第1讲 预测预报方法及其应用

汪晓银 教授

课后辅导微博 **http:// weibo.com/wxywxq**

1.1 一元线性回归

1.1.1 一元线性回归参数估计

一元线性回归可用来分析自变量$\boldsymbol{x}$取值与因变量$\boldsymbol{Y}$取值的内在联系，不过这里的自变量$\boldsymbol{x}$是确定性的变量，因变量Y是随机性的变量。

进行n次独立试验，测得数据如下：

X	x_1	x_2	$\cdots$	x_n
Y	y_1	y_2	$\cdots$	y_n

1.1 一元线性回归

力图建立回归方程的估计式或经验回归方程

$$\hat{\boldsymbol{y}} = \hat{\alpha} + \hat{\beta}\boldsymbol{x},$$

$\hat{\alpha} = a, \hat{\beta} = b$及$\hat{y}_i = a + bx_i$使

使用最小二乘法进行参数估计

$$\boldsymbol{Q} = \sum_{i=1}^{n} (\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i)^2 = \sum_{i=1}^{n} [\boldsymbol{y}_i - (\boldsymbol{a} + \boldsymbol{b}\boldsymbol{x}_i)]^2$$

的值最小，所求出的a称为经验截距，简称为截距，b称为经验回归系数，简称为回归系数。

1.1 一元线性回归

根据最小二乘法的要求由

$$\frac{\partial Q}{\partial a}=0,\frac{\partial Q}{\partial b}=0,\text{得}$$

$$a=\overline{Y}-b\overline{x},b=\frac{l_{xy}}{l_{xx}},\quad l_{xy}=\sum(x-\overline{x})(y-\overline{y})$$

$$l_{xy}=\sum(x-\overline{x})^2$$

1.1 一元线性回归

1.1.2　一元回归方程检验

(1) F检验法： $\dfrac{SSE}{\sigma^2} \sim \chi^2(n-2)$,

当H$_0$为真时， $\dfrac{SSE}{\sigma^2} \sim \chi^2(1)$;

且SSR与SSE相互独立；因此，当H_0为真时，

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

当$F \geqslant F_{1-\alpha}(1,n-2)$时应该放弃原假设$H_0$。

1.1 一元线性回归

(2) t检验法：

$$\because b \sim N(\beta, \frac{\sigma^2}{l_{xx}}), \frac{SSE}{\sigma^2} \sim \chi^2(n-2),$$

当H_0为真时，

$$t = b\sqrt{\frac{l_{xx}}{SSE/(n-2)}} \sim T(n-2),$$

当$|t| \geqslant t_{1-0.5\alpha}(n-2)$时应该放弃原假设$H_0$。

1.1 一元线性回归

(3) **r检验法**： 根据x与Y的观测值的相关系数

$$r=\frac{l_{xy}}{\sqrt{l_{xx}l_{yy}}},r^2=\frac{l_{xy}^2}{l_{xx}l_{yy}},$$

可以推出 $$r^2=\frac{SSR}{SST}.$$

当H_0为真时，

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

1.1 一元线性回归

当$F \geqslant F_{1-\alpha}(1,n-2)$或$|r| \geqslant r_{\alpha}(n-2)$时应该放弃原假设$H_0$，式中的

$$r_{\alpha}(n-2)=\sqrt{\frac{F_{1-\alpha}(1,n-2)}{F_{1-\alpha}(1,n-2)+(n-2)}}$$

可由r检验用表中查出。

$$\because \ r^2=\frac{SSR}{SST},$$

因此，r常常用来表示x与Y的线性关系在x与Y的全部关系中所占的百分比，又称为x与Y的观测值的决定系数。

1.1 一元线性回归

1.1.3　利用回归方程进行点预测和区间预测

若线性回归作显著性检验的结果是放弃H_0，也就是放弃回归系数β＝0的假设，便可以利用回归方程进行点预测和区间预测，这是人们关注线性回归的主要原因之一。

(1) 当$\boldsymbol{x}=\boldsymbol{x_0}$时，　用$\hat{y}_0 = a + bx_0$预测$Y_0$的观测值

y_0称为点预测。

由于$E(\hat{y}_0) = \alpha + \beta x_0 = E(Y_0)$,

$\boldsymbol{Y_0}$的观测值$\boldsymbol{y_0}$的点预测是无偏的。

1.1 一元线性回归

（2）当$x = x_0$时，用适合不等式$P\{Y_0 \in (G, H)\} \geq 1-\alpha$的统计量G和H所确定的随机区间(G, H)预测$Y_0$的取值范围称为区间预测，而(G, H)称为$Y_0$的$1-\alpha$预测区间。

若Y与样本中的各Y相互独立，则根据$Z = Y_0-(a+bx_0)$服从正态分布，$E(Z)=0$，

$$D(Z)=\sigma^2\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{l_{xx}}\right),$$

及$\dfrac{SSE}{\sigma^2} \sim \chi^2(n-2)$，Z与SSE相互独立，

1.1 一元线性回归

可以导出

$$t=\frac{Z}{\sqrt{\frac{SSE}{n-2}\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{l_{xx}}\right)}}\sim t(n-2).$$

因此，Y_0的1-α 预测区间为 $a+bx_0\pm\Delta(x_0)$，

$$\Delta(x_0)=t_{1-0.5\alpha}(n-2)\sqrt{\frac{SSE}{n-2}\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{l_{xx}}\right)}.$$

1.1 一元线性回归

例1.1 《吸附方程》某种物质在不同温度下可以吸附另一种物质，如果温度x(单位：℃)与吸附重量Y(单位：mg)的观测值如下表所示：

温度x	1.5	1.8	1.4	3.0	3.5	3.9	4.4	4.8	5.0
重量y	4.8	5.7	1.0	8.3	1.8.9	1.74	13.1	13.6	11.3

试求线性回归方程并用三种方法作显著性检验，若x_0＝2，求Y_0的0.95预测区间。

解：根据上述观测值得到n＝9，

1.1 一元线性回归

```
/*代码以及结果的解释见教材*/
data ex;
input x y@@;
cards;
1.5 4.8 1.8 5.7 1.4 7 3 8.3 3.5 1.8.9 3.9 1.74 4.4 13.1 4.8 13.6
5 11.3 2 .
;
proc gplot;plot y*x;symbol i=rl v=dot;proc reg;model
y=x/cli;
run;
```

1.1 一元线性回归

The SAS System　　　　23:52 Tuesday, March 13, 2007

The REG Procedure
Model: MODEL1
Dependent Variable: y

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	112.48368	112.48368	387.52	<.0001
Error	7	2.03188	0.29027		
Corrected Total	8	114.51556			

Root MSE	0.53877	R-Square	0.9823
Dependent Mean	10.12222	Adj R-Sq	0.9797
Coeff Var	5.32260		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.25695	0.53235	0.48	0.6441
x	1	2.93028	0.14886	19.69	<.0001

1.1 一元线性回归

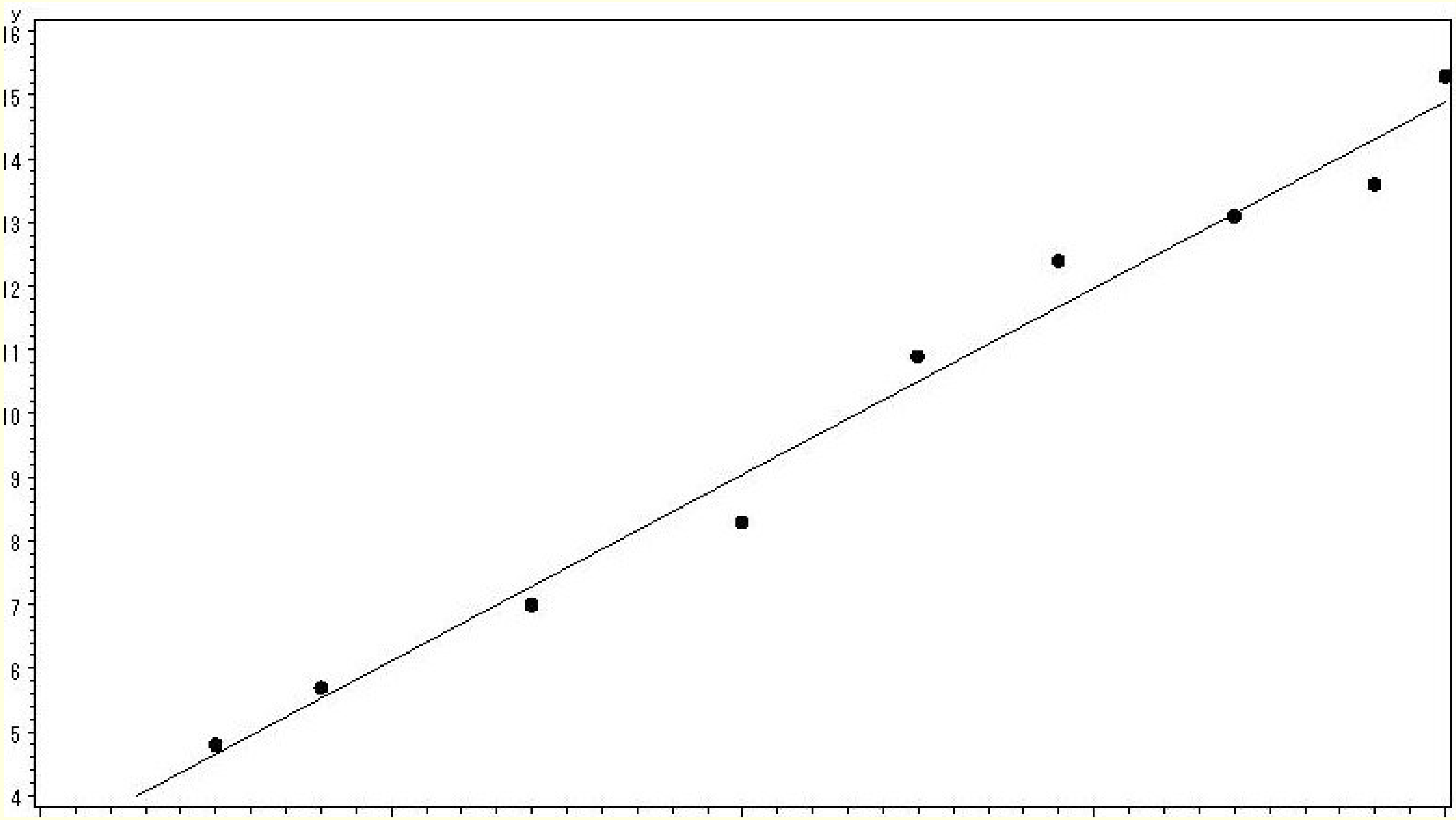
The REG Procedure
Model: MODEL1
Dependent Variable: y

Output Statistics

Obs	Dep Var y	Predicted Value	Std Error Mean Predict	95% CL Predict		Residual
1	4.8000	4.6524	0.3308	3.1574	6.1474	0.1476
2	5.7000	5.5315	0.2943	4.0797	6.9832	0.1685
3	7.0000	7.2896	0.2301	5.9043	8.6749	-0.2896
4	8.3000	9.0478	0.1877	7.6987	10.3969	-0.7478
5	10.9000	10.5129	0.1807	9.1692	11.8566	0.3871
6	12.4000	11.6850	0.1964	10.3291	13.0410	0.7150
7	13.1000	13.1502	0.2365	11.7589	14.5415	-0.0502
8	13.6000	14.3223	0.2789	12.8878	15.7568	-0.7223
9	15.3000	14.9083	0.3023	13.4476	16.3691	0.3917
10	.	6.1175	0.2714	4.6911	7.5440	.

Sum of Residuals	0
Sum of Squared Residuals	2.03188
Predicted Residual SS (PRESS)	3.13772

1.1 一元线性回归



1.2 一元非线性回归

例1.2 假设变量x与y的9组观测值如下

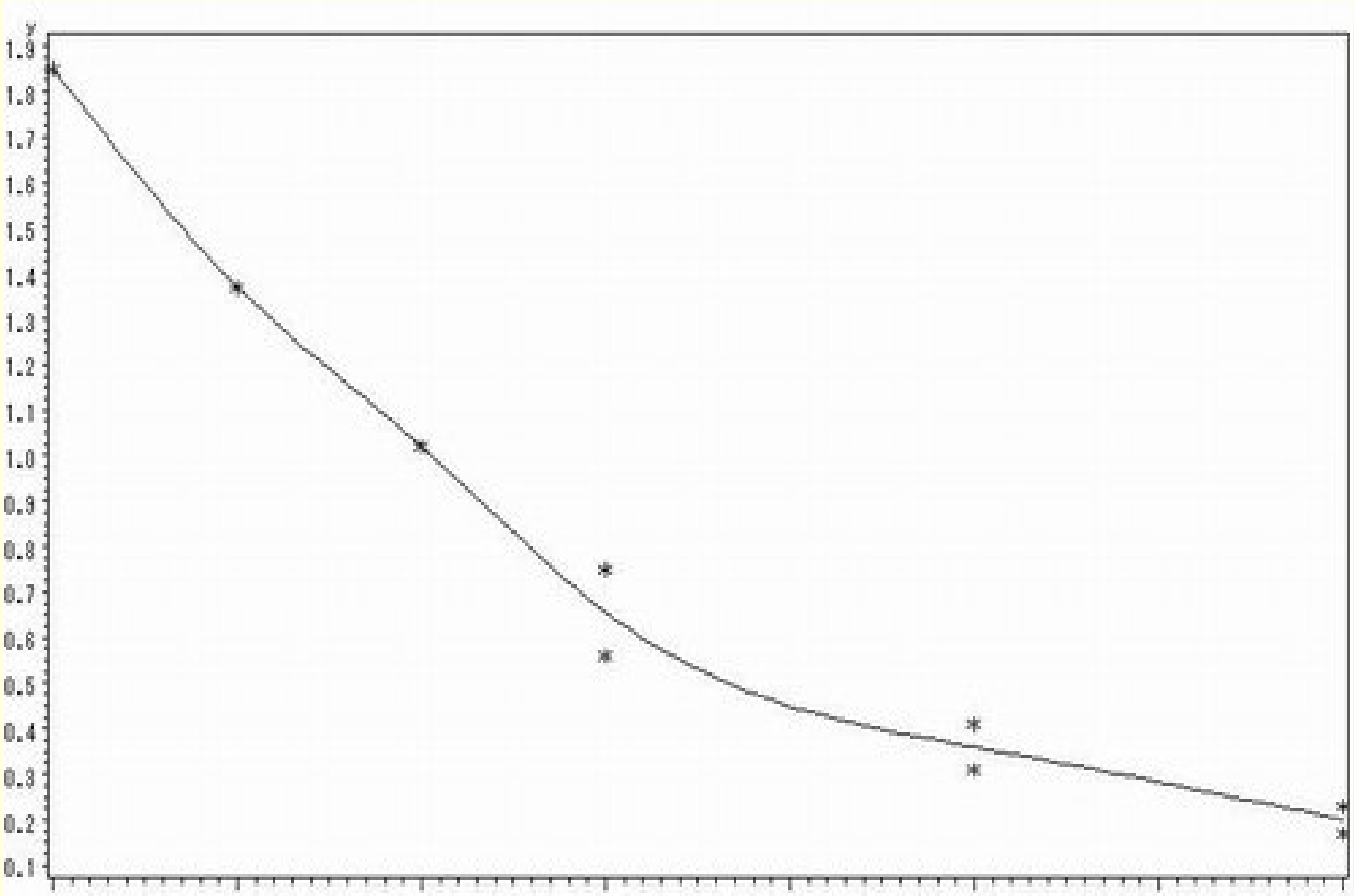
x	1	2	3	4	4	6	6	8	8
y	1.85	1.37	1.02	0.75	0.56	0.41	0.31	0.23	0.17

试选用多个线性方程进行拟合，并比较。

$$y = a + \frac{b}{x} \qquad y = ax^{b} \qquad y = ae^{bx}$$

方法主要是：将非线性化为线性

1.2 一元非线性回归



1.2 一元非线性回归

```
data ex;input x y @@;
x1=1/x;lx=log(x);ly=log(y);
cards;
1 1.85 2 1.37 3 1.02 4 0.75 4 0.56
6 0.41 6 0.31 8 0.23 8 0.17
;
proc gplot;plot y*x;symbol i=spline v=star;
proc reg;model y=x1;
proc reg;model ly=lx;
proc reg;model ly=x;
run;
```

1.2 一元非线性回归

The SAS System 10:45 Friday, March 16, 2007

The REG Procedure
Model: MODEL1
Dependent Variable: y

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2.33605	2.33605	57.86	0.0001
Error	7	0.28264	0.04038		
Corrected Total	8	2.61869			

Root MSE	0.20094	R-Square	0.8921
Dependent Mean	0.74111	Adj R-Sq	0.8767
Coeff Var	27.11326		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.11593	0.10603	1.09	0.3104
x1	1	1.92915	0.25362	7.61	0.0001

1.2 一元非线性回归

The REG Procedure
Model: MODEL1
Dependent Variable: ly

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	4.80864	4.80864	64.16	<.0001
Error	7	0.52460	0.07494		
Corrected Total	8	5.33324			

Root MSE	0.27376	R-Square	0.9016
Dependent Mean	-0.58024	Adj R-Sq	0.8876
Coeff Var	-47.18028		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.96379	0.21326	4.52	0.0027
lx	1	-1.12915	0.14096	-8.01	<.0001

1.2 一元非线性回归

The SAS System　　　10:45 Friday, March 16, 20

The REG Procedure
Model: MODEL1
Dependent Variable: ly

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	5.18785	5.18785	249.77	<.0001
Error	7	0.14539	0.02077		
Corrected Total	8	5.33324			

Root MSE	0.14412	R-Square	0.9727
Dependent Mean	-0.58024	Adj R-Sq	0.9688
Coeff Var	-24.83823		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.92296	0.10656	8.66	<.0001
x	1	-0.32211	0.02038	-15.80	<.0001

1.2 一元非线性回归

第一个方程 $\hat{y}=a+\frac{b}{x}$，设 $w=\frac{1}{x}$ 后化为 $\hat{y}=a+bw$

$$\hat{y}=0.1159+\frac{1.9291}{x}$$

第二个方程 $\hat{y}=ax^{b}$，变换形式为 $\ln\hat{y}=\ln a+b\ln x$

$$\hat{z}=0.9638-1.1292w \qquad \hat{y}=2.6216+x^{-1.1292}$$

第三个方程 $\hat{y}=ae^{bx}$，变换形式为 $\ln\hat{y}=\ln a+bx$

$$\hat{z}=0.9230-0.3221x \qquad \hat{y}=2.5168e^{-0.3221x}$$

1.2 一元非线性回归

```
data ex;input x y @@;
x1=1/x;lx=log(x);ly=log(y);
y1=0.1159+1.9291*x1;q1+(y-y1)**2;
y2=exp(0.9638-1.1292*lx);q2+(y-y2)**2;
y3=exp(0.9230-0.3221*x);q3+(y-y3)**2;
cards;
1 1.85 2 1.37 3 1.02 4 0.75 4 0.56
6 0.41 6 0.31 8 0.23 8 0.17
;
proc print;var q1-q3;run;
```

1.2 一元非线性回归

The SAS System   10:

Obs	q1	q2	q3
1	0.03802	0.59543	0.000689
2	0.12186	0.62483	0.003037
3	0.19002	0.69335	0.006927
4	0.21307	0.73418	0.010072
5	0.21453	0.73432	0.028007
6	0.21528	0.73834	0.030089
7	0.23151	0.73968	0.033045
8	0.24765	0.74010	0.034541
9	0.28264	0.74658	0.034996

1.3 多元线性回归

- 人的体重与身高、胸围
- 血压值与年龄、性别、劳动强度、饮食习惯、吸烟状况、家族史
- 糖尿病人的血糖与胰岛素、糖化血红蛋白、血清总胆固醇、甘油三脂
- 射频治疗仪定向治疗脑肿瘤过程中，脑皮质的毁损半径与辐射的温度、与照射的时间

1.3 多元线性回归

多元回归模型：含两个以上解释变量的回归模型

多元线性回归模型：一个应变量与多个解释变量

之间设定的是线性关系

多元线性回归模型一般形式为：

$$\underbrace{\boldsymbol{Y}=\boldsymbol{b}_0}_{\text{截距}}+\underbrace{\boldsymbol{b}_1\boldsymbol{X}_1+\boldsymbol{b}_2\boldsymbol{X}_2+\cdots+\boldsymbol{b}_k\boldsymbol{X}_k}_{\text{偏回归系数}}+\underbrace{\boldsymbol{\varepsilon}}_{\text{残差}}$$

1.3 多元线性回归

多元线性回归模型的假设:

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_k X_k + u$$

解释变量 $\boldsymbol{X_i}$ 是确定性变量，不是随机变量；
解释变量之间互不相关，即无多重共线性。
随机误差项不存在序列相关关系
随机误差项与解释变量之间不相关
随机误差项服从**0**均值、同方差的正态分布

1.3 多元线性回归

多元模型的解析表达式：

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_k X_k + \varepsilon$$

$$\begin{cases} Y_1 = b_0 + b_1 X_{11} + b_2 X_{21} + \cdots + b_k X_{k1} + u_1 \\ Y_2 = b_0 + b_1 X_{12} + b_2 X_{22} + \cdots + b_k X_{k2} + u_2 \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ Y_n = b_0 + b_1 X_{1n} + b_2 X_{2n} + \cdots + b_k X_{kn} + u_n \end{cases}$$

1.3 多元线性回归

多元模型的矩阵表达式:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & X_{11} & X_{21} & \cdots & X_{k1} \\ 1 & X_{12} & X_{22} & \cdots & X_{k2} \\ \ldots & \ldots & \ldots & & \ldots \\ 1 & X_{1n} & X_{2n} & \cdots & X_{kn} \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ \ldots \\ b_k \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$$\boldsymbol{Y} = \boldsymbol{XB} + \boldsymbol{\varepsilon}$$

1.3 多元线性回归

参数值估计：最小二乘估计

$$Q=\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2$$

$$=\sum_{i=1}^{n}\left(Y_i-\left(\hat{b}_0+\hat{b}_1X_{1i}+\cdots+\hat{b}_kX_{ki}\right)\right)^2 \longleftarrow \begin{cases}\dfrac{\partial Q}{\partial \hat{b}_0}=0\\ \dfrac{\partial Q}{\partial \hat{b}_1}=0\\ \dfrac{\partial Q}{\partial \hat{b}_2}=0\\ \cdots\cdots\cdots\cdots\\ \dfrac{\partial Q}{\partial \hat{b}_k}=0\end{cases}$$

参数估计公式：

$$\boxed{\hat{B}=(X^TX)^{-1}X^TY}$$

1.3 多元线性回归

多元线性回归模型的检验

主要介绍：

拟合优度检验（判定系数）

回归方程的显著性检验（**F**－检验）

回归参数的显著性检验（**t**－检验）

1.3 多元线性回归

拟合优度检验

目的：构造一个不含单位，可以相互比较，而且能直观判断拟合优劣的指标。

$$R^2 = \frac{SSR}{SST}$$

判定系数的定义：

判定系数的定义

意义：判定系数越大，自变量对因变量的解释程度越高，自变量引起的变动占总变动的百分比高。观察点在回归直线附近越密集。 取值范围：**0-1**

1.3 多元线性回归

回归方程的显著性检验

检验的目的

检验Y与解释变量$\mathbf{x_1}$，$\mathbf{x_2}$，……$\mathbf{x_k}$之间的线性关系是否显著。

检验的步骤

第一步，提出假设：

$$\begin{cases} \text{原假设：} H_0\text{：} b_1=b_2=\cdots\cdots b_k=0 \\ \text{备择假设：} H_1\text{：} b_i \text{不全为} 0 \quad (i=1,\ 2,\ \ldots,\ k) \end{cases}$$

1.3 多元线性回归

第二步，计算统计量：

$$F = \frac{SSR/k}{SSE/(n-k-1)} \sim F(k, n-k-1)$$

第三步，查表，得：

$$F_\alpha = F_\alpha(k, n-k-1)$$

第四步，做检验：

检验法则

$F > F_\alpha$ → 拒绝H_0，回归方程显著

$F < F_\alpha$ → 接受H_0，回归方程不显著

1.3 多元线性回归

回归系数的显著性检验

回归方程显著，并不意味着每个解释变量对因变量**Y**的影响都重要**,**因此需要进行检验。

1.3 多元线性回归

回归系数显著性的检验的步骤

第一步，提出假设：

原假设：H_0：　$b_i=0$　（i=1，2，……k）

备择假设：H_1：$b_i \neq 0$　（i=1，2，……k）

第二步，构造并计算统计量 ：

$$T_i = \frac{\hat{b}_i}{s(\hat{b}_i)} \quad (i = 1, 2, \cdots, k)$$

1.3 多元线性回归

第三步，查表得 ：$$t_{\alpha/2} = t_{\alpha/2}(n-k-1)$$

第四步，做检验：

检验法则

$|T_i| < t_{\frac{\alpha}{2}}$ → 接受H_0

$|T_i| > t_{\frac{\alpha}{2}}$ → 拒绝H_0

1.3 多元线性回归

例 某品种水稻糙米含镉量**y(mg/kg)**与地上部生物量 **x1(10g/**盆**)**及土壤含镉量**x2(100mg/kg)**的**8**组观测值如表**1.1**。试建立多元线性回归模型。

x1	1.37	11.34	9.67	0.76	11.67	15.91	15.74	1.41
x2	9.08	1.89	3.06	1.8.2	0.05	0.73	1.03	6.25
y	4.93	1.86	1.33	5.78	0.06	0.43	0.87	3.86

1.3 多元线性回归

```
/*代码以及结果的解释见教材*/
data ex;
input  x1-x2 y@@;
cards;
…
;
proc reg;
model y=x1 x2;
run;
```

1.3 多元线性回归

回归方程显著性检验：

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	31.46291	31.46291	494.06	<.0001
Error	6	0.38209	0.06368		
Corrected Total	7	31.84500			

Root MSE	0.25235	R-Square	$\boxed{0.9880}$
Dependent Mean	2.51500	Adj R-Sq	0.9860
Coeff Var	10.03390		

拟合度很高

由方差分析表可知，其**F value=494.06,pr>F**的值**<0.0001**，远小于**0.05**，故拒绝原假设，接受备择假设，认为**y1**与**x1,x2**之间具有显著性的线性关系；

1.3 多元线性回归

参数显著性检验:

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	3.61051	0.95915	3.76	0.0131
x1	1	-0.19828	0.05822	-3.41	0.0191
x2	1	0.20675	0.09769	2.12	0.0879

由参数估计 表可知，对自变量x_2检验t值分别为t=1.12、,Pr>|t|的值=0.0879,大于0.05，因此，拒绝原假设认为x_2的系数应为0，说明x_2的系数没有通过检验。为此，需要在程序中**model y1=x1 x2**中去掉x_2

1.3 多元线性回归

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	5.62117	0.16580	33.90	<.0001
x1	1	-0.31911	0.01436	-22.23	<.0001

对常数检验**t**值分别为**t=33.9**、,**Pr>|t|**的值**<0.0001,**远小于**0.05**，说明截距项通过检验，估计值为**5.62117**，同理可知x_1的系数通过检验，估计值为**-0.31911**

回归方程： $y = -0.31911x_1 + 5.62117$

1.3 多元线性回归

许多实际问题中可能还会出现某几个变量的系数并没有通过检验，此时，可以在原程序中的**model y1=x1-x2**中去掉没用通过的变量,直到所有的系数均通过检验。或者使用逐步回归方法，让软件自动保留通过检验的变量。

1.4 多元非线性回归

建立多元非线性回归方程在科学研究中应用广泛，其重要方法是将非线性回归方程转化为线性回归方程。转化时应首先选择适合的非线性回归形式，并将其线性化。再确定线性化回归方程的系数，最后确定非线性回归方程中未知的系数或参数。

1.4 多元非线性回归

实例：湖北省油菜投入与产出的统计分析

1.投入指标

（**1**）土地（**S**）。土地用播种面积来表示。农作物播种面积是指当年从事农业

（**2**）劳动（**L**）。劳动用劳动用工数（成年劳动力一人劳动一天为一个工）来表示。劳动用工中包含着直接和间接生产用工。

（**3**）资本（**K**）。资本用物质费用来表示。物质费用包含直接费用和间接费用。主要有种子秧苗费、农家肥费、化肥费、农药费、畜力、固定资产折旧费和管理及其他费用等。

1.产出指标

产出指标用湖北省历年油菜生产的总产量**(Y)**来表示。

1.4 多元非线性回归

年份	产量（万吨）Y	物质费用（万元）K	播种面积（万亩）S	劳动用工（万个）L	年份序号 t
1990	70.8972	40076.5884	821.1305	15341.4273	1
1991	83.7506	48008.7690	911.1500	15831.0950	2

$$Y = A_0 e^{\lambda t} K^{\alpha} L^{\beta} S^{\gamma}$$

$$\ln Y = \ln A_0 + \lambda t + \alpha \ln K + \beta \ln L + \gamma \ln S + \mu$$

1.4 多元非线性回归

```
data ex;input y k s l t  @@;
x1=log(k);x2=log(s);x3=log(l);y1=(y);
cards;
70.8972      40076.5884    821.1305     15341.4273    1
83.7506      48008.7690    911.1500     15831.0950    2
70.8627      44593.8425    805.6150     13306.8090    3
78.3451      43460.3229    783.2100     13314.5700    4
98.0749      72651.2633    923.8050     14596.1190    5
134.8767     146108.3421   1281.8900    20911.1070    7
141.5315     162433.3500   1244.7000    18670.5000    8
154.7607     166979.6325   1330.5150    18621.2100    9
159.9743     190391.5262   1501.4600    20771.3480    10
198.4942     205914.6645   1738.4100    22599.3300    11
194.7943     189761.7335   1671.0900    20963.6250    12
181.1013     193461.5610   1761.9450    21936.2153    14
231.1184     183768.4035   1779.1500    19606.2330    15
;
proc reg;model y1=x1 x2 x3 t ; /*/selection=stepwise*/
run;
```

1.4 多元非线性回归

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	2.15231	0.53808	148.95	<.0001
Error	8	0.02890	0.00361		
Corrected Total	12	2.18121			

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.93016	1.91920	0.48	0.6409
x1	1	0.24781	0.09610	2.58	0.0327
x2	1	1.28223	0.57122	2.24	0.0550
x3	1	-0.82102	0.55591	-1.48	0.1780
t	1	-0.00168	0.02437	-0.07	0.9466

变量t 的显著性概率为**0.9466**，远大于**0.05**，因此将**model y1=x1 x2 x3 t;**去掉他，即改为**model y1=x1 x2 x3 ;**

1.4 多元非线性回归

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	2.15229	0.71743	223.29	<.0001
Error	9	0.02892	0.00321		
Corrected Total	12	2.18121			

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.87950	1.67253	0.53	0.6117
x1	1	0.24554	0.08518	2.88	0.0181
x2	1	1.24568	0.20239	6.15	0.0002
x3	1	-0.78798	0.26689	-2.95	0.0162

截距项**Intercept** 的显著性概率为**0.6117**，大于**0.05**，因此将**model y1=x1 x2 x3 ;** 改为**model y1=x1 x2 x3 /noint;**

1.4 多元非线性回归

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	309.07415	103.02472	34565.8	<.0001
Error	10	0.02981	0.00298		
Uncorrected Total	13	309.10395			

Root MSE	0.05459	R-Square	0.9999
Dependent Mean	4.85895	Adj R-Sq	0.9999
Coeff Var	1.12358		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
x1	1	0.22851	0.07588	3.01	0.0131
x2	1	1.21016	0.18376	6.59	<.0001
x3	1	-0.65225	0.06539	-9.98	<.0001

1.4 多元非线性回归

$$\ln \hat{Y} = 0.244189 \ln K + 1.172185 \ln S - 0.643284 \ln L$$

F=34565.8 R^2=0.9999

K，S，L的t值分别为（3.01） (6.59) (-9.98)

$$\hat{Y} = K^{0.22851} S^{1.21016} L^{-0.65225}$$

要善于解释经济含义、本模型虽然满足数学规则，但不能通过经济检验。助于如何继续修正模型，需要学习数学与经济的交叉学科《计量经济学》。

1.5 逐步回归

逐步回归的基本思想是，从当前在圈外的全部变量中，挑选其偏回归平方和贡献最大的变量，用方差比进行显著性检验的办法，判别是否选入；而当前在圈内的全部变量中，寻找偏回归平方和贡献最小的变量，用方差比进行显著性检验的办法，判别是否从回归方程中剔除。选入和剔除循环反复进行，直至圈外无符合条件的选入项，圈内无符合条件的剔除项为止。

逐步回归选择变量快捷，但对于存在多重共线的自变量选择，有时并不准确，使用时注意分辨。

还是用上面的例子，将**model y1=x1 x2 x3 t ;**
改为**model y1=x1 x2 x3 t /selection=stepwise;**

1.5 逐步回归

注意，为了筛选变量宽容，程序中默认显著度为**0.15**，而不是**0.05**，以避免条件过于严格只用筛选无法进行。

All variables left in the model are significant at the 0.1500 level.

No other variable met the 0.1500 significance level for entry into the model.

Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	x2		1	0.9668	0.9668	11.0349	320.51	<.0001
2	t		2	0.0086	0.9754	7.8417	3.50	0.0909
3	x1		3	0.0077	0.9831	5.1812	4.12	0.0730

从程序结果中不难看出，**x2**、**x1**、**t**进入模型。因此**model y1=x1 x2 x3 t /selection=stepwise;**改为**model y1=x1 x2 t /noint;**再运行一遍即可。

1.5 逐步回归

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	309.06293	103.02098	25111.1	<.0001
Error	10	0.04103	0.00410		
Uncorrected Total	13	309.10395			

Root MSE	0.06405	R-Square	0.9999
Dependent Mean	4.85895	Adj R-Sq	0.9998
Coeff Var	1.31822		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
x1	1	0.21081	0.09041	2.33	0.0419
x2	1	0.29433	0.14500	2.03	0.0698
t	1	0.04175	0.00501	8.34	<.0001

思考：为什么这个结果与前面计算的结果不一样？

1. 6 灰色预测:概述

1.6.1 概述

灰色系统是指“部分信息已知, 部分信息未知”的“小样本”,“贫信息”的不确定性系统, 它通过对“部分”已知信息的生成、开发去了解、认识现实世界，实现对系统运行行为和演化规律的正确把握和描述.

灰色系统模型的特点：对试验观测数据及其分布没有特殊的要求和限制，是一种十分简便的新理论，具有十分宽广的应用领域。

1.6 灰色预测:概述

灰色系统理论经过20年的发展，已基本建立起一门新兴的结构体系，其研究内容主要包括：灰色系统建模理论、灰色系统控制理论、灰色关联分析方法、灰色预测方法、灰色规划方法、灰色决策方法等。

我们主要介绍灰色GM(1.1)模型预测。即灰色生成、GM（1.1）模型建模机理、GM(1.1)模型的精度检验

1.6 灰色预测:概述

1.62 GM(1,1)模型

1.令 $X^{(0)}$为**GM(1,1)**建模序列,

$$X^{(0)}=(x^{(0)}(1),x^{(0)}(2),\ldots,x^{(0)}(n))$$

$X^{(1)}$为 $X^{(0)}$的**1-AGO**序列,

$$X^{(1)}=(x^{(1)}(1),x^{(1)}(2),\ldots,x^{(1)}(n))$$

$$x^{(1)}(k)=\sum_{i=1}^{k}x^{(0)}(i)\qquad k=1,2,\ldots,n$$

1.6 灰色预测:概述

令$Z^{(1)}$为 $X^{(1)}$的紧邻均值（**MEAN**）生成序列

$$Z^{(1)}=(z^{(1)}(2),z^{(1)}(3),\ldots,z^{(1)}(n))$$

$$z^{(1)}(k)=0.5\,x^{(1)}(k)+0.5\,x^{(1)}(k-1)$$

则**GM(1,1)**的灰微分方程模型为

$$x^{(0)}(k)+az^{(1)}(k)=b$$

1.6 灰色预测:概述

记 $\hat{\alpha}=(a,b)^T$

则灰微分方程的最小二乘估计参数列满足

$$\hat{\alpha}=(B^TB)^{-1}B^TY_n$$

其中

$$\boldsymbol{B}=\begin{bmatrix} -z^{(1)}(2) & 1 \\ -z^{(1)}(3) & 1 \\ \cdots & \cdots \\ -z^{(1)}(n) & 1 \end{bmatrix} \qquad \boldsymbol{Y_n}=\begin{bmatrix} x^{(0)}(2) \\ x^{(0)}(3) \\ \cdots \\ x^{(0)}(n) \end{bmatrix}$$

1.6 灰色预测:概述

称 $\frac{dx^{(1)}}{dt}+ax^{(1)}=b$ 为灰色微分方程

$$x^{(0)}(k)+az^{(1)}(k)=b$$

的白化方程，也叫影子方程。

综上所述，则有

1. 白化方程 $\frac{dx^{(1)}}{dt}+ax^{(1)}=b$ 的解也称

时间响应函数为

$$\hat{x}^{(1)}(t)=(x^{(1)}(0)-\frac{b}{a})e^{-at}+\frac{b}{a}$$

1.6 灰色预测:概述

1.GM(1,1)灰色微分方程$x^{(0)}(k)+az^{(1)}(k)=b$

的时间响应序列为

$$\hat{x}^{(1)}(k+1)=[x^{(1)}(0)-\frac{b}{a}]e^{-ak}+\frac{b}{a}$$

3.取$x^{(1)}(0)=x^{(0)}(1)$ ，则 $k=1,2,\ldots,n$

$$\hat{x}^{(1)}(k+1)=[\,x^{(0)}(1)-\frac{b}{a}\,]e^{-ak}+\frac{b}{a}$$

$$k=1,2,\ldots,n$$

1.6 灰色预测:概述

4.还原值

$$\hat{x}^{(0)}(k+1) = \hat{x}^{(1)}(k+1) - \hat{x}^{(1)}(k)$$

上式即为预测方程。

GM(1,1)模型的检验分为三个方面:

残差检验;

关联度检验;

后验差检验。

1.6 灰色预测:概述

后验差检验判别参照表

C	模型精度
<0.35	优
<0.5	合格
<0.65	勉强合格
>0.65	不合格

其中 $C = \dfrac{S_1}{S_2}$

S_1 ⟶ 残差序列均方差

S_2 ⟶ 原序列均方差

1.6 灰色预测:概述

给定原始时间 1990-2001 年资料列：

$x^{(0)}=(x^{(0)}(1), x^{(0)}(2), x^{(0)}(3), x^{(0)}(4), x^{(0)}(5), x^{(0)}(6), x^{(0)}(7), x^{(0)}(8), x^{(0)}(9), x^{(0)}(10),$
$x^{(0)}(11), x^{(0)}(12))$

$=(19519, 19578, 19637, 19695, 16602, 25723, 30379, 34473, 38485, 40514,$
$42400, 48337),$

对 $x^{(0)}$ 做 AGO 生成，有 $x^{(1)}=\mathrm{AGO}x^{(0)}$， $x^{(1)}(k)=\sum^{k} x^{(0)}(m)$，则

$x^{(1)}=(x^{(1)}(1), x^{(1)}(2), x^{(1)}(3), x^{(1)}(4), x^{(1)}(5), x^{(1)}(6), x^{(1)}(7), x^{(1)}(8), x^{(1)}(9), x^{(1)}(10),$
$x^{(1)}(11), x^{(1)}(12))$

$=(19519, 39097, 58734, 78429, 95031, 120754, 151133, 185606, 224091,$
$264605, 307005, 355342),$

1.6 灰色预测:概述

对上述 $x^{(0)}$的 GM（1，1）参数 a,b，按下述算式辨识：

$=(B^TB)^{-1}B^Ty_N$

基于 $x^{(0)}$ $x^{(0)}$ 与 $x^{(1)}$，有

$$B=\begin{bmatrix} -z^{(1)}(2) & 1 \\ -z^{(1)}(3) & 1 \\ -z^{(1)}(4) & 1 \\ -z^{(1)}(5) & 1 \\ -z^{(1)}(6) & 1 \\ -z^{(1)}(7) & 1 \\ -z^{(1)}(8) & 1 \\ -z^{(1)}(9) & 1 \\ -z^{(1)}(10) & 1 \\ -z^{(1)}(11) & 1 \\ -z^{(1)}(12) & 1 \end{bmatrix}=\begin{bmatrix} -0.5(x^{(1)}(1)+x^{(1)}(2)) & 1 \\ -0.5(x^{(1)}(2)+x^{(1)}(3)) & 1 \\ -0.5(x^{(1)}(3)+x^{(1)}(4)) & 1 \\ -0.5(x^{(1)}(4)+x^{(1)}(5)) & 1 \\ -0.5(x^{(1)}(5)+x^{(1)}(6)) & 1 \\ -0.5(x^{(1)}(6)+x^{(1)}(7)) & 1 \\ -0.5(x^{(1)}(7)+x^{(1)}(8)) & 1 \\ -0.5(x^{(1)}(8)+x^{(1)}(9)) & 1 \\ -0.5(x^{(1)}(9)+x^{(1)}(10)) & 1 \\ -0.5(x^{(1)}(10)+x^{(1)}(11)) & 1 \\ -0.5(x^{(1)}(11)+x^{(1)}(12)) & 1 \end{bmatrix}=\begin{bmatrix} -29308 & 1 \\ -48915.5 & 1 \\ -68581.5 & 1 \\ -86730 & 1 \\ -107892.5 & 1 \\ -135943.5 & 1 \\ -168369.5 & 1 \\ -204848.5 & 1 \\ -244348 & 1 \\ -331173.5 & 1 \\ -5236.2 & 1 \end{bmatrix}$$

1.6 灰色预测:概述

y_N=[$x^{(0)}(2)$, $x^{(0)}(3)$, $x^{(0)}(4)$, $x^{(0)}(5)$, $x^{(0)}(6)$, $x^{(0)}(7)$, $x^{(0)}(8)$, $x^{(0)}(9)$, $x^{(0)}(10)$, $x^{(0)}(11)$ $x^{(0)}(12)]^T$=(39097, 58734, 78429, 95031, 120754, 151133, 185606, 224091, 264605 307005, 355342)^T。

将 B,y_N 代入辨识算式，有

$$\hat{a}=\begin{bmatrix} a \\ b \end{bmatrix}=(B^TB)^{-1}B^Ty_N=\begin{bmatrix} -0.1062105 \\ 13999.9 \end{bmatrix},$$

$a=-0.1062105$,

$b=13999.9$

得 GM(1，1)模型为

1）灰微分方程

$$x^{(0)}(k)-0.1062105z^{(1)}(k)=13999.9;$$

2）白化方程

$$\frac{dx^{(1)}}{dt}-0.1062105z^{(1)}(k)=13999.9;$$

3）白化方程的时间响应式

$$\hat{x}^{(1)}(t+1)=(x^{(0)}(1)-\frac{b}{a})e^{-at}+\frac{b}{a}$$

$$=151332.5e^{0.1062105t}-131813.5,$$

$$\hat{x}^{(0)}(t+1)=\hat{x}^{(1)}(t+1)-\hat{x}^{(1)}(t),\ \hat{x}^{(0)}(1)=\hat{x}^{(1)}(1)=19519。①$$

得还原方程 $\hat{x}^{(0)}(t+1)=15248.968e^{0.1062105t}$，②

1.6 灰色预测:概述

年份	预测值（10^4t）	实际值（10^4t）	残差 q(10^4t)	相对误差 ξ_1(%)
1991	16957.69	19578	2620.3	13.38
1992	18857.91	19637	779.09	3.967
1993	20971.05	19695	-1276	-6.48
1994	23320.96	16602	-6719	-40.5
1995	25934.29	25723	-211.3	-0.82
1996	28840.3	30379	1538.7	5.065
1997	32072.06	34473	2400.9	6.965
1998	35665.88	38485	2819.1	7.325
1999	39662.47	40514	851.53	2.102
2000	44106.89	42400	-1707	-4.03
2001	49049.32	48337	-712.3	-1.47

注： $q = x^{(0)}(k) - \hat{x}^{(0)}(k)$，$\xi = \dfrac{x^{(0)}(k) - \hat{x}^{(0)}(k)}{x^{(0)}(k)}\%$

1.6 灰色预测:概述

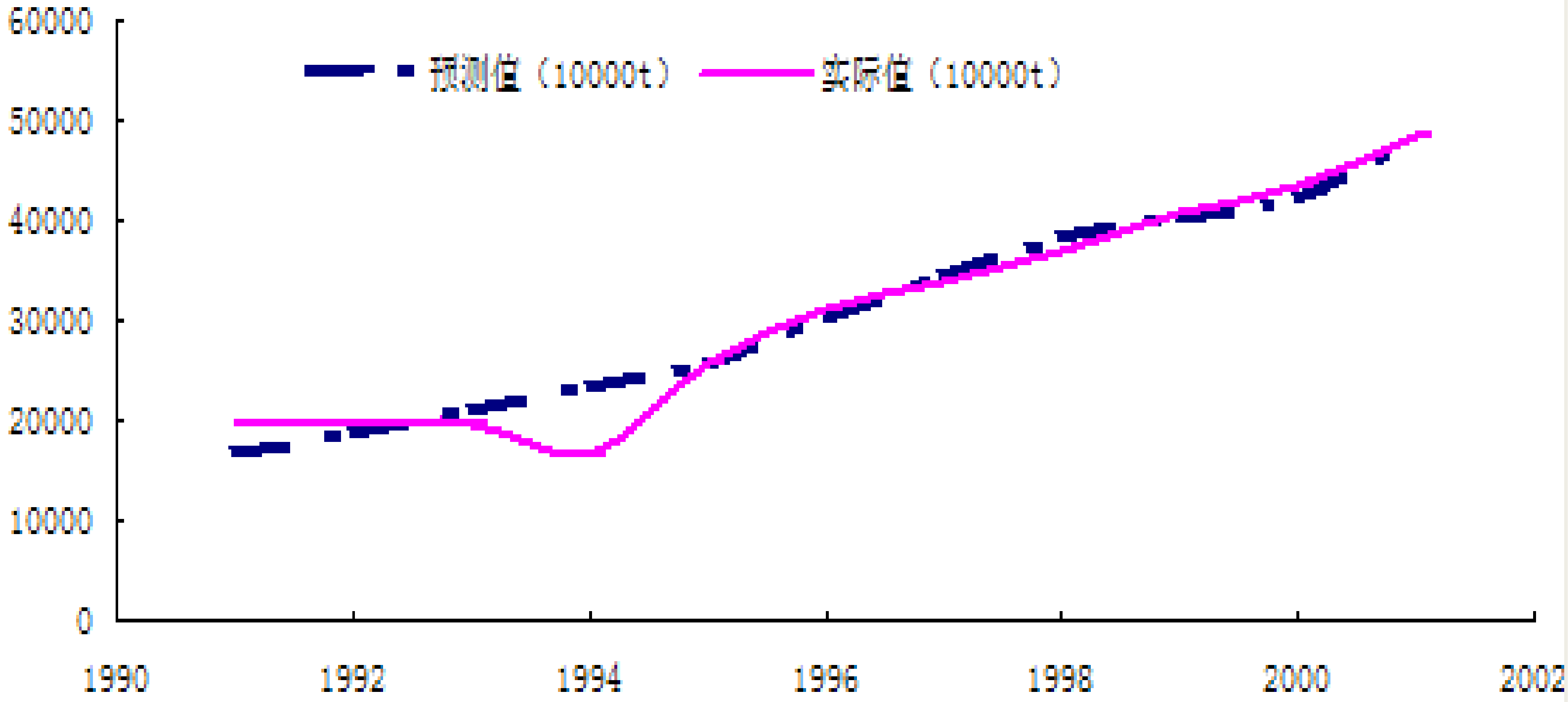


图 2-10 1990-2002 年蔬菜产出水平的灰色预测值与实际值比较

1.6 灰色预测:概述

敲击“Enter”键就可以得出结果

379,34473,38485,40514,42400,48337];gm(x)

1.7 马尔科夫链预测：理论

理论背景：

- 马尔柯夫链是一种特殊的随机时间序列。它的特点是：序列将来的状态只与现在的状态有关而与过去的状态无关。这种特性称为无后效性或称马氏性。

- 马尔柯夫链的研究对象是某一系统的状态与状态转移。设该系统有**n**个状态，以

$$P\{X_{t+1}=j \mid X_t=i\}=p_{ij}, \quad i,j=1,2,\ldots,n$$

1.7 马尔科夫链预测：理论

- 表示一个系统在时刻**t**处于状态**i**，于下一时刻**t+1**转变为状态**j**的概率，并称之为一步转移概率。
- 同时p_{ij}满足条件（1）$p_{ij} \geq 0$，（2）$\sum_{j=1}^{n} p_{ij}$ =1（i,j=1,2,…,n）。
- 由一步转移概率**pij**构成的矩阵成为一步转移概率矩阵，记为**P(1)**，即

1.7 马尔科夫链预测：理论

$$P(1)=\begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ p_{n1} & p_{n2} & \cdots & p_{nn} \end{bmatrix}$$

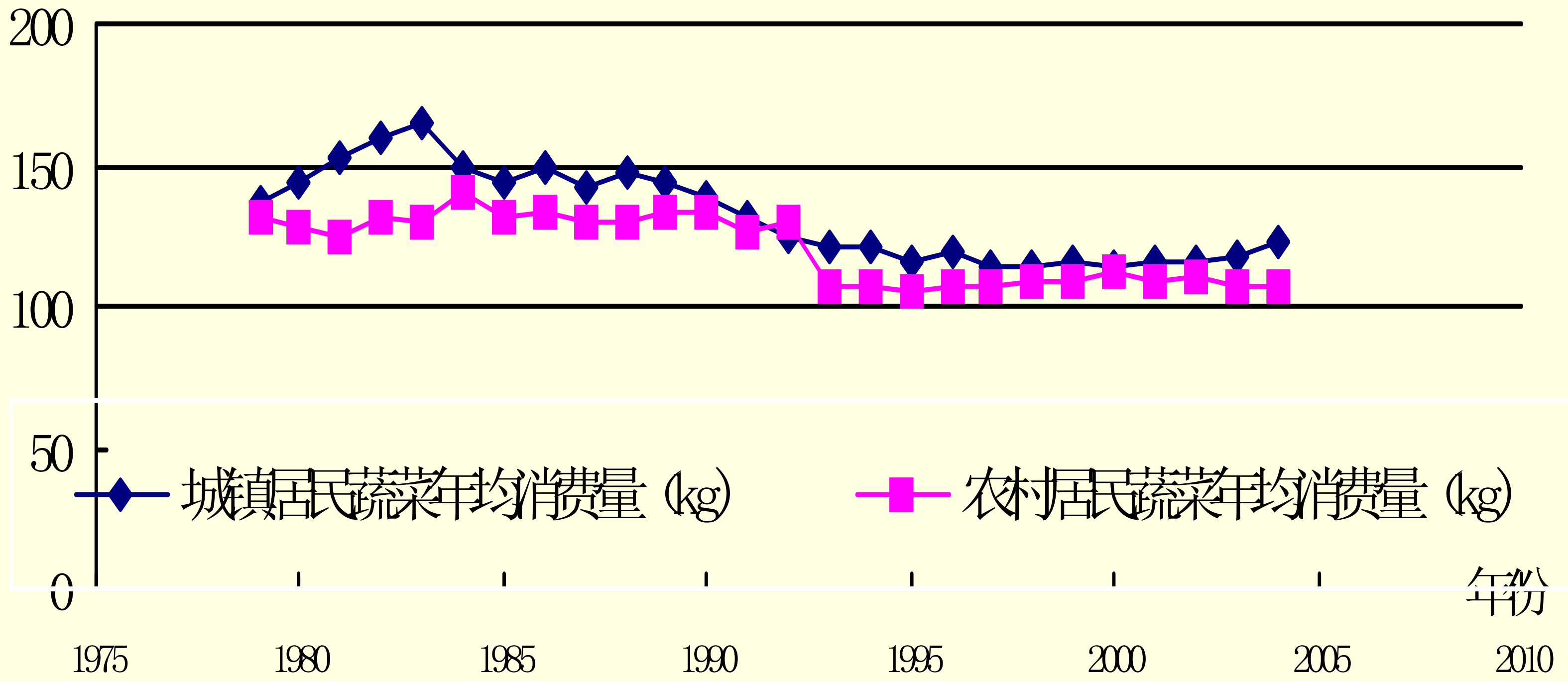
若以$P\{X_{t+k}=j \mid X_t=i\}=p_{ij}(k)$，$i,j=1,2,\cdots,n$表示一个系统从状态之径$k(k>1)$步到达状态j的转移概率，则称$p_{ij}(k)$为k步转移概率。

1.7 马尔科夫链预测：理论

➢ 由**k**步转移概率构成的矩阵成为**k**步转移概率矩阵，记为**P(k)**，即

$$P(k)=\begin{bmatrix} p_{11}(k) & p_{12}(k) & \cdots & p_{1n}(k) \\ p_{21}(k) & p_{22}(k) & \cdots & p_{2n}(k) \\ \cdots & \cdots & \cdots & \cdots \\ p_{n1}(k) & p_{n2}(k) & \cdots & p_{nn}(k) \end{bmatrix}$$

1.7 马尔科夫链预测：案例



1978-2004年城乡居民蔬菜年人均消费量趋势图

1.7 马尔科夫链预测：案例

➤长期趋势预测

➤第一步：状态划分及构造

表 蔬菜人均消费量状态分类及各状态下的平均增减率

等级	大减	平稳减	平稳增	大增
环比增减率（%）	(-∞，－3)	(－3，0)	(0，3)	(3，+∞)
状态	1	2	3	4
城镇各状态下的平均增减率 CI	-4.72	-1.41	1.14	4.35
农村各状态下的平均增减率 NI	-6.99	-1.41	1.63	6.89

1.7 马尔科夫链预测：案例

➤第二步：计算状态转移次数及转移概率矩阵

表 蔬菜年人均消费量的状态时间序列表

年份	城镇 环比增减率（%）	城镇 状态	农村 环比增减率（%）	农村 状态	年份	城镇 环比增减率（%）	城镇 状态	农村 环比增减率（%）	农村 状态
1975					1990	-4.10	1	0.46	3
1976	3.85	4			1991	-4.70	1	-5.20	1
1977	-2.22	2			1992	-5.50	1	1.69	3
1978	-1.52	4			1993	-3.40	1	-17.00	1
1979	5.38	4	-7.30	1	1994	0.08	3	0.40	3
1980	4.37	4	-3.02	1	1995	-3.50	1	-3.00	1
1981	6.29	4	-2.53	2	1996	1.75	3	1.57	3
1982	4.42	4	6.54	4	1997	-4.40	1	0.89	3
1983	4.02	4	-1.10	2	1998	0.37	3	1.63	3
1984	-9.90	1	7.23	4	1999	1.04	3	-0.10	2
1985	-3.10	1	-6.40	1	2000	-0.20	2	2.84	3
1986	2.74	3	1.92	3	2001	0.98	3	-2.40	2
1987	-3.90	1	-2.40	2	2002	0.57	3	1.14	3
1988	3.11	4	-0.30	2	2003	1.56	3	2.85	3
1989	-1.70	2	2.54	3	2004	3.35	4	-0.74	2

1.7 马尔科夫链预测：案例

➢ 可得城乡居民年人均蔬菜消费量的各状态的转移次数

表　中国城乡居民年人均蔬菜消费量的状态转移次数

城镇状态转移次数	1	2	3	4	Σ		农村状态转移次数	1	2	3	4	Σ
1	4	0	4	1	9		1	1	1	4	0	6
2	1	1	1	1	4		2	0	1	3	2	6
3	3	1	3	1	8		3	3	4	4	0	11
4	1	2	0	4	7		4	1	1	0	0	2

1.7 马尔科夫链预测：案例

- 可得城乡居民年人均蔬菜消费量的一步转移概率矩阵

表　中国城乡居民年人均蔬菜消费量的状态转移概率矩阵

城镇状态转移概率	1	2	3	4		农村状态转移概率	1	2	3	4
1	0.444	0	0.444	0.111		1	0.167	0.167	0.667	0
2	0.25	0.25	0.25	0.25		2	0	0.167	0.5	0.333
3	0.375	0.125	0.375	0.125		3	0.273	0.364	0.364	0
4	0.143	0.286	0	0.571		4	0.5	0.5	0	0

1.7 马尔科夫链预测：案例

由表**5**可知，城镇状态一步转移概率矩阵中，***p*31=0.375**， ***p*13=0.444**，***p*23= 0.25**，***p*32=0.125**，***p*14= 0.111**，***p*41=0.143**都不为**0**，说明城镇**Markov**链的状态**1**与**3**、**2**与**3**、**1**与**4**是互通的，从而状态**1**，**2**，**3**，**4**全部互通，因此城镇的**Markov**链是不可约的。另外，由于城镇状态转移概率矩阵主对角线上的概率全部大于**0**，显然四个状态是非周期的。

1.7 马尔科夫链预测：案例

- 第三步：中国城乡居民年人均蔬菜消费量的长期趋势

$$
\begin{cases}
p_1 = 0.44\,p_1 + 0.25\,p_2 + 0.375\,p_3 + 0.143\,p_4 \\
p_2 = 0.25\,p_2 + 0.125\,p_3 + 0.286\,p_4 \\
p_3 = 0.44\,p_1 + 0.25\,p_2 + 0.375\,p_3 \\
p_4 = 0.11\,p_1 + 0.25\,p_2 + 0.125\,p_3 + 0.571\,p_4 \\
p_1 + p_2 + p_3 + p_4 = 1
\end{cases}
$$

1.7 马尔科夫链预测：案例

- 得出解

$$\begin{bmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \end{bmatrix} = \begin{bmatrix} 0.318 \\ 0.144 \\ 0.282 \\ 0.255 \end{bmatrix}$$