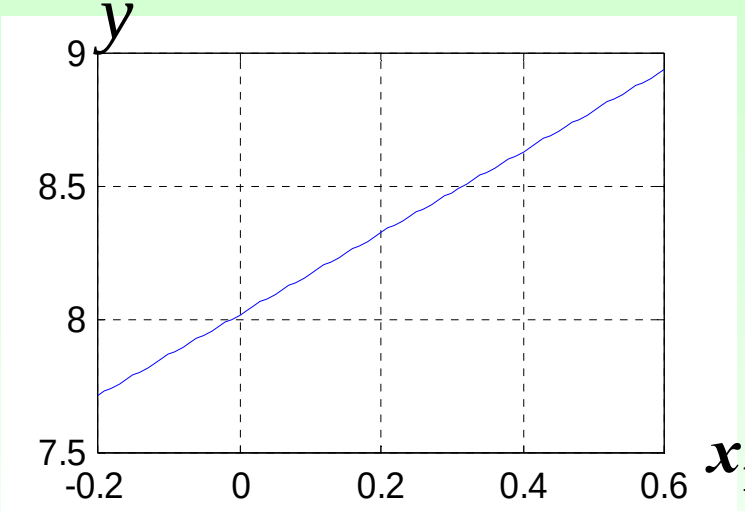
预测区间长度更短

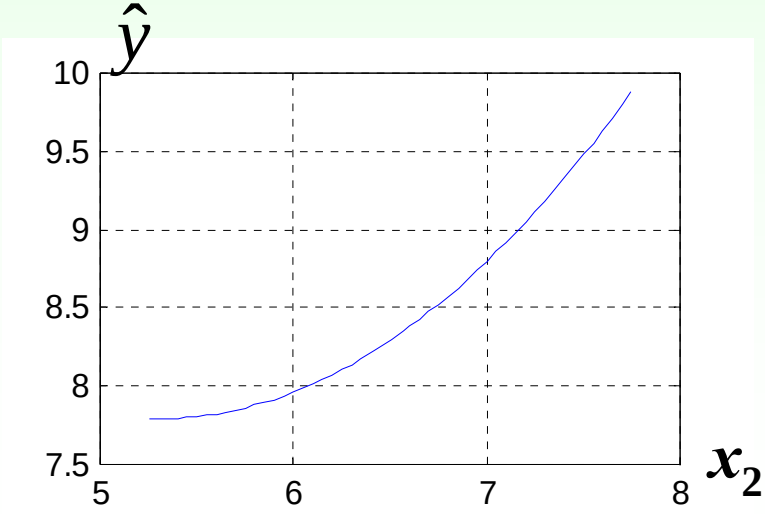
两模型 $\hat{y}$ 与x_1, x_2关系的比较

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_2^2 \qquad \hat{y} = \beta_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_2^2 + \hat{\beta}_4 x_1 x_2$$

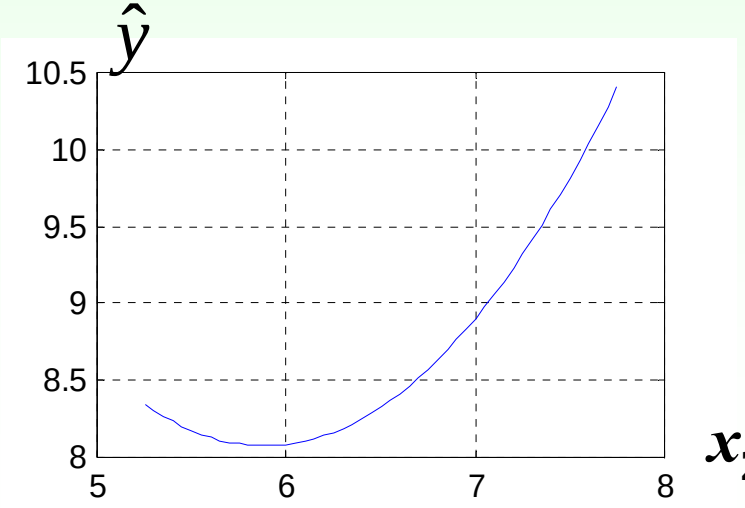


x_2=6.5





x_1=0.2



交互作用影响的讨论

$$\hat{y} = \beta_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_2^2 + \hat{\beta}_4 x_1 x_2$$

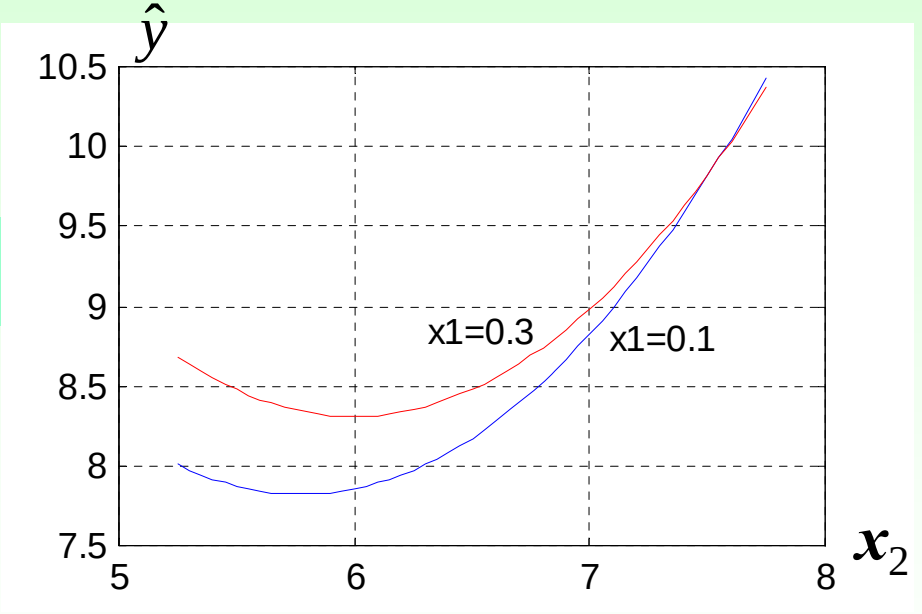
价格差 x_1=0.1

$$\hat{y}\big|_{x_1=0.1} = 30.2267 - 7.7558\,x_2 + 0.6712\,x_2^2$$

价格差 x_1=0.3

$$\hat{y}\big|_{x_1=0.3} = 32.4535 - 8.0513\,x_2 + 0.6712\,x_2^2$$

$$x_2 < 7.5357 \Rightarrow \hat{y}\big|_{x_1=0.3} > \hat{y}\big|_{x_1=0.1}$$



⇨ 价格优势会使销售量增加

加大广告投入使销售量增加
（x_2大于6百万元）

价格差较小时增加的速率更大 ⇨ 价格差较小时更需要靠广告来吸引顾客的眼球

4.2 投资额与国民生产总值和物价指数

问题

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

该地区连续20年的统计数据

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
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

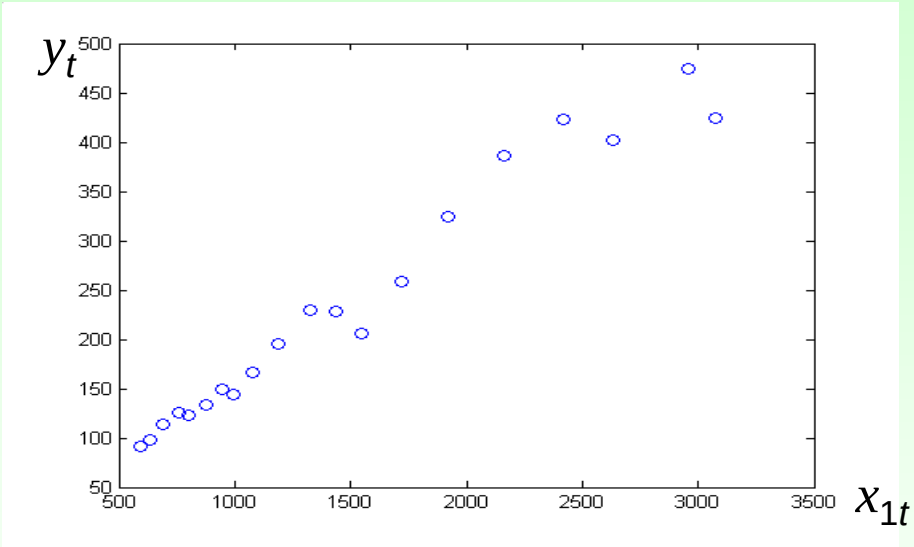
时间序列中同一变量的顺序观测值之间存在自相关

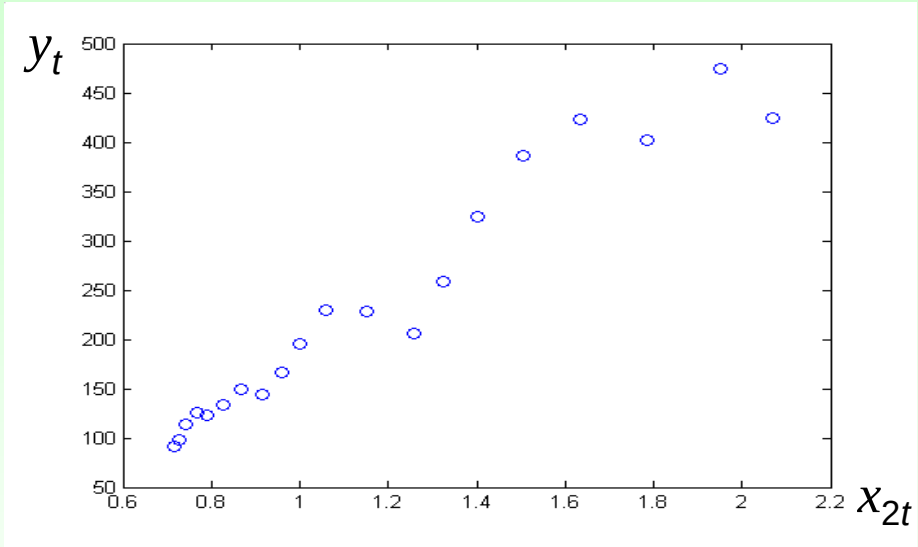
若采用普通回归模型直接处理，将会出现不良后果

需要诊断并消除数据的自相关性，建立新的模型

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

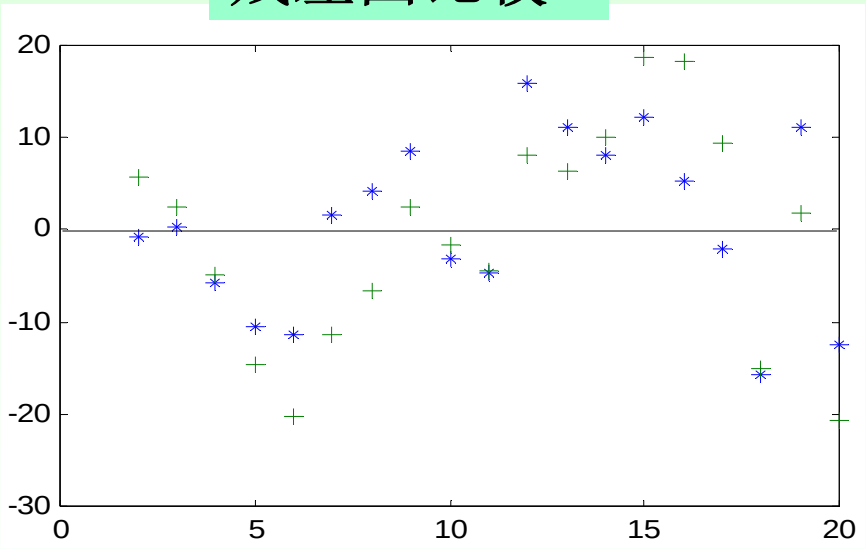
ε_t ~对t相互独立的零均值正态随机变量

模型结果比较

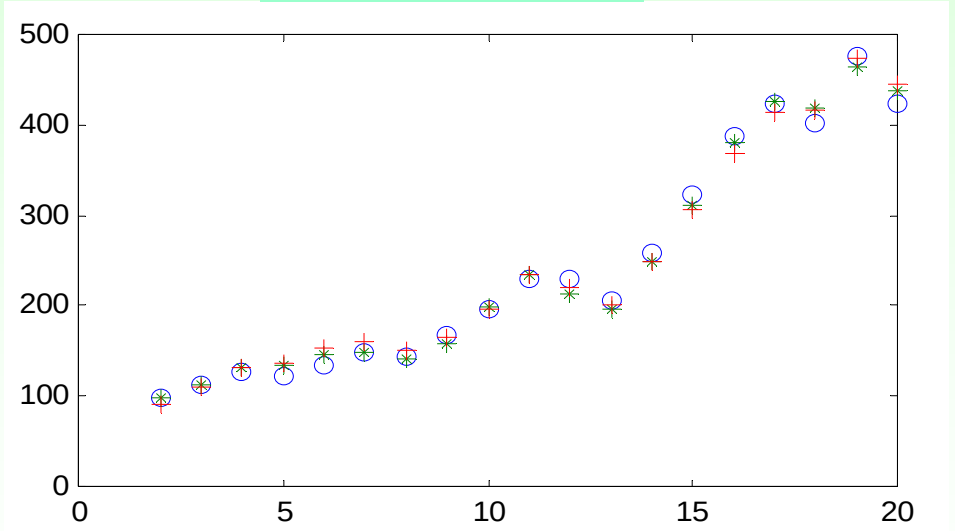
基本回归模型 $\hat{y}_t = 322.725 + 0.6185x_{1t} - 859.479x_{2t}$

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

五　马氏链模型

5.1　健康与疾病

5.2　钢琴销售的存贮策略

马氏链模型

马氏链模型描述一类重要的随机动态系统（过程）的模型

- 系统在每个时期所处的状态是随机的
- 从一时期到下时期的状态按一定概率转移
- 下时期状态只取决于本时期状态和转移概率

已知现在，将来与过去无关（无后效性）

马氏链 (Markov Chain)

——时间、状态均为离散的随机转移过程

5.1 健康与疾病

人的健康状态随着时间的推移会随机地发生转变

保险公司要对投保人未来的健康状态作出估计, 以制订保险金和理赔金的数额

例5.1 人的健康状况分为健康和疾病两种状态，设对特定年龄段的人，今年健康、明年保持健康状态的概率为0.8, 而今年患病、明年转为健康状态的概率为0.7，

若某人投保时健康, 问10年后他仍处于健康状态的概率

状态与状态转移

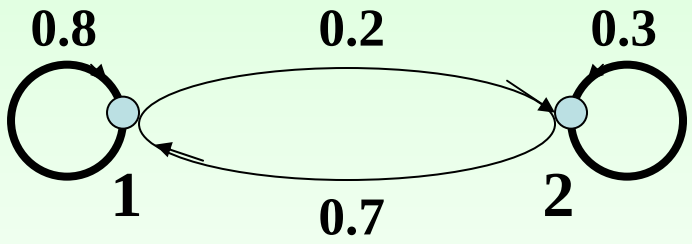
状态 $X_n=\begin{cases}1, & \text{第}n\text{年健康}\\ 2, & \text{第}n\text{年疾病}\end{cases}$　状态概率 $a_i(n)=P(X_n=i),$

$i=1,2,\ n=0,1,\cdots$

转移概率 $p_{ij}=P(X_{n+1}=j\,|\,X_n=i),\ i,j=1,2,\ n=0,1,\cdots$

$p_{11}=0.8 \quad p_{12}=1-p_{11}=0.2$

$p_{21}=0.7 \quad p_{22}=1-p_{21}=0.3$



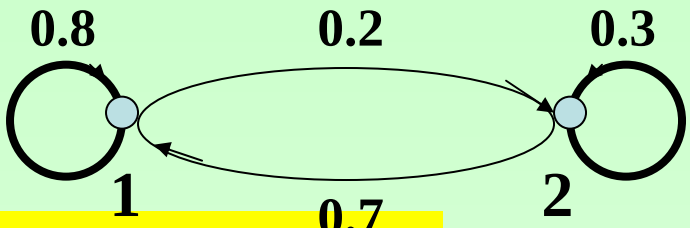
X_{n+1}只取决于X_n和p_{ij}, 与X_{n-1}, …无关

状态转移具有无后效性

$$a_1(n+1)=a_1(n)p_{11}+a_2(n)p_{21}$$

$$a_2(n+1)=a_1(n)p_{12}+a_2(n)p_{22}$$

状态与状态转移



$$\begin{cases} a_1(n+1) = a_1(n)\,p_{11} + a_2(n)\,p_{21} \\ a_2(n+1) = a_1(n)\,p_{12} + a_2(n)\,p_{22} \end{cases}$$

给定$a(0)$, 预测$a(n)$, n=1,2…

设投保时健康

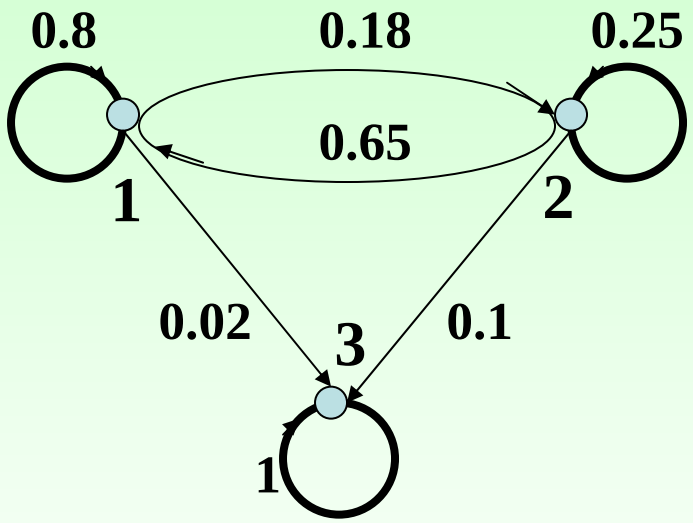
n	0	1	2	3	…	∞
$a_1(n)$	1	0.8	0.78	0.778	…	7/9
$a_2(n)$	0	0.2	0.22	0.222	…	2/9

设投保时疾病

$a_1(n)$	0	0.7	0.77	0.777	…	7/9
$a_2(n)$	1	0.3	0.33	0.333	…	2/9

$n\to\infty$时状态概率趋于稳定值，稳定值与初始状态无关

例5.2 健康和疾病状态同上，X_n=1~ 健康，X_n=2~ 疾病死亡为第3种状态，记Xn=3



p_{11}=0.8, p_{12}=0.18, p_{13}=0.02

p_{21}=0.65, p_{22}=0.25, p_{23}=0.1

p_{31}=0, p_{32}=0, p_{33}=1

$$a_1(n+1) = a_1(n)\,p_{11} + a_2(n)\,p_{21} + a_3(n)\,p_{31}$$

$$a_2(n+1) = a_1(n)\,p_{12} + a_2(n)\,p_{22} + a_3(n)\,p_{32}$$

$$a_3(n+1) = a_1(n)\,p_{13} + a_2(n)\,p_{23} + a_3(n)\,p_{33}$$

状态与状态转移

设投保时处于健康状态，预测 $a(n)$, n=1,2…

n	0	1	2	3	…	50	…	∞
$a_1(n)$	1	0.8	0.757	0.7285	…	0.1293	…	0
$a_2(n)$	0	0.18	0.189	0.1835	…	0.0326	…	0
$a_3(n)$	0	0.02	0.054	0.0880	…	0.8381	…	1

- 不论初始状态如何，最终都要转到状态3；
- 一旦$a_1(k)$= $a_2(k)$=0, $a_3(k)$=1, 则对于$n>k$, $a_1(n)$=0, $a_2(n)$=0, $a_3(n)$=1, 即从状态3不会转移到其它状态。

马氏链的基本方程

状态 $X_n = 1,2,\cdots k \quad (n = 0,1,\cdots)$

状态概率 $a_i(n) = P(X_n = i),$

$i = 1,2,\cdots,k,\ n = 0,1,\cdots$

$$\sum_{i=1}^{k} a_i(n) = 1$$

转移概率 $p_{ij} = P(X_{n+1} = j \,|\, X_n = i)$

$$p_{ij} \geq 0, \sum_{j=1}^{k} p_{ij} = 1, i = 1,2,\cdots,k$$

基本方程

$$a_i(n+1) = \sum_{j=1}^{k} a_j(n)\, p_{ji}, \quad i = 1,2,\cdots,k$$

$a(n) = (a_1(n), a_2(n), \cdots, a_k(n))$

~ 状态概率向量

$P = \{p_{ij}\}_{k\times k}$ ~ 转移概率矩阵

（非负，行和为1）

$$a(n+1) = a(n)P$$

$$a(n) = a(0)P^n$$

马氏链的两个重要类型

$$a(n+1)=a(n)P$$

1. 正则链 ~ 从任一状态出发经有限次转移能以正概率到达另外任一状态（如例**1**）。

$$\text{正则链} \Leftrightarrow \exists N, P^N > 0$$

$$\text{正则链} \Rightarrow \exists w,\ a(n) \to w (n \to \infty)$$

w ~ 稳态概率

w满足 $wP = w$

例1. $P = \begin{bmatrix} 0.8 & 0.2 \\ 0.7 & 0.3 \end{bmatrix}$

$$\begin{cases} 0.8w_1 + 0.7w_2 = w_1 \\ 0.2w_1 + 0.3w_2 = w_2 \end{cases} \Rightarrow 0.2w_1 = 0.7w_2$$

w满足 $\sum_{i=1}^{k} w_i = 1$

$$w_1 + w_2 = 1 \Rightarrow w = (7/9, 2/9)$$

马氏链的两个重要类型

2. 吸收链 ~ 存在吸收状态（一旦到达就不会离开的状态i, $p_{ii}=1$）,且从任一非吸收状态出发经有限次转移能以正概率到达吸收状态（如例**2**）。

有r个吸收状态的吸收链的转移概率阵标准形式

$$P=\begin{bmatrix} I_{r\times r} & 0 \\ R & Q \end{bmatrix}$$

R有非零元素

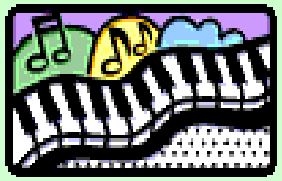
$$M=(I-Q)^{-1}=\sum_{s=0}^{\infty}Q^{s}$$

$$y=(y_1,y_2,\cdots y_{k-r})=Me$$

$$e=(1,1,\cdots,1)^{T}$$

y_i~ 从第 i 个非吸收状态出发，被某个吸收状态吸收前的平均转移次数。

5.2 钢琴销售的存贮策略



背景与问题

钢琴销售量很小，商店的库存量不大以免积压资金

一家商店根据经验估计，平均每周的钢琴需求为1架

存贮策略：每周末检查库存量，仅当库存量为零时，才订购3架供下周销售；否则，不订购。

估计在这种策略下失去销售机会的可能性有多大，以及每周的平均销售量是多少。

问题分析

顾客的到来相互独立，需求量近似服从波松分布，其参数由需求均值为每周1架确定，由此计算需求概率

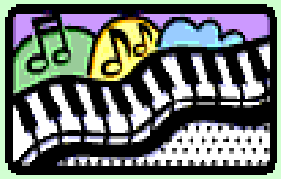
存贮策略是周末库存量为零时订购3架 →周末的库存量可能是0, 1, 2, 3，周初的库存量可能是1, 2, 3。

用马氏链描述不同需求导致的周初库存状态的变化。

动态过程中每周销售量不同，失去销售机会（需求超过库存）的概率不同。

可按稳态情况（时间充分长以后）计算失去销售机会的概率和每周的平均销售量。

模型假设



钢琴每周需求量服从泊松分布，均值为每周1架

存贮策略：当周末库存量为零时，订购3架，周初到货；否则，不订购。

以每周初的库存量作为状态变量，状态转移具有无后效性。

在稳态情况下计算该存贮策略失去销售机会的概率，和每周的平均销售量。

模型建立

D_n~第n周需求量，均值为1的泊松分布

$$P(D_n = k) = e^{-1}/k! \quad (k = 0,1,2\cdots)$$

D_n	0	1	2	3	>3
P	0.368	0.368	0.184	0.061	0.019

S_n~第n周初库存量(状态变量) $S_n \in \{1,2,3\}$

状态转移规律

$$S_{n+1} = \begin{cases} S_n - D_n, & D_n < S_n \\ 3, & D_n \geq S_n \end{cases}$$

状态转移阵

$$P = \begin{bmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{bmatrix}$$

$$p_{11} = P(S_{n+1} = 1 \mid S_n = 1) = P(D_n = 0) = 0.368$$

$$p_{12} = P(S_{n+1} = 2 \mid S_n = 1) = 0$$

$$p_{13} = P(S_{n+1} = 3 \mid S_n = 1) = P(D_n \geq 1) = 0.632$$

…… ……

$$p_{33} = P(S_{n+1} = 3 \mid S_n = 3) = P(D_n = 0) + P(D_n \geq 3) = 0.448$$

$$= \begin{bmatrix} 0.368 & 0 & 0.632 \\ 0.368 & 0.368 & 0.264 \\ 0.184 & 0.368 & 0.448 \end{bmatrix}$$

模型建立

状态概率 $a_i(n) = P(S_n = i),\ i = 1,2,3$

马氏链的基本方程 $a(n+1) = a(n)P$

$$P = \begin{bmatrix} 0.368 & 0 & 0.632 \\ 0.368 & 0.368 & 0.264 \\ 0.184 & 0.368 & 0.448 \end{bmatrix}$$

已知初始状态，可预测第 n 周初库存量 $S_n = i$ 的概率

正则链 $\Leftrightarrow \exists N, P^N > 0$ $\quad P^2 > 0$ ⇨ 正则链

⇨ 稳态概率分布 w 满足 $wP = w$

$$w = (w_1, w_2, w_3) = (0.285, 0.263, 0.452)$$

$n \to \infty$, 状态概率 $a(n) = (0.285, 0.263, 0.452)$

模型求解

1. 估计在这种策略下失去销售机会的可能性

第*n*周失去销售机会的概率

$$P(D_n > S_n) = \sum_{i=1}^{3} P(D_n > i \mid S_n = i) P(S_n = i)$$

*n*充分大时

$$P(S_n = i) = w_i$$

$$= P(D>1)w_1 + P(D>2)w_2 + P(D>3)w_3$$

$$= 0.264 \times 0.285 + 0.080 \times 0.263 + 0.019 \times 0.452 = 0.105$$

D	0	1	2	3	>3
P	0.368	0.368	0.184	0.061	0.019

$$w = (0.285, 0.263, 0.452)$$

从长期看，失去销售机会的可能性大约 **10%**。

模型求解

2. 估计这种策略下每周的平均销售量

第n周平均售量

需求不超过存量,销售需求　　需求超过存量,销售存量

$$R_n=\sum_{i=1}^{3}[\sum_{j=1}^{i} jP(D_n=j,S_n=i)+iP(D_n>i,S_n=i)]$$

$$=\sum_{i=1}^{3}[\sum_{j=1}^{i} jP(D_n=j\,|\,S_n=i)+iP(D_n>i\,|\,S_n=i)]P(S_n=i)$$

$$=0.632\times0.285+0.896\times0.263+0.977\times0.452=0.857$$

n充分大时　$P(S_n=i)=w_i$

从长期看，每周的平均销售量为 **0.857(架)**

敏感性分析

当平均需求在每周1 (架) 附近波动时，最终结果有多大变化。

设D_n服从均值为λ的泊松分布

$$P(D_n=k)=\lambda^k e^{-\lambda}/k!,\ (k=0,1,2\cdots)$$

状态转移阵

$$P=\begin{bmatrix} e^{-\lambda} & 0 & 1-e^{-\lambda} \\ \lambda e^{-\lambda} & e^{-\lambda} & 1-(1+\lambda)e^{-\lambda} \\ \lambda^2 e^{-\lambda}/2 & \lambda e^{-\lambda} & 1-(\lambda+\lambda^2/2)e^{-\lambda} \end{bmatrix}$$

第n周(n充分大)失去销售机会的概率 $P=P(D_n>S_n)$

λ	0.8	0.9	1.0	1.1	1.2
P	0.073	0.089	0.105	0.122	0.139

当平均需求增长（或减少）10%时，失去销售机会的概率将增长（或减少）约12% 。