第二讲 回归分析

- 一元线性回归模型
- 多元线性回归模型
- 非线性回归模型
- 回归变量的选择方法
- 逐步回归分析

- **回归分析**（regression analysis)是确定两种或两种以上变量间相互依赖的定量关系的一种统计分析方法，应用广泛。
- 回归分析按照涉及的变量的多少，分为**一元回归**和**多元回归分析**；按照因变量的多少，可分为**简单回归分析**和**多重回归分析**；按照自变量和因变量之间的关系类型，可分为**线性回归分析**和**非线性回归分析**。
- 如果在回归分析中，只包括一个自变量和一个因变量，且二者的关系可用一条直线近似表示，这种回归分析称为**一元线性回归分析**。
- 如果回归分析中包括两个或两个以上的自变量，且自变量之间存在线性相关，则称为**多重线性回归分析**。

一、一元线性回归模型

引例：探究财政收入与税收依存关系。

表 4－1　1978—2008 年税收与财政收入数据（数据见 RstatM. xls：d4. 3）

年份	y	x	年份	y	x
1978	11. 326 2	5. 192 8	1994	52. 181 0	51. 268 8
1979	11. 463 8	5. 378 2	1995	62. 422 0	60. 380 4
1980	11. 599 3	5. 717 0	1996	74. 079 9	69. 098 2
1981	11. 757 9	6. 298 9	1997	86. 511 4	82. 340 4
1982	12. 123 3	7. 000 2	1998	98. 759 5	92. 628 0
1983	18. 669 5	7. 555 9	1999	114. 440 8	106. 825 8
1984	16. 428 6	9. 473 5	2000	133. 952 3	125. 815 1
1985	20. 048 2	20. 407 9	2001	163. 860 4	153. 013 8
1986	21. 220 1	20. 907 3	2002	189. 036 4	176. 364 5
1987	21. 993 5	21. 403 6	2003	217. 152 5	200. 173 1
1988	23. 572 4	23. 904 7	2004	263. 964 7	241. 656 8
1989	26. 649 0	27. 274 0	2005	316. 492 9	287. 785 4
1990	29. 371 0	28. 218 7	2006	387. 602 0	348. 043 5
1991	31. 494 8	29. 901 7	2007	513. 217 8	456. 219 7
1992	34. 833 7	32. 969 1	2008	613. 303 5	542. 196 2
1993	43. 489 5	42. 553 0			

一、一元线性回归模型

定义：假设有两个变量x和y，x为自变量，y为因变量。则一元线性回归模型的基本结构形式为

$$y_a = a + bx_a + \varepsilon_a \qquad (2.1)$$

式中：a和b为待定参数；

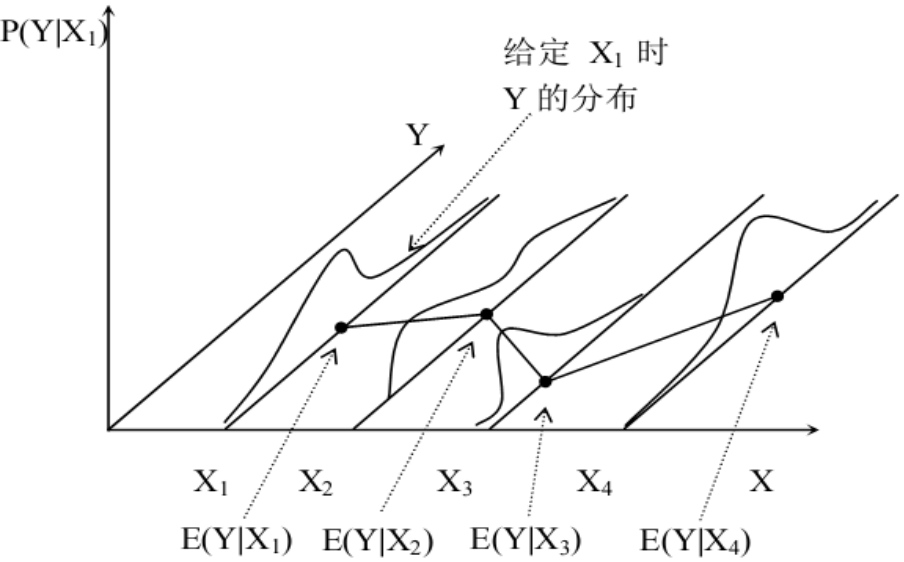
$\alpha = 1,2,\ldots,n$ 为各组观测数据的下标；

ε_a为随机变量。

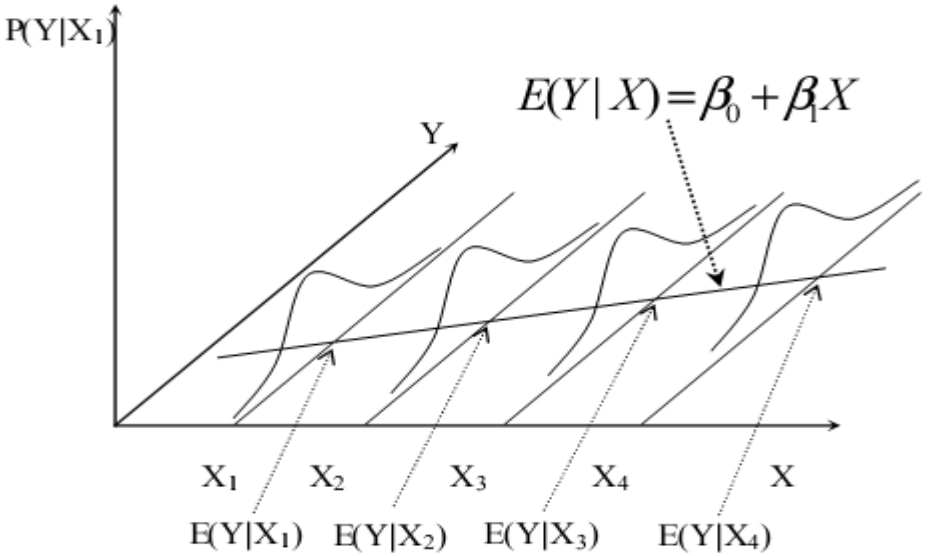
记$\hat{a}$和$\hat{b}$分别为参数a与b的拟合值，则一元线性回归模型为:

$$\hat{y}= \hat{a}+ \hat{b}x \qquad (2.2)$$

（2.2）式代表x与y之间相关关系的拟合直线，称为**回归直线**； $\hat{y}$是y的**估计值**，亦称**回归值**。



一般情况下的总体回归模型



假定条件下的总体回归模型

1)所有的Y的分布的均值都正好在一条直线上，称之为总体的（真实的）回归直线：$E(Y|X) = \beta_0 + \beta_1 X$
总体参数β_0和β_1确定了该直线，他们要通过样本信息估计。

2）所有的Y分布都具有同样的形状。这意味着对所有的$X_i(i = 1,2,\dots,n)$,概率分布$P(Y_i|X_i)$都有着相同的方差σ^2，即$Var(Y_i|X_i)=\sigma^2, i = 1,2,\dots,n$。

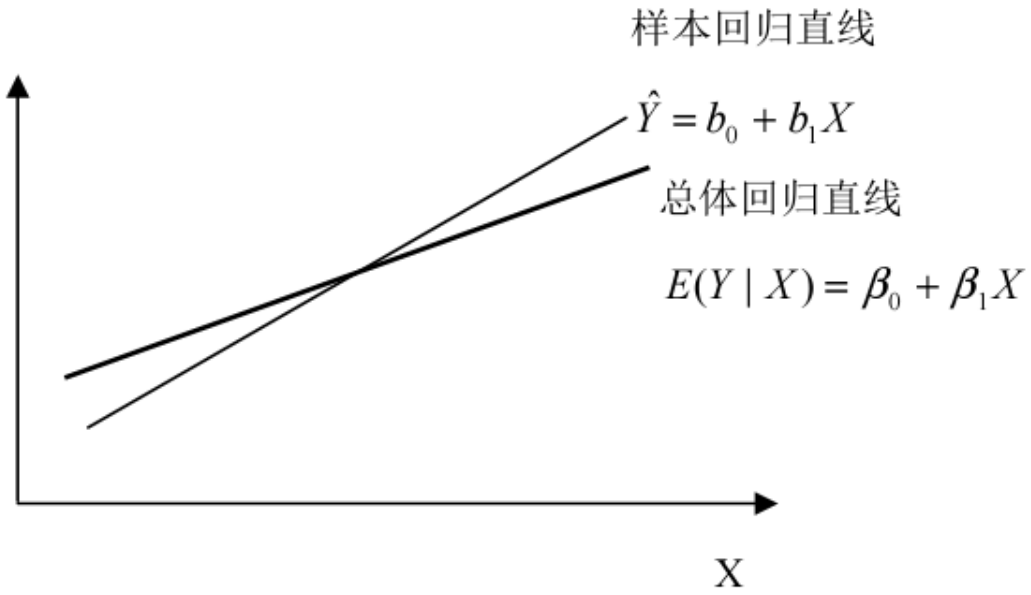
3）随机变量Y是相互独立的。也就是说,Y2和Y1没有统计关系，Y3和Y4也没有统计关系，等等。

4）给定X时Y分布的形状是正态的，即Y服从正态分布。

我们把符合以上假定的回归模型成为标准(古典）的回归模型。

•回归分析的主要任务就是要采用适当的方法，充分利用样本所提供的信息，使得样本回归函数尽可能地接近于真实的总体回归函数。

•所估计的样本回归直线都不可能与真实的总体回归直线完全一致。



（一）最小二乘估计

$$y_a = a + bx_a + \varepsilon_a$$

① 参数a与b的最小二乘拟合原则要求y_i与$\hat{y}_i$的误差e_i的平方和达到最小，即

$$Q = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2 \rightarrow min$$

② 根据取极值的必要条件，有

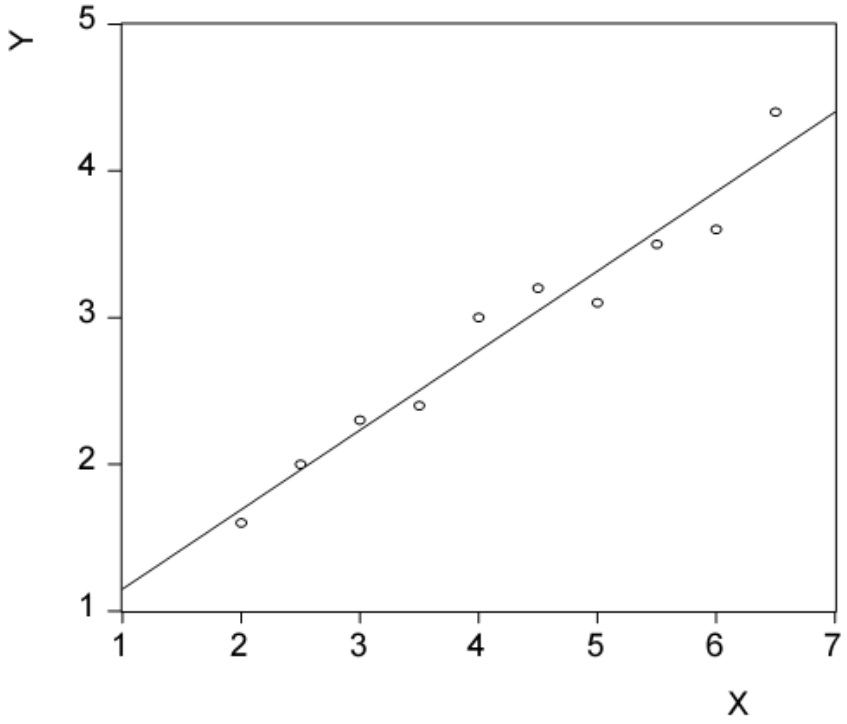
$$\begin{cases} \sum_{i=1}^{n} (y_i - a - bx_i) = 0 \\ \sum_{i=1}^{n} (y_i - a - bx_i)\, x_i = 0 \end{cases}$$

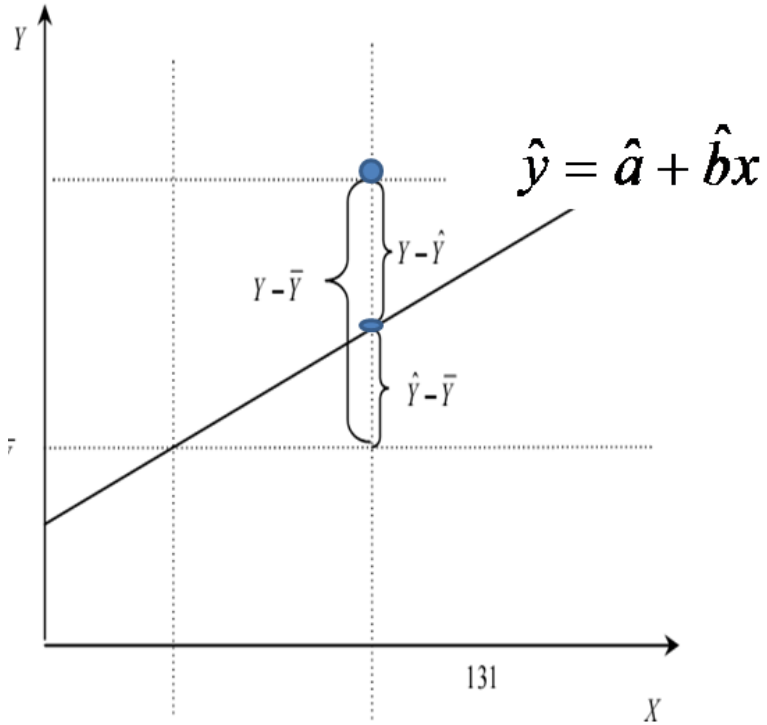
（一）最小二乘估计

$$y_a = a + bx_a + \varepsilon_a$$

$$\sum \varepsilon_i^2 \rightarrow min$$

观测值的散点图及其拟合直线





③ 解上述正规方程组，得到参数a与b的拟合值

$$\hat{a} = \bar{y} - \hat{b}\bar{x}$$

$$\hat{b} = \frac{L_{xy}}{L_{xx}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

x	0	1	2	3	4
y	1	4	3	8	9

$$\hat{y} = 1 + 2x \quad \hat{y} = \begin{pmatrix} 1 \\ 3 \\ 5 \\ 7 \\ 9 \end{pmatrix}$$

（二）一元线性回归模型的显著性检验

一元线性回归模型检验的种类

◆实际意义检验

参数估计值的符号和取值范围

消费支出与可支配收入：
如果估计出来的 b小于 0

◆统计检验

检验样本回归方程的可靠性

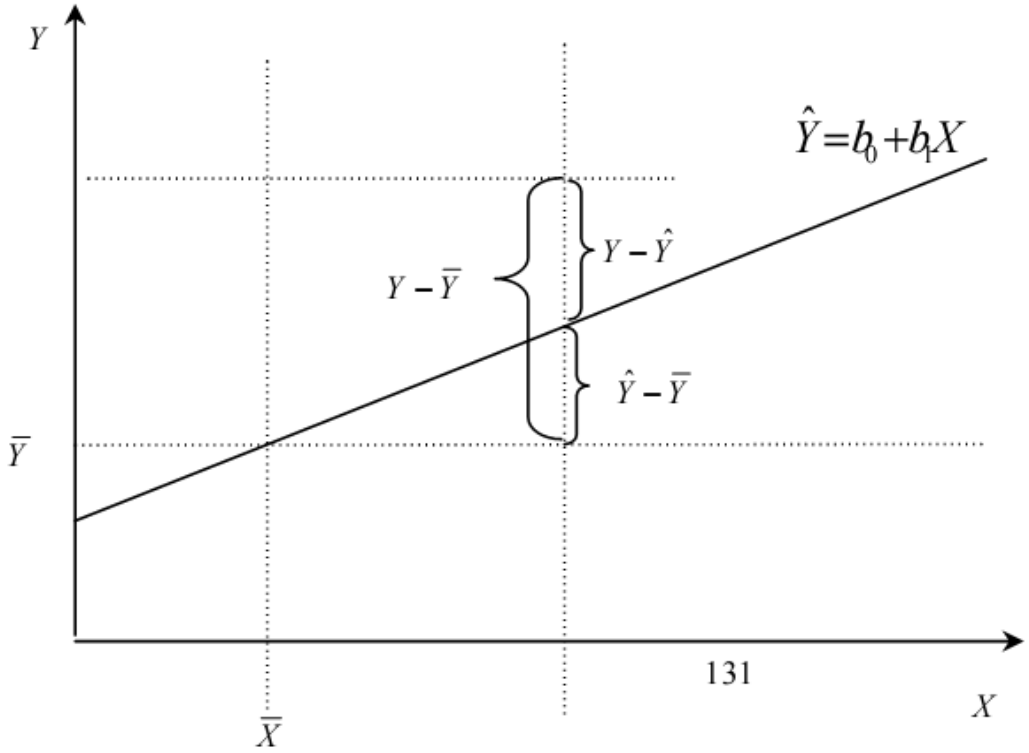
$$\hat{y} = \hat{a} + \hat{b}x$$

支出　　收入

- 拟合优度检验；
- 相关系数检验；
- 参数显著性检验(t检验)；
- 回归方程显著性检验（F 检验）

1 拟合优度检验

所谓拟合程度，是指样本观测值聚集在样本回归直线周围的紧密程度。判断回归模型拟合程度优劣最常用的数量指标是判定系数（**Coefficient of Determination**）



$$R^2=\frac{SS_R}{SS_T}$$

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

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

总的离差平方和：

在回归分析中，表示y的n次观测值之间的差异，记为

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

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

$SS_E=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$ 称为误差平方和或剩余平方和

$SS_R=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$ 称为回归平方和

各个样本观测点与样本回归直线靠得越紧，SS_R在 SS_T 中所占的比例就越大。因此，可定义这一比例为**判定系数**，即有：

$$R^2 = \frac{SS_R}{SS_T}$$

性质：

1、具有非负性，分子分母均是不可能为负值。

2、判定系数的取值范围为$0 \leq R^2 \leq 1$。

3、判定系数是样本观测值的函数，它也是一个统计量。

2 相关系数的显著性检验

X和 Y之间真实的线性相关程度用总体相关系数ρ来表示

$$\rho = \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$$

由于总体未知，ρ无法计算，我们利用样本相关系数

$$r = S_{XY}/S_X S_Y$$

```
x1<-c(171,175,159,155,152,158,154,164,168,166,159,164);
x2<-c(57,64,41,38,35,44,41,51,57,49,47,46);
```

2 相关系数的显著性检验

（1）建立检验假设：$H_0: \rho = 0$；$H_1: \rho \neq 0$

（2）计算样本相关系数 r；

r=cor(x1,x2)

（3）根据给定的显著性水平α（=0.05）和样本容量 n，计算相关系数r的t值：

$$t_r = \frac{r = 0}{\sqrt{\frac{1 - r^2}{n - 2}}}$$

查相关系数表得到临界值 r 。

n=length(x1) ; tr=r/sqrt((1-r^2)/(n-2))

（4）若|r|大于临界值，则 X与 Y有显著的线性关系，否则 X 与Y 的线性相关关系不显著。

cor.test(x1,x2)

3 回归系数的显著性检验（t检验）

根据样本估计的结果对总体回归参数的有关假设进行检验

1、提出假设。

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

2、计算回归参数t统计量值

$$t = \frac{b_1 - \beta_1}{S(b_1)} = \frac{b_1 - 0}{S(b_1)} = \frac{b_1}{S(b_1)}$$

3、根据给定的显著水平α确定临界值，或者计算t值所对应的p值。

4、做出判断。

4 回归方程的显著性检验

① 方法：*F* 检验法。

② 总的离差平方和：在回归分析中，表示*y*的*n*次观测值之间的差异，记为

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

③ 统计量*F*

$$F = SS_R/\frac{SS_E}{n-2}$$

④ *F* 越大，模型的效果越佳。统计量$F\sim F$（1，$n-2$）。在显著水平α下，若$F>F_\alpha$，则认为回归方程效果在此水平下显著。一般地，当$F<F_{0.10}(1,n-2)$时，则认为方程效果不明显。

表 4-1　1978—2008 年税收与财政收入数据（数据见 RstatM. xls：d4. 3）

年份	y	x	年份	y	x
1978	11. 326 2	5. 192 8	1994	52. 181 0	51. 268 8
1979	11. 463 8	5. 378 2	1995	62. 422 0	60. 380 4
1980	11. 599 3	5. 717 0	1996	74. 079 9	69. 098 2
1981	11. 757 9	6. 298 9	1997	86. 511 4	82. 340 4
1982	12. 123 3	7. 000 2	1998	98. 759 5	92. 628 0
1983	18. 669 5	7. 555 9	1999	114. 440 8	106. 825 8
1984	16. 428 6	9. 473 5	2000	133. 952 3	125. 815 1
1985	20. 048 2	20. 407 9	2001	163. 860 4	153. 013 8
1986	21. 220 1	20. 907 3	2002	189. 036 4	176. 364 5
1987	21. 993 5	21. 403 6	2003	217. 152 5	200. 173 1
1988	23. 572 4	23. 904 7	2004	263. 964 7	241. 656 8
1989	26. 649 0	27. 274 0	2005	316. 492 9	287. 785 4
1990	29. 371 0	28. 218 7	2006	387. 602 0	348. 043 5
1991	31. 494 8	29. 901 7	2007	513. 217 8	456. 219 7
1992	34. 833 7	32. 969 1	2008	613. 303 5	542. 196 2
1993	43. 489 5	42. 553 0			

（1）读入数据##在d4.3中选取B1：C32区域，然后拷贝

```
> yx=read.table("clipboard",header=T)
> attach(yx)    ##解析变量成y，x
```

（2）拟合模型

```
> fm=lm(y~x)
> fm

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)          x
     -1.197       1.116
```

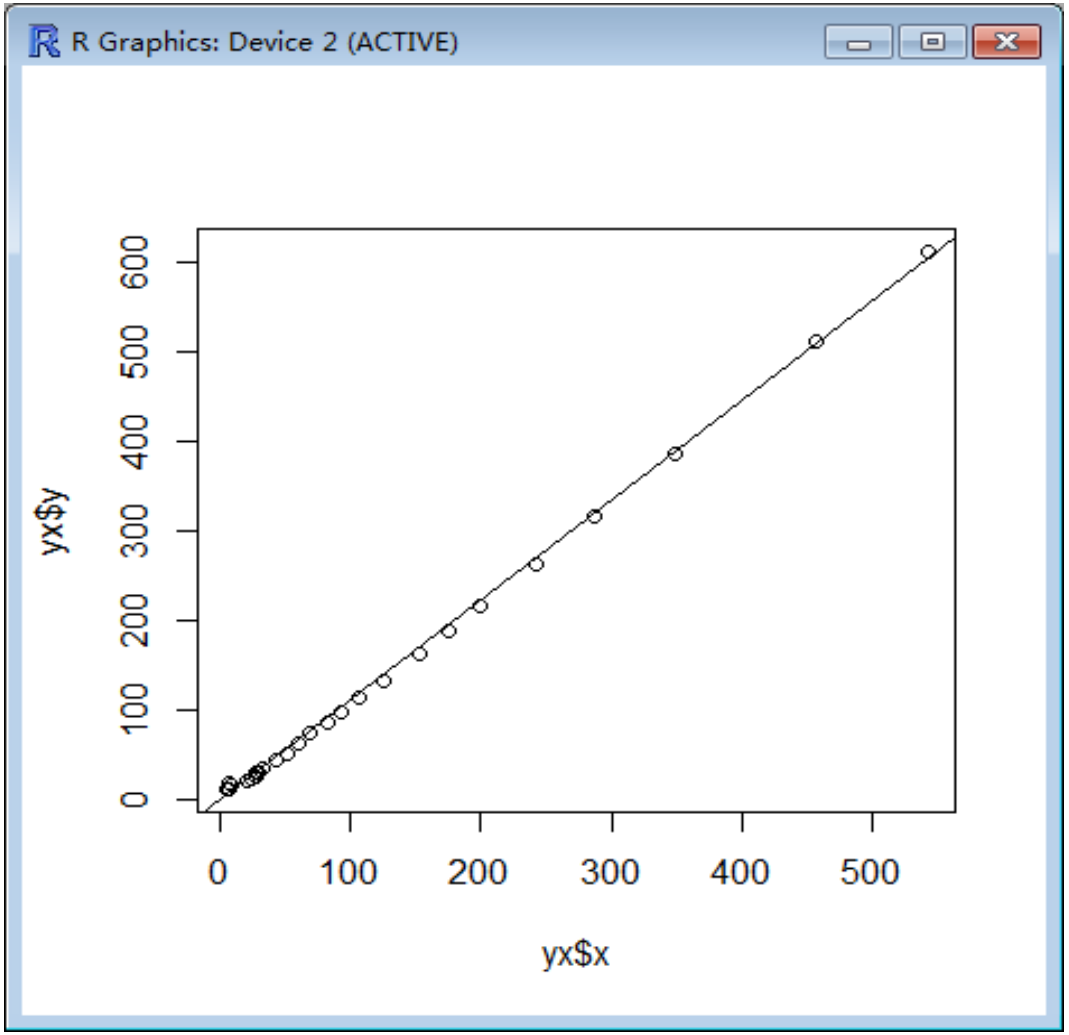
于是得到回归方程：

$$\hat{y} = -1.197 + 1.116x$$

（3）作回归直线

> plot(x,y) # 散点图

> abline(fm) # 添加回归线



（4）回归方程的假设检验

1）模型的方差分析

```
> anova(fm)
Analysis of Variance Table

Response: y
          Df   Sum Sq   Mean Sq   F value          Pr(>F)
x          1   712077    712077     27428  < 2.2e-16  ***
Rds       29   753       26
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’
1
```

由于p<0.05，于是在0.05水平处拒绝原假设，即本例回归系数有统计学意义，x与y间存在直线回归关系。

2）回归系数的显著性检验（t检验）

```
> summary(fm)
Call:
lm(formula = y ~ x)
Residuals:
   Min      1Q  Median      3Q     Max
-6.630  -3.692  -1.535   5.338  11.432

Coefficients:
             Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)  -1.19656   1.16125     -1.03    0.311
yx$x          1.11623   0.00674    165.61    <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.095 on 29 degrees of freedom
Multiple R-squared: 0.9989,    Adjusted R-squared: 0.9989

F-statistic: 2.743e+04  on 1 and 29 DF,  p-value: < 2.2e-16
```

由于p<0.05，于是在0.05水平处拒绝原假设，即本例回归系数有统计学意义，x与y间存在回归关系。

二、多元线性回归模型

1 多元线性回归模型的结构形式为

$$y_a = \beta_0 + \beta_1 x_{1a} + \beta_2 x_{2a} + \cdots + \beta_k x_{ka} + \varepsilon_a \quad （2.11）$$

式中： β_0，$\beta_2 \cdots \beta_k$为待定参数；ε_a为随机变量。

例子：Y=住房的当前市场价值

X1=居住面积

X2=位置

X3=建筑质量

线性模型：

$$y_a = \beta_0 + \beta_1 x_{1a} + \beta_2 x_{2a} + \cdots + \beta_k x_{ka} + \varepsilon_a$$

观测值线性模型：

$$y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_k x_{1k} + \varepsilon_a$$

$$y_2 = \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_k x_{2k} + \varepsilon_a$$

$$\vdots$$

$$y_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_k x_{nk} + \varepsilon_a$$

性质：

$1.E(\varepsilon_i) = 0$

$2.Var(\varepsilon_i) = \delta^2$（常数）

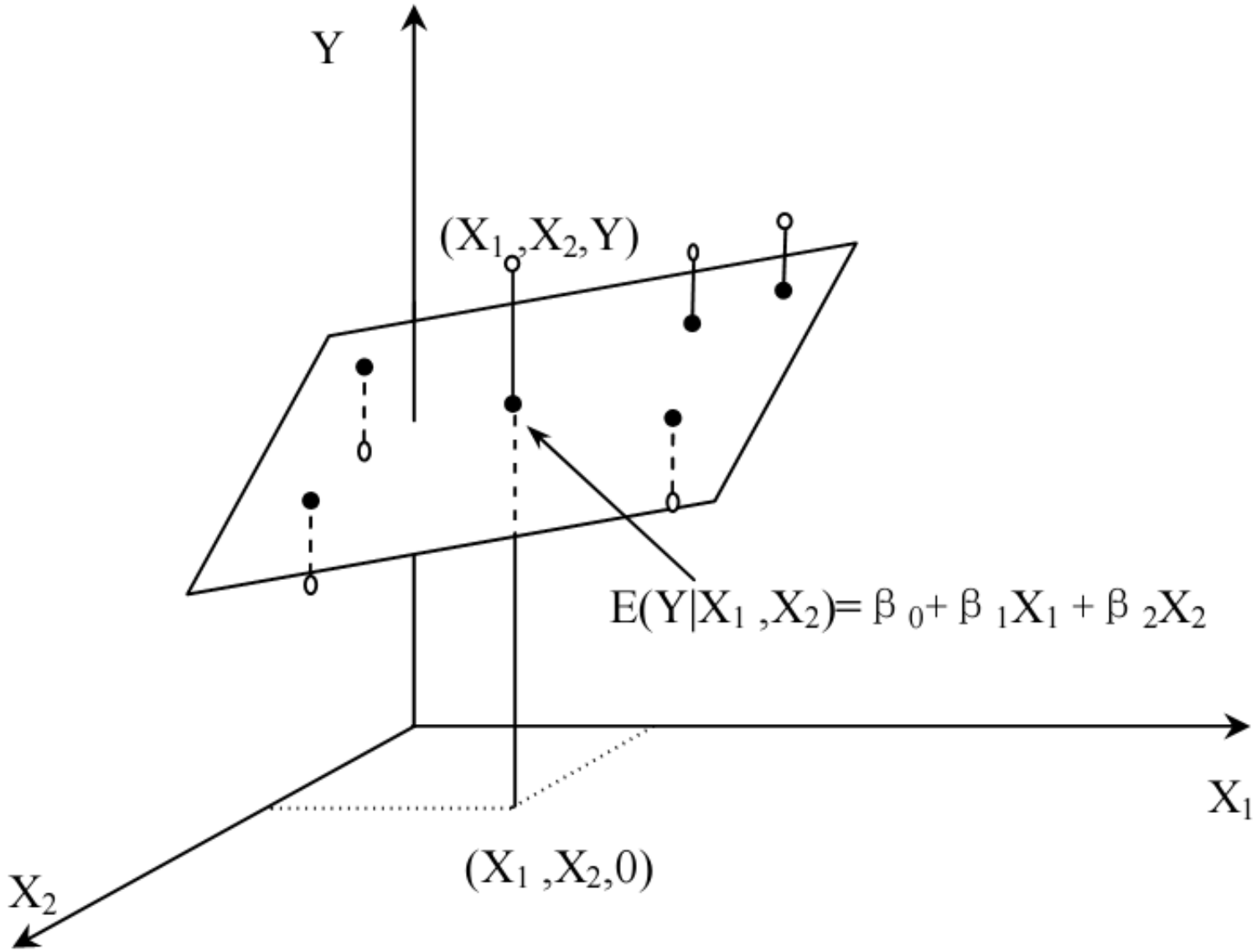
$3.Cov(\varepsilon_i, \varepsilon_j) = 0, i \neq j$

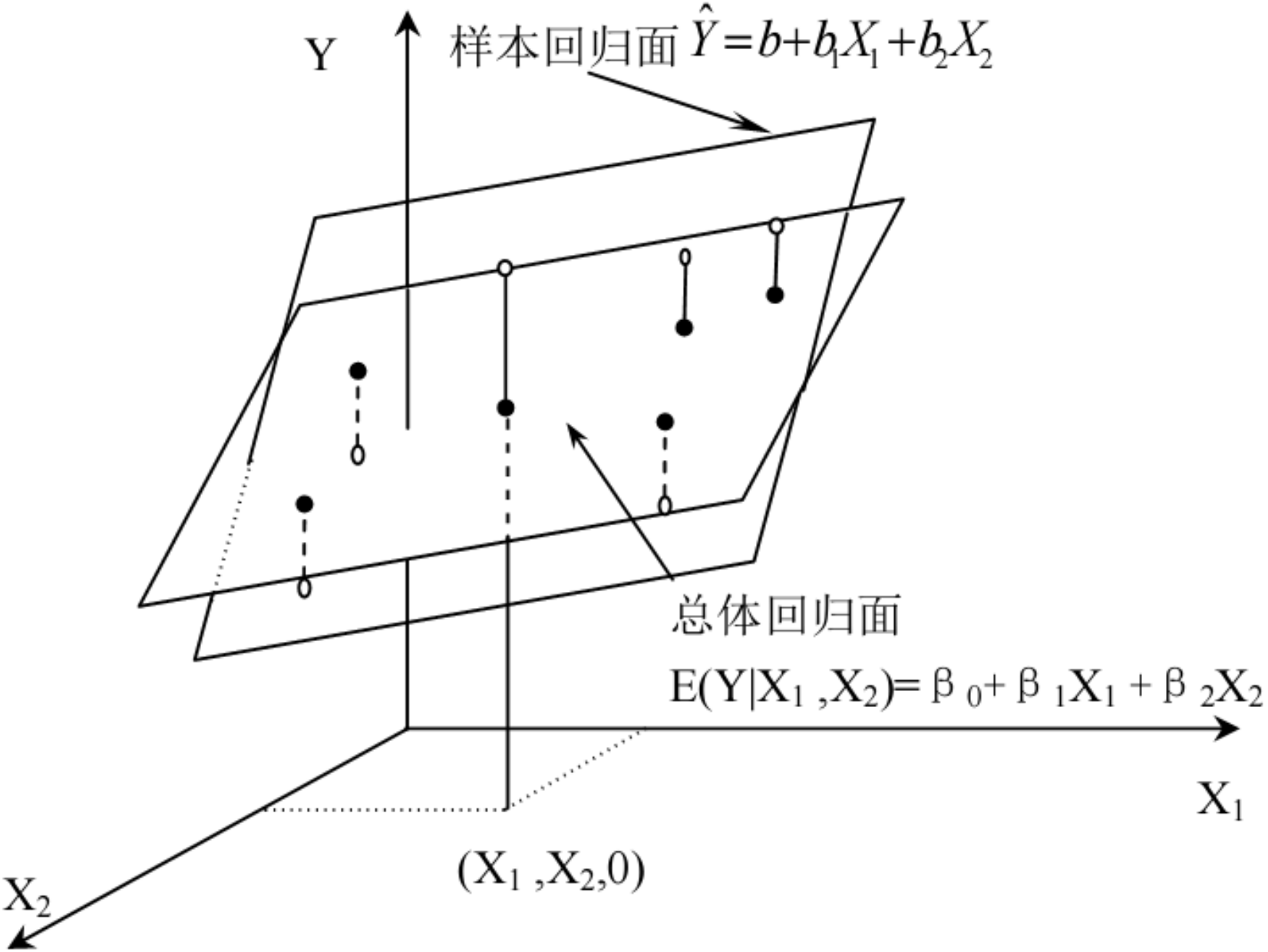
用矩阵形式表达

$$Y=\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad X=\begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ 1 & x_{31} & x_{32} & \cdots & x_{3k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} \quad \beta=\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}$$

$$y_1=\beta_0+\beta_1 x_{11}+\beta_2 x_{12}+\cdots+\beta_k x_{1k}+\varepsilon_a$$
$$y_2=\beta_0+\beta_1 x_{21}+\beta_2 x_{22}+\cdots+\beta_k x_{2k}+\varepsilon_a$$
$$\vdots$$
$$y_n=\beta_0+\beta_1 x_{n1}+\beta_2 x_{n2}+\cdots+\beta_k x_{nk}+\varepsilon_a$$

$$\boldsymbol{Y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{E}$$


Y
(X1 ,X2, Y)
E(Y|X1 ,X2)= β 0+ β 1X1 + β 2X2
X1
(X1 ,X2,0)
X2


Y
样本回归面 $\hat{Y}=b+b_1X_1+b_2X_2$
总体回归面
$E(Y|X_1,X_2)=\beta_0+\beta_1X_1+\beta_2X_2$
X_1
X_2
$(X_1,X_2,0)$

2 多元线性回归模型的基本假定

	假定名称	假定条件	说明
对扰动项ε的假定	1、正态性	$\varepsilon \sim N(0,\sigma^2)$ 且 Cov(εi, εj)=0 (i≠j)	$Y \sim N(\beta_0+\beta_1 X_1+\mathrm{L}++\beta_p X_p, \sigma^2)$且 $Cov(Y_i,Y_j)=0 \quad (i \neq j)$
	2、零均值		
	3、同方差		
	4、互独立		
对自变量X的假定	5、非随机	解释是确定性变量	
	6、不相关	解释变量间不存在线性相关关系	
对X与ε的假定	7、不相关	Cov(X, ε)=0	

3 回归方程的估计:

如果$b_0, b_1, b_2, \cdots, b_k$分别为$\beta_0, \beta_1, \beta_2, \cdots, \beta_k$的拟和值，则回归方程为

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k \qquad (2.12)$$

b_0为常数，$b_1, b_2, \ldots b_k$称为偏回归系数。偏回归系数的意义是：当其他自变量都固定时，自变量x_i每变化一个单位而使因变量平均改变的数值。

偏回归系数的推导过程:根据最小二乘法原理，

估计值$b_i(i=0,1,2,\ldots,k)$

应该使

$$SS_T=\sum_{a=1}^{n}(y_a-\widehat{y_a})^2=\sum_{a=1}^{n}(y_a-(b_0+b_1x_{1a}+b_2x_{2a}+\ldots+b_kx_{ka}))^2\rightarrow min \tag{2.13}$$

由求极值的必要条件得

$$\begin{cases}\dfrac{\partial Q}{\partial b_0}=-2\sum_{a=1}^{n}(y_a-\widehat{y_a})=0\\ \dfrac{\partial Q}{\partial b_j}=-2\sum_{a=1}^{n}(y_a-\widehat{y_a})x_{ja}=0\ \ (j=1,2,\ldots,k)\end{cases} \tag{2.14}$$

$$
\begin{cases}
nb_0+(\sum_{a=1}^{n}x_{1a})b_1+(\sum_{a=1}^{n}x_{2a})b_2+\cdots+(\sum_{a=1}^{n}x_{ka})b_k=\sum_{a=1}^{n}y_a \\
(\sum_{a=1}^{n}x_{1a})b_0+(\sum_{a=1}^{n}x_{1\alpha}^2)b_1+(\sum_{a=1}^{n}x_{1a}x_{2a})b_2+\cdots+(\sum_{a=1}^{n}x_{1a}x_{ka})b_k=\sum_{a=1}^{n}x_{1a}y_a \\
(\sum_{a=1}^{n}x_{2a})b_0+(\sum_{a=1}^{n}x_{1a}x_{2a})b_1+\sum_{a=1}^{n}(x_{2a}^2)b_2+\cdots+(\sum_{a=1}^{n}x_{2a}x_{ka})b_k=\sum_{a=1}^{n}x_{2a}y_a \\
\cdots\cdots\cdots\cdots \\
(\sum_{a=1}^{n}x_{ka})b_0+(\sum_{a=1}^{n}x_{1a}x_{ka})b_1+(\sum_{a=1}^{n}x_{2a}x_{ka})b_2+....+(\sum_{a=1}^{n}x_{ka}^2)b_k=\sum_{a=1}^{n}x_{ka}y_a
\end{cases}
\qquad (2.15)
$$

方程组（2.15）式称为正规方程组。

引入矩阵 $AX=b$

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ 1 & x_{31} & x_{32} & \cdots & x_{3k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}$$

$$A = X^T X = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & x_{31} & \cdots & x_{n1} \\ x_{12} & x_{22} & x_{32} & \cdots & x_{n2} \\ \vdots & \vdots & \vdots & & \vdots \\ x_{1k} & x_{2k} & x_{3k} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ 1 & x_{31} & x_{32} & \cdots & x_{3k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}$$

$$= \begin{bmatrix} n & \sum_{a=1}^{n} x_{1a} & \sum_{a=1}^{n} x_{2a} & \cdots & \sum_{a=1}^{n} x_{ka} \\ \sum_{a=1}^{n} x_{1a} & \sum_{a=1}^{n} x_{1\alpha}^{2} & \sum_{a=1}^{n} x_{1a}x_{2a} & \cdots & \sum_{a=1}^{n} x_{1a}x_{ka} \\ \sum_{a=1}^{n} x_{2a} & \sum_{a=1}^{n} x_{1a}x_{2a} & \sum_{a=1}^{n} x_{2a}^{2} & \cdots & \sum_{a=1}^{n} x_{2a}x_{ka} \\ \vdots & \vdots & \vdots & & \vdots \\ \sum_{a=1}^{n} x_{ka} & \sum_{a=1}^{n} x_{1a}x_{ka} & \sum_{a=1}^{n} x_{2a}x_{ka} & \cdots & \sum_{a=1}^{n} x_{ka}^{2} \end{bmatrix}$$

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad b = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$$

$$B = X^T Y = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ x_{11} & x_{12} & x_{13} & \cdots & x_{1n} \\ x_{21} & x_{22} & x_{23} & \cdots & x_{2n} \\ \vdots & \vdots & \vdots & & \vdots \\ x_{k1} & x_{k2} & x_{k3} & \cdots & x_{kn} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \sum_{a=1}^{n} y_a \\ \sum_{a=1}^{n} x_{1a} y_a \\ \sum_{a=1}^{n} x_{2a} y_a \\ \vdots \\ \sum_{a=1}^{n} y_{ka} y_a \end{bmatrix}$$

则正规方程组（2.15）式可以进一步写成矩阵形式 $Ab = B$

求解得 $b = A^{-1}B = (X^T X)^{-1} X^T Y$

$$b = A^{-1}B = (X^T X)^{-1} X^T Y$$

计算最小二乘估计、残差、残差平方和

x	0	1	2	3	4
y	1	4	3	8	9

$$A = X^T X = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 & 4 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{pmatrix} = \begin{pmatrix} 5 & 10 \\ 10 & 30 \end{pmatrix} \qquad A^{-1} = \begin{pmatrix} 0.6 & -0.2 \\ -0.2 & 0.1 \end{pmatrix}$$

$$B = X^T Y = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 & 4 \end{pmatrix} \begin{pmatrix} 1 \\ 4 \\ 3 \\ 8 \\ 9 \end{pmatrix} = \begin{pmatrix} 25 \\ 70 \end{pmatrix}$$

x	0	1	2	3	4
y	1	4	3	8	9

$$b = A^{-1}B = \begin{pmatrix} 1 \\ 2 \end{pmatrix} \qquad y = 1 + 2x \qquad y = \begin{pmatrix} 1 \\ 3 \\ 5 \\ 7 \\ 9 \end{pmatrix}$$

残差

$$\varepsilon = y - y = \begin{pmatrix} 1 \\ 4 \\ 3 \\ 8 \\ 9 \end{pmatrix} - \begin{pmatrix} 1 \\ 3 \\ 5 \\ 7 \\ 9 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ -2 \\ 1 \\ 0 \end{pmatrix}$$

残差平方和

$$\varepsilon'\varepsilon = \begin{pmatrix} 0 & 1 & -2 & 1 & 0 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \\ -2 \\ 1 \\ 0 \end{pmatrix} = 6$$

4 回归方程的显著性检验

① 方法：F 检验法。

② 总的离差平方和：在回归分析中，表示y的n次观测值之间的差异，记为

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

③ 统计量F

$$F = \frac{SS_R}{\mathrm{p}} \Big/ \frac{SS_E}{n-\mathrm{p}-1}$$

④ F 越大，模型的效果越佳。统计量$F \sim F$（p，n-p-1）。在显著水平α下，若$F>F_\alpha(p，n\text{-}p\text{-}1)$或者(p<α)，则认为回归方程效果在此水平下显著。

拟合优度检验

$$R^2 = \frac{SS_R}{SS_T}$$

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

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

参数的显著性检验（t检验）

1、提出假设

$$H_{0j}: \beta_j = 0 \; ; H_{1j}: \beta_j \neq 0$$

2、计算回归参数t统计量值

$$t_j = \frac{\hat{\beta}_j - \beta_j}{s(\hat{\beta}_j)} \quad (j=1,\ldots,p) \sim df(n-p-1)$$

3、计算t值所对应的p值，做出判断。

当$|t_j| \geq t_{1-\alpha/2}$时，拒绝零假设$H_{0j}$，认为$\beta_j$显著不为0，自变量$x_j$对因变量y的线性效果显著;

当$|t_j| < t_{1-\alpha/2}$时，不拒绝零假设H_{0j}，认为β_j=0，自变量x_j对因变量y的线性效果不显著.

5 例题表 8-40 某商品的统计资料

年份	需求量Y（吨）	价格X1（元）	收入X2（元）
1	59190	23.56	76200
2	65450	24.44	91200
3	62360	32.07	106700
4	64700	32.46	111600
5	67400	31.15	119000
6	64440	34.14	129200
7	68000	35.3	143400
8	72400	38.7	159600
9	75710	39.63	180000
10	70680	46.68	193000

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.950	.902	.874	1738.9846

a Predictors: (Constant), X2, X1

R=0.950，说明 Y 与自变量 X1、X2 之间的相关程度为 95.0%。

样本判定系数0.902 说明 Y的变动有 90.2%可以由自变量 X1 和 X2 解释。

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	195318937.424	2	97659468.712	32.294	.000
	Residual	21168472.576	7	3024067.511		
	Total	216487410.000	9			

a Predictors: (Constant), X2, X1

b Dependent Variable: Y

1、提出假设

$$\mathrm{H}_0\colon \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

$$\mathrm{H}_1\colon \beta_j \text{不全为} 0\ (j = 1,2,\ldots,p)$$

2、$F = \frac{SS_R}{\mathrm{p}} / \frac{SS_E}{n-\mathrm{p}-1} \sim F(p,\ n-p-1)$

若$F > F_\alpha(p,\ n-p-1)$ (或者$\mathrm{p} < \alpha$)，就拒绝H_0, 回归方程显著成立。

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	62650.928	4013.010		15.612	.000
	X1	-979.057	319.784	-1.381	-3.062	.018
	X2	.286	.058	2.211	4.902	.002

a Dependent Variable: Y

$$\hat{Y} = 62650.928 - 979.057X_1 + 0.286X_2$$

考察财政收入y和国内生产总值x1，税收x2，进出口贸易总额x3，经济活动人口x4之间的数量关系(例题4-4)

y	x1	x2	x3	x4
11.3262	36.241	5.1928	3.55	406.82
11.4638	40.382	5.3782	4.12	415.92
11.5993	45.178	5.717	5.7	429.03
11.7579	48.603	6.2989	8.904	441.65
12.1233	53.018	7.0002	12.801	456.74
18.6695	59.574	7.5559	15.903	467.07
16.4286	72.067	9.4735	18.202	484.33
20.0482	89.891	20.4079	20.667	501.12
21.2201	102.014	20.9073	26.019	515.46
21.9935	119.545	21.4036	32.202	530.6
23.5724	149.223	23.9047	41.6	546.3
26.649	169.178	27.274	49.802	557.07

```
##在d4.4中选取B1：F32区域，然后拷贝
yX=read.table("clipboard",header=T)
(fm=lm(y~x1+x2+x3+x4,data=yX))
```

Call:
lm(formula = y ~ x1 + x2 + x3 +x4)

Coefficients:

(Intercept)	x1	x2	x3	x4
23.5321088	-0.0033866	1.1641150	0.0002919	-0.0437416

于是得到多元线性回归方程：

$$\hat{y} = 23.5321088 - 0.0033866x_1 + 1.164115x_2 + 0.0002919x_3 - 0.0437416x_4$$

标准化后的模型：

```
>library(mvstats)
> coef.sd(fm)
$coef.sd
         x1          x2          x3          x4
-0.0174513678  1.0423522972  0.0009628564 -0.0371053994
```

由标准化偏回归系数可见，税收对财政收入的线性影响最大。

```
> anova(fm)
Analysis of Variance Table

Response: y
          Df Sum Sq   Mean Sq    F value        Pr(>F)
x1         1 694627    694627   89259.0016   < 2.2e-16 ***
x2         1  17803    17803     2287.6286    < 2.2e-16 ***
x3         1     24       24        3.0569         0.0922 .
x4         1    174      174       22.2954       7.005e-05 ***
Residuals 26    202        8
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

由方差分析结果可见，模型的p<0.0001，故本例回归模型是有意义的。

```
> summary(fm) # 多元线性回归系数t检验

Call:
lm(formula = y ~ x1 + x2 + x3 + x4)

Residuals:
    Min      1Q  Median      3Q     Max
-5.0229 -2.1354  0.3297  1.2639  6.9690

Coefficients:
             Estimate       Std.Error      t value    Pr(>|t|)
(Intercept) 23.5321088     4.5990714     5.117      2.47e-05 ***
     x1      -0.0033866     0.0080749    -0.419      0.678
     x2       1.1641150     0.0404889     28.751     < 2e-16 ***
     x3       0.0002919     0.0085527     0.034      0.973
     x4      -0.0437416     0.0092638     -4.722     7.00e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’
0.1 ‘ ’ 1

Residual standard error: 2.79 on 26 degrees of freedom
Multiple R-squared: 0.9997,     Adjusted R-squared: 0.9997
F-statistic: 2.289e+04 on 4 and 26 DF,  p-value: < 2.2e-16
```

由t检验结果可见，偏回归系数b2，b4的p值都小于0.01，可认为解释变量税收x2和经济活动人口x4显著；b1，b3的p值大于0.50，不能否定b1=0，b3=0的假设。可认为国内生产总值x1和进出口贸易总额x3对财政收入y没有显著的影响。

我们可以看到，国内生产总值x1、经济活动人口x4所对应的偏回归系数都为负，这与经济现实是不相符的。出现这种结果的可能原因是这些解释变量之间存在高度的共线性。

6 多重共线性： 线性回归模型中的若干解释变量或全部解释变量的样本观测值之间具有某种线性关系。

多重共线性产生的主要后果有：

1、各个解释变量的影响很难精确鉴别。

某个解释变量的变动将也引起其它解释变量的变动，从而难以区分各个解释变量对被解释变量的影响大小。

2、模型回归参数估计量的方差会很大

这将使得进行显著性检验时认为回归参数的值与零无显著差异。从而导致将相应的解释变量从模型中剔除，但这并不是因为该解释变量对被解释变量无影响作用，而只是由于样本数据不适于精确区分各解释变量的单独影响。

3、模型参数的估计量不稳定

对删除或增加少量的观测值以及删除一个不显著的解释变量都可能非常敏感。

共线性检验

对于多于两个解释变量的模型，可以分别利用其中一个解释变量对其他解释变量进行线性回归，并计算出拟合优度R_1^2，R_2^2，... R_k^2，如果某一个拟合优度较大，则说明对应的解释变量与其他解释变量之间存在共线性。

如果在对多元线性回归模型进行统计检验时，发现参数估计值的太小，或者判定系数、F 检验值很大（p 值小）而各个偏回归参数的 t 检验值均偏小（其 p 值大于α），那么很有可能是因为解释变量之间存在多重共线性。

常用的检验方法：容限度法和方差扩大因子法。

共线性检验

容限度是由每个自变量 **Xj**作为因变量对其他自变量回归时得到的余差比例，即：

$$Tolerance_j = 1 - R_j^2$$

通常当容限度小于 **0.1**（这里**R >0.9**）时，多重共线性超过了容许界限.

方差扩大因子（以下简称 **VIF**）是容限度的倒数。即：

$$VIF_j = \frac{1}{Tolerance_j} = \frac{1}{1 - R_j^2}$$

它表示所对应的偏回归系数的方差由于多重共线性而扩大的倍数。

当容限度为 **0.1** 时，**VIF**为 **10**（倍）。通常当**VIF>10**时，便认为变量 **X** 与其他变量之间存在多重共线性。

表8-44　例8-2某商品需要量对价格和消费者收入回归的部分输出结果

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
(Constant)	62650.928	4013.010		15.612	.000		
X1	-979.057	319.784	-1.381	-3.062	.018	.069	14.559
X2	.286	.058	2.211	4.902	.002	.069	14.559

多重共线性的处理
- 删除不重要的解释变量
- 追加样本信息
- 利用非样本先验信息
- 改变解释变量的形式
- 逐步回归法

变量选择
- 变量多增加了模型的复杂度
- 计算量增大
- 估计和预测的精度下降
- 模型应用费用增加

自相关(Autocorrelation)

是对随机扰动项之间相互独立假定的违背，指扰动项序列相邻之间不是随机独立而是存在相关关系，又称为序列相关。自相关主要表现在时间序列中。

例如：t,t-1,t-2,…表示观测数据的时期，作为扰动项的下标。

因此对于线性回归模型：$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} \ldots + \beta_p X_{pt} + \varepsilon_t$

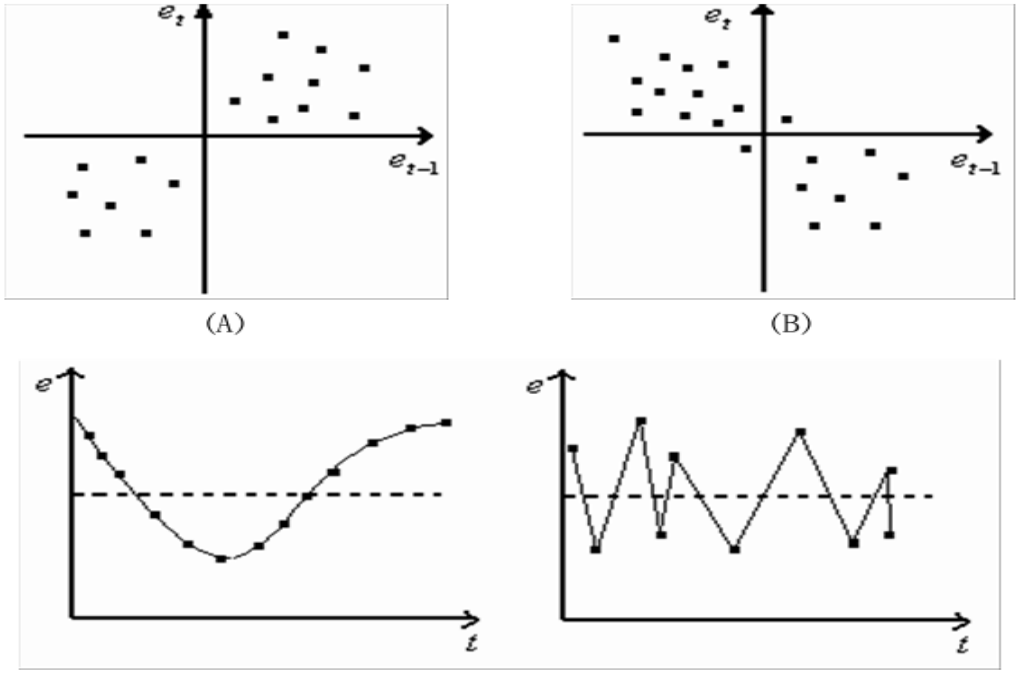
自相关可表示为：$Cov(\varepsilon_t, \varepsilon_{t-i}) \neq 0 (i = 1,2, \ldots, s)$

当扰动项存在自相关时，就违背了标准线性回归模型的基本假定，如果仍直接用 **OLS**法估计参数，将产生一系列后果。其主要影响后果有：

1、参数的估计量的方差增大，不再具有最小方差性。

2、导致常用的 **F**检验和 **t**检验失效。

3、如果不加处理地运用 **OLS**法估计模型参数，用此模型进行预测时会带来较大的方差的错误的解释。



- 产生自相关的原因有多种。如果自相关是由于遗漏重要变量，或者设定的模型形式错误，那么就应该引入新变量或者修改模型形式。
- 排除了上述原因后，经检验仍然存在自相关，就必须采用一定的方法来处理。
- 其基本思想是通过差分变换，对原始数据进行整理，变自相关为无自相关。常见的方法有广义差分法等。

回归变量的选择方法

全局择优法

对每组子集，RSS（残差平方和）越小、R2越大、校正R2越大、AIC BIC越小，模型越好。

```
> library(leaps)  ##安装包leaps
> varsel=regsubsets(y~ x1+ x2+ x3+ x4,data=yX)
> result=summary(varsel)
>data.frame(result$outmat,RSS=result$rss,R2=result$rsq,adjR2=result$adjr2,Cp=result$cp,BIC=result$bic)
```

回归变量选择的常用准则：

1.平均残差平方和最小准则 $RMS_p = \frac{RSS_p}{n-p}$

2.误差均方根最小准则 $MSE_p = \sqrt{RMS_p}$

3.校正决定系数（复相关系数平方）（越大越好）

$$adjR^2 = 1 - \frac{n-1}{n-p}(1-R^2) = 1 - \frac{\left(\frac{RSS_p}{n-p}\right)}{\left(\frac{SS_T}{n-1}\right)} = 1 - \frac{n-1}{SS_T} RMS_p$$

以上三条准则实质上是等价的

4. C_p准则

$$C_p=\frac{RSS_p}{s^2}-(n-2p)=\frac{RSS_p}{RMS}-(n-2p)=\frac{(n-p)RMS_p}{RMS}-(n-2p)$$

这里，C 即criterion，p 为所选模型中变量的个数，C_p 接近 p 模型为最优。其中，为全模型的均方误差RMS。

5. AIC 准则 和 BIC 准则

$$AIC=nln\left(\frac{RSS_p}{n}\right)+2p$$

$$BIC=nln\left(\frac{RSS_p}{n}\right)+pln(n)$$

AIC和BIC选择变量的准则是：按“AIC或BIC愈小愈好”选取自变量。

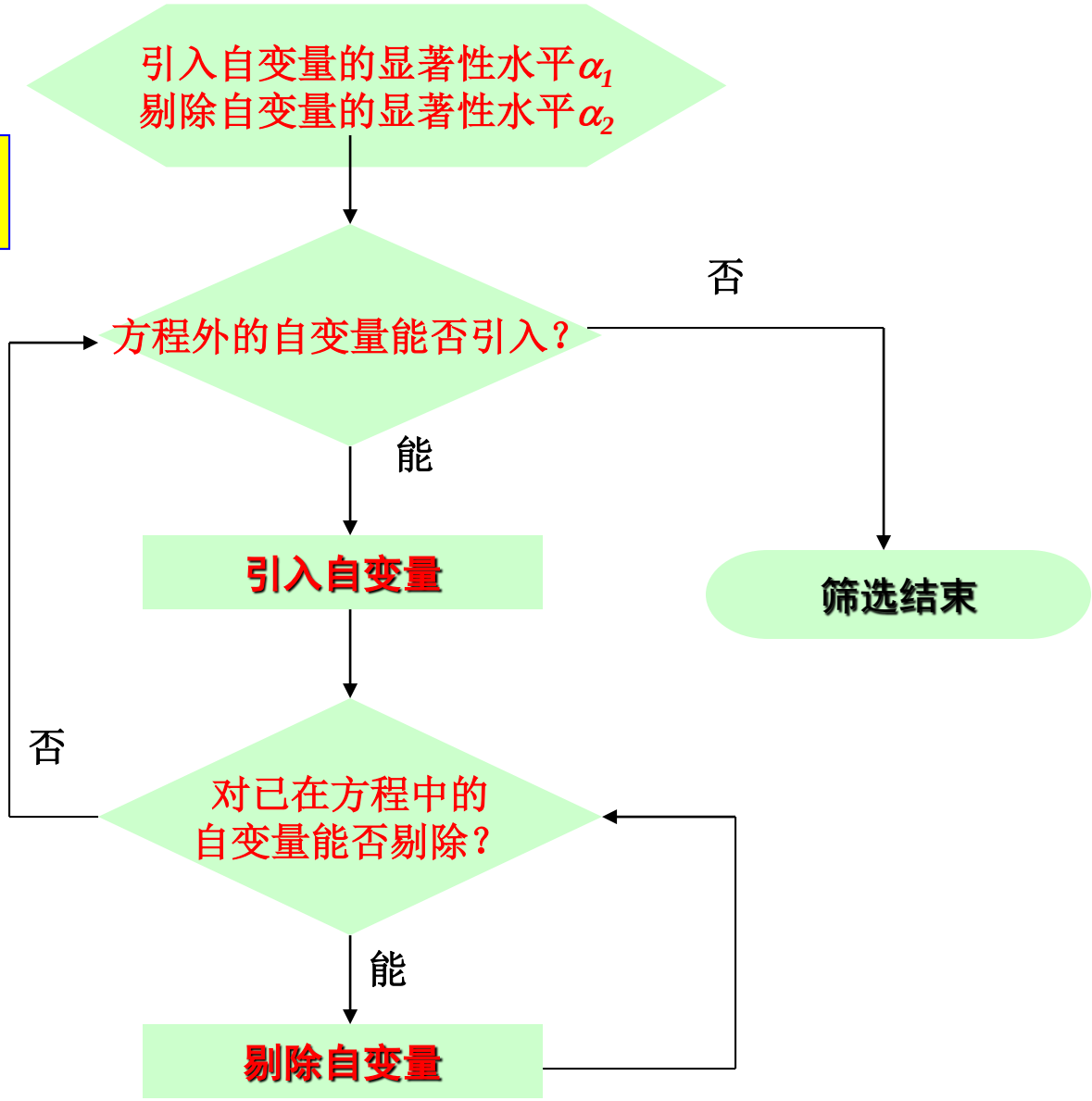
三、逐步回归分析

1. **前进法（forward selection）：**
   自变量从无到有、从少到多

2. **后退法（backward elimination）**
   先将全部自变量放入方程，然后逐步剔除

3. **逐步回归法（stepwise regression）**
   双向筛选：引入有意义的变量（前进法），
   剔除无意义变量（后退法）

逐步回归的基本步骤


引入自变量的显著性水平α_1
剔除自变量的显著性水平α_2
方程外的自变量能否引入？
否
能
引入自变量
筛选结束
否
对已在方程中的
自变量能否剔除？
能
剔除自变量

```
> fm.step=step(fm,direction="backward") #direction=" backward "表示向后引入法；
  Start:  AIC=68.15          #最开始四个变量都选入的AIC值为68.15
  y ~ x1 + x2 + x3 + x4

       Df   Sum of Sq    RSS        AIC
- x3   1       0.0          202.3        66          #只去除x3之后的AIC值为66.156
- x1   1       1.4          203.7        66          #只去除x1之后的AIC值为66.363
<none>                      202.3        68          #不做任何操作的AIC值为68.154
-x4    1     173.5        375.8        85          #只去除x4之后的AIC值为85.351
-- x2  1     6433.1     6635.4       174.352       #只去除x2之后的AIC值为174.352
```

```
#按照AIC值越小模型效果约好的原则，选择去掉x3的模型，即y ~ x1 + x2 + x4，进行下一步
Step:  AIC=66.16          #去除x3之后的AIC值为66.156
y ~ x1 + x2 + x4          #去除x3之后的模型

       Df  Sum of Sq     RSS        AIC
- x1    1     1.5          203.9        64     #在y ~ x1 + x2 + x4基础上再去除x1的AIC值为64.390
<none>                     202.3        66     #在y ~ x1 + x2 + x4基础上不做操作的AIC值为66.156
- x4    1    197.3         399.6        85.    #在y ~ x1 + x2 + x4基础上再去除x4的AIC值为85.253
-x2     1   7382.2        7584.5      176     #在y ~ x1 + x2 + x4基础上再去除x2的AIC值为176.496
```

```
Step:  AIC=64.39#去除x3、x1之后的AIC值为64.39
y ~ x2 + x4
       Df   Sum of Sq     RSS        AIC
<none>                     204          64
-x4      1     549          753          103
-x2      1   367655     367859       295
```

三、非线性回归模型

指数曲线 $y = de^{bx}$ $y' = lny$ $x' = x$ $a = lnd$

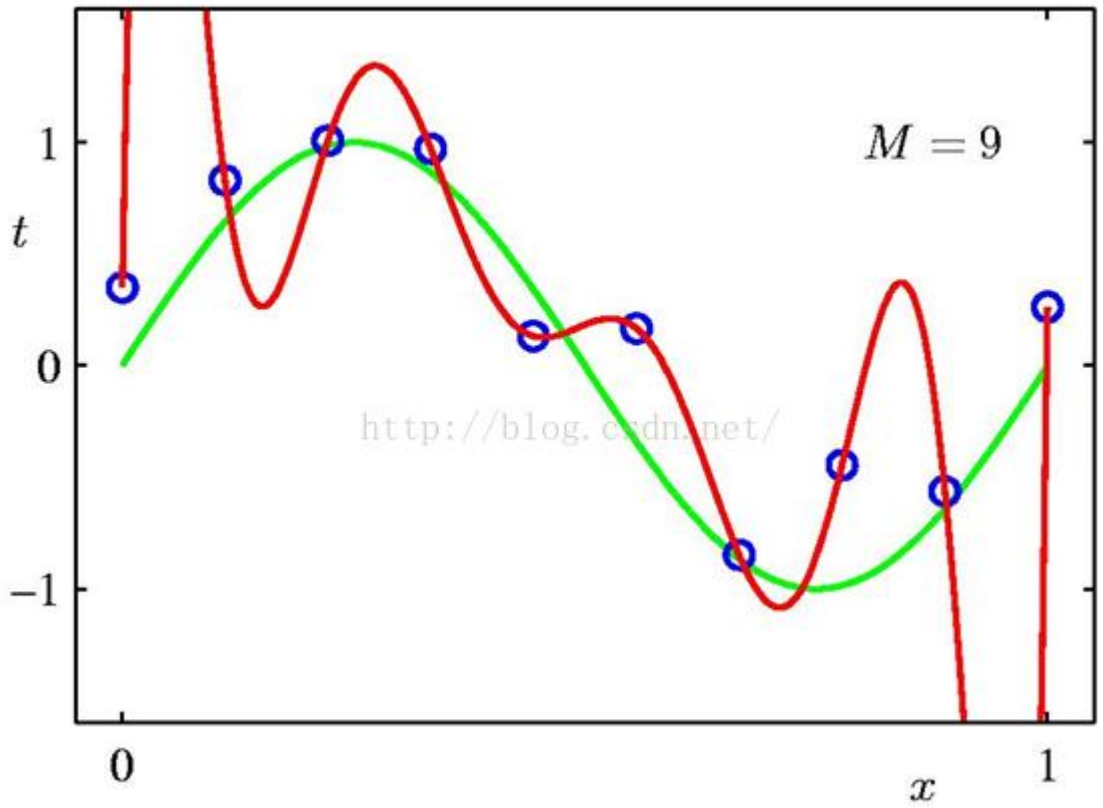
$$y' = a + b\,x'$$

幂函数曲线 $y = de^{bx}$ $y' = lny$ $x' = lnx$

$$y' = a + b\,x'$$

对数曲线 $y = a + blnx$ $y' = y$ $x' = lnx$

$$y' = a + b\,x'$$



红色的线存在明显的过拟合
绿色的线才是合理的拟合曲线