本节内容与思路

$$\begin{cases} y = x\beta + \varepsilon \\ E\varepsilon = 0, D\varepsilon = \sigma^2 \end{cases} \xRightarrow[(y_i;x_{i1},x_{i2},\cdots,x_{ip})]{n\text{个独立观测值}} ⓪\begin{cases} \hat{\beta} \\ \hat{\sigma}^2 \end{cases}$$

$(\beta,\sigma^2$未知$)$

$\varepsilon \sim N(0,\sigma^2)$

$\Rightarrow$

①检验回归方程是否显著 $(H_0: \beta_1 = \beta_2 = \cdots \beta_p = 0)$

②检验回归系数是否显著 $(H_{0j}: \beta_j = 0)$

③对β_j,σ^2进行区间估计

④预测 $\xrightarrow{x=x_0} \begin{cases} \hat{y}_0 \\ [\hat{y}_0^{(1)}, \hat{y}_0^{(2)}] \end{cases}$

其中, $x = (1, x_1, \cdots, x_p)$, $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$.

一、多元线性回归模型

下面的讨论中，自变量$x_1,\cdots,x_p$为非随机变量，因变量y为随机变量.

1. 总体模型:

$$\begin{cases} y=\beta_0+\beta_1 x_1+\cdots+\beta_p x_p+\varepsilon \\ E\varepsilon=0,\ D\varepsilon=\sigma^2<\infty\ (\sigma^2\text{未知}) \end{cases} \qquad (4.38)$$

回归变量——自变量$x_1,\cdots,x_p$;

回归系数——固定的未知参数$\beta_0,\beta_1,\cdots\beta_p$，称$\beta_j$为因子$x_j$的效应.

$E(y\,|\,x)=\beta_0+\beta_1 x_1+\cdots+\beta_p x_p$——y对$x_1,\cdots,x_p$的回归函数

$\hat{y}=\hat{\beta}_0+\hat{\beta}_1 x_1+\cdots+\hat{\beta}_p x_p$ ——y对$x_1,\cdots,x_p$的回归方程 (4.59)

回归平面——p=2时的回归方程$\hat{y}=\hat{\beta}_0+\hat{\beta}_1x_1+\hat{\beta}_2x_2$所表示的平面;

回归值——对固定的$x_1,\cdots,x_p$, y的估计值$\hat{y}=\hat{\beta}_0+\hat{\beta}_1x_1+\cdots+\hat{\beta}_px_p$.

2. 样本模型:

设有n组独立的观测值$(y_i;x_{i1},\cdots,x_{ip}),i=1,2,\cdots,n$, 则由(4.38)有

$$\begin{cases} y_i=\beta_0+\beta_1x_{i1}+\cdots+\beta_px_{ip}+\varepsilon_i, \quad i=1,\cdots,n \\ E\varepsilon_i=0, D\varepsilon_i=\sigma^2\text{且}\varepsilon_1,\cdots,\varepsilon_n\text{相互独立} \end{cases} \text{——样本模型} \quad (4.39)$$

若记 $Y=\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}_{n\times 1}, X=\begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}_{n\times(p+1)}, \beta=\begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}_{(p+1)\times 1}, \varepsilon=\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}_{n\times 1}$

其中, Y——已知的观测向量;

X——设计矩阵, 假定X列满秩, 即rank(X)=p+1;

β——未知参数向量;

ε——误差向量.

则(4.39)可写成

$$\begin{cases} Y = X\beta + \varepsilon \\ E\varepsilon = \mathbf{0}_{n\times 1},\ D\varepsilon = \sigma^2 I_n \end{cases} \quad \text{——矩阵模型} \qquad (4.40)$$

二、最小二乘估计及统计性质

1. β的最小二乘估计:

(1) 最小二乘法的原理:

误差平方和: $Q(\beta)=\sum_{i=1}^{n}\varepsilon_i^2=\sum_{i=1}^{n}(y_i-\sum_{j=0}^{p}x_{ij}\beta_j)^2$

$$\boxed{Q(\hat{\beta})=\min_{\beta}Q(\beta)}$$

(2) β的最小二乘估计 (LS估计):

$$X'X\hat{\beta}=X'Y \qquad \text{——正规方程} \qquad (4.43)$$

$$\boxed{\hat{\beta}=(X'X)^{-1}X'Y} \qquad \text{——}\beta\text{的LS估计} \qquad (4.44)$$

2. LS估计$\hat{\beta}$的性质:

(1) $\hat{\beta}$是$(y_1,y_2,\cdots,y_n)$的线性函数.

(2-3) $E(\hat{\beta})=\beta$, $D(\hat{\beta})=\sigma^2(X'X)^{-1}$.

(4) (Gauss-Markov定理) $\hat{\beta}$是β的最小方差线性无偏估计.

(7) 若$\varepsilon \sim N(0_{n\times 1}, \sigma^2 I_n)$, 则$\hat{\beta}$也是$\beta$的极大似然估计.

3. σ^2的无偏估计:

残差: $e_i = y_i - \hat{y}_i$

残差平方和: $Q_e = Q(\hat{\beta}) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

性质6 $EQ_e = (n-p-1)\sigma^2 \quad \Rightarrow \quad E\left(\dfrac{Q_e}{n-p-1}\right) = \sigma^2$.

结论1: $\hat{\sigma}_e^2 = \dfrac{Q_e}{n-p-1}$(剩余方差、残差的方差)是$\sigma^2$的无偏估计.

三、最小二乘的几何解释

四、回归方程和回归系数的显著性检验

现假定$\varepsilon \sim N_n(0,\sigma^2 I_n)$, 则矩阵模型(4.40)可写成

$$\begin{cases} Y = X\beta + \varepsilon \\ \varepsilon \sim N_n(0_{n\times 1},\sigma^2 I_n) \end{cases} \tag{4.60}$$

定理4.5

$$\left.\begin{array}{l} \hat\beta \sim N_{p+1}\left(\beta,\sigma^2(X'X)^{-1}\right) \\ \dfrac{Q_e}{\sigma^2} \sim \chi^2(n-p-1) \\ \bar y, \hat\beta_j (j=1,\cdots,p), Q_e\text{相互独立} \\ \hat\beta_0\text{与}Q_e\text{相互独立} \end{array}\right\} \xRightarrow[\text{的第 } j+1 \text{ 个对角元}]{\text{记 } c_{jj} \text{ 为} C=(X'X)^{-1}} \begin{cases} \dfrac{(\hat\beta_j-\beta_j)\big/\sqrt{c_{jj}}}{\sqrt{Q_e/(n-p-1)}} \sim t(n-p-1) \\ \dfrac{Q_e}{\sigma^2} \sim \chi^2(n-p-1) \end{cases}$$

(一) 回归方程的显著性检验

y与$x_1,\cdots,x_p$之间有无线性关系 $y=\beta_0+\beta_1x_1+\cdots+\beta_px_p+\varepsilon$, 即检验假设:

$$H_0:\beta_1=\beta_2=\cdots=\beta_p=0 \tag{4.61}$$

1. 总离差平方和的分解式:

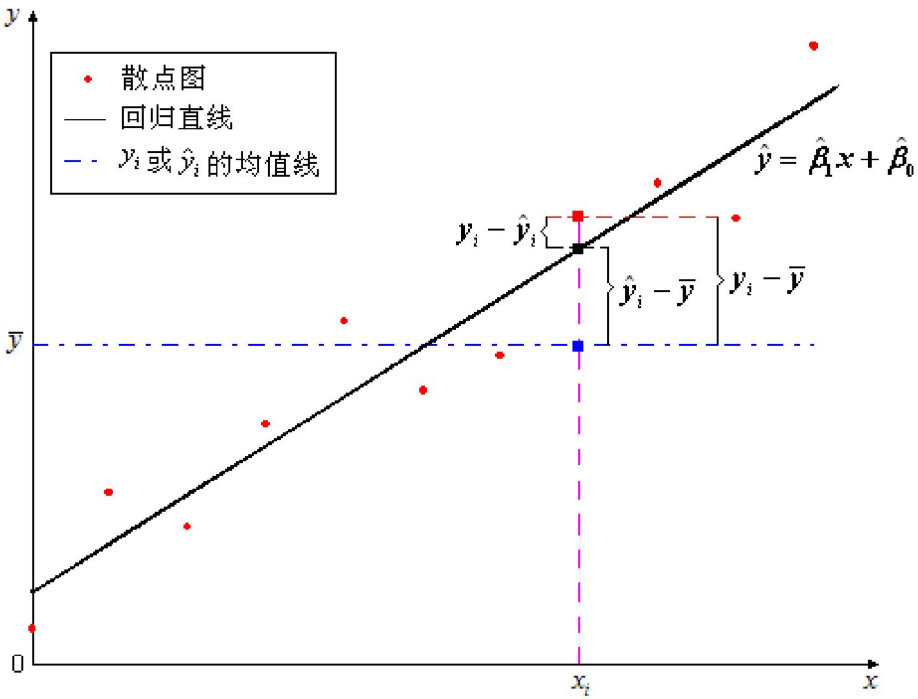


图4-7 总离差平方和分解式的几何意义

$$L_{yy}=\sum_{i=1}^{n}(y_i-\bar{y})^2$$ ——总离差平方和

$$Q_e=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$ ——残差平方和

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

则

$$\boxed{L_{yy}=Q_e+U} \tag{4.62}$$

2. 分解式中各元素的统计意义:

$$EQ_e = (n-p-1)\sigma^2,$$

$$EU = p\sigma^2 + B'[(I_n - \frac{11'}{n})\hat{X}]'[(I_n - \frac{11'}{n})\hat{X}]B$$ (此结论与 $\varepsilon \sim N(0,\sigma^2 I_n)$ 无关)

其中, $B = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}, \hat{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}, 1 = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}_{n\times 1}$.

Q_e——反映了误差引起数据 $y_1,\cdots,y_n$ 的波动程度大小;

U——除反映了误差的作用外, 还反映了回归因子 $x_1,\cdots,x_p$ 对y的总的线性影响.

3. 回归方程的显著性检验:

(1) F检验法

① 定理4.6 U与Q_e独立, 且$U/\sigma^2 \overset{H_0:\beta_1=\cdots=\beta_p=0\text{成立}}{\sim\!\sim\!\sim\!\sim\!\sim\!\sim\!\sim\!\sim\!\sim\!\sim\!\sim\!\sim} \chi^2(p)$.

② 检验统计量: $\boxed{F=\dfrac{U/p}{Q_e/(n-p-1)} \overset{H_0:\beta_1=\cdots=\beta_p=0\text{成立}}{\sim\!\sim\!\sim\!\sim\!\sim\!\sim\!\sim\!\sim\!\sim\!\sim\!\sim\!\sim} F(p,n-p-1)}$.

③ $H_0:\beta_1=\beta_2=\cdots=\beta_p=0$的拒绝域: $\boxed{F>F_{1-\alpha}(p,n-p-1)}$.

表4-5 回归分析的方差分析表

方差来源	平方和	自由度	均方	F值
回归	U	p	U/p	$\dfrac{U/p}{Q_e/(n-p-1)}$
残差	Q_e	n-p-1	$Q_e/(n-p-1)$	
总和	L_{yy}	n-1		

(2) R检验法

$$R=\sqrt{\frac{U}{L_{yy}}}=\sqrt{\frac{U}{Q_e+U}}$$

——y与 $x_1,\cdots,x_p$ 的多元相关系数 (或复相关系数) (4.65)

易证明: $F=\frac{n-p-1}{p}\frac{R^2}{1-R^2}$, 故用 F 和用 R 检验$H_0$是等效的.

(二) 回归系数的显著性检验

若方程是显著的,则还需对各x_j的显著性进行检验,即检验假设:

$$H_{0j}:\beta_j=0,\quad (j=1,\cdots,p)$$

若拒绝H_{0j},就认为x_j对y的作用显著;否则应剔除变量x_j.

1. t检验法:

(1) 检验统计量: $\boxed{T_j=\dfrac{\hat{\beta}_j\big/\sqrt{c_{jj}}}{\sqrt{Q_e\big/(n-p-1)}}\overset{H_{0j}\text{成立}}{\sim\!\sim\!\sim\!\sim}t(n-p-1)}$.

(2) $H_{0j}:\beta_j=0,(j=1,\cdots,p)$的拒绝域: $\boxed{|T_j|>t_{1-\alpha/2}(n-p-1)}$.

若存在不显著变量,取$|T_k|=\min\limits_{1\le j\le p}\{|T_j|\}$,从方程中剔除最不显著变量$x_k$.

2. 建立只包含显著变量的回归方程的步骤(只出不进法):

(1) 由样本值$(y_i; x_{i1}, \cdots, x_{ip}), i = 1, 2, \cdots, n$, 建立p元回归方程

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_{k-1} x_{k-1} + \hat{\beta}_k x_k + \hat{\beta}_{k+1} x_{k+1} + \cdots + \hat{\beta}_p x_p \qquad (4.59)$$

(2) 对p元回归方程作显著性检验, 若拒绝H_0, 就认为此方程有意义;

(3) 对回归系数β_j作显著性检验, 剔除掉最不显著的变量x_k;

(4) 对余下的p-1个变量重新建立回归方程, 重复步骤(1)(3)(4), 直到回归方程中所有的变量都显著为止.

【★例4.5(P_{178})】在平炉炼钢中，由于矿石与炉气的氧化作用，铁水的总含碳量在不断降低，一炉钢在冶炼初期总的去碳量y与所加的两种矿石的量x_1, x_2及熔化时间x_3有关. 经实测某号平炉的49组数据如下表4-6所列：

编号	x_1	x_2	x_3	y	编号	x_1	x_2	x_3	y
1	2	18	50	4.3302	26	9	6	39	2.7066
2	7	9	40	3.6485	27	12	5	51	5.6314
3	5	14	46	4.4830	28	6	13	41	5.8152
4	12	3	43	5.5468	29	12	7	47	5.1302
5	1	20	64	5.4970	30	0	24	61	5.3910
6	3	12	40	3.1125	31	5	12	37	4.4533
7	3	17	64	5.1182	32	4	15	49	4.6569
8	6	5	39	3.8759	33	0	20	45	4.5212
9	7	8	37	4.6700	34	6	16	42	4.8650
10	0	23	55	4.9536	35	4	17	48	5.3566
11	3	16	60	5.0060	36	10	4	48	4.6098
12	0	18	49	5.2701	37	4	14	36	2.3815
13	8	4	50	5.3772	38	5	13	36	3.8746
14	6	14	51	5.4849	39	9	8	51	4.5919
15	0	21	51	4.5960	40	6	13	54	5.1588
16	3	14	51	5.6645	41	5	8	100	5.4373

17	7	12	56	6.0795	42	5	11	44	3.9960
18	16	0	48	3.2194	43	8	6	63	4.3970
19	6	16	45	5.8076	44	2	13	55	4.0622
20	0	15	52	4.7306	45	7	8	50	2.2905
21	9	0	40	4.6805	46	4	10	45	4.7115
22	4	6	32	3.1272	47	10	5	40	4.5310
23	0	17	47	2.6104	48	3	17	64	5.3637
24	9	0	44	3.7174	49	4	15	72	6.0771
25	2	16	39	3.8946					

设y与x_1,x_2,x_3之间有线性关系

$$\begin{cases} y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\varepsilon, \\ \varepsilon\sim N(0,\sigma^2) \end{cases}$$

求($\alpha=0.01$):

(1)y与x_1,x_2,x_3的回归方程;

(2)检验回归方程和回归系数的显著性;

(3)如有不显著的变量，请剔除之并求剔除不显著的变量之后的回归方程.

五、回归系数的置信区间($\hat{\sigma}_e = \sqrt{Q_e/(n-p-1)}$)

1. $\beta_j\ (j=0,1,\cdots,p)$的置信水平为$1-\alpha$的置信区间:

$$\boxed{\frac{(\hat{\beta}_j-\beta_j)\big/\sqrt{c_{jj}}}{\sqrt{Q_e/(n-p-1)}} \sim t(n-p-1)}$$

$$\Rightarrow [\hat{\beta}_j \pm \hat{\sigma}_e \cdot \sqrt{c_{jj}} \cdot t_{1-\alpha/2}(n-p-1)]$$

2. (补)σ^2的置信水平为$1-\alpha$的置信区间:

$$\boxed{\frac{Q_e}{\sigma^2} \sim \chi^2(n-p-1)}$$

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

【★例4.6(P182)】在例4.5中, 求β_1,β_2,β_3的置信水平为95%的置信区间.

六、利用多元回归方程进行预测

设y与$x_1,\cdots,x_p$满足模型$\begin{cases} y=\beta_0+\beta_1x_1+\cdots+\beta_px_p+\varepsilon \\ \varepsilon\sim N(0,\sigma^2) \end{cases}$(4.60). 并已得回归系数均显著的回归方程为$\hat{y}=\hat{\beta}_0+\hat{\beta}_1x_1+\cdots+\hat{\beta}_px_p$.

令$x_0=(x_{01},x_{02},\cdots,x_{0p})$为$(x_1,x_2,\cdots,x_p)$的一组固定值, 设$y_0,y_1,\cdots,y_n$相互独立. 求:

(1) y_0的预测值:

$\hat{y}_0=\hat{\beta}_0+\hat{\beta}_1x_{01}+\cdots+\hat{\beta}_px_{0p}$, 且$\hat{y}_0$是$Ey_0$的无偏估计.

(2) y_0的置信水平为$1-\alpha$的预测区间:

记$\tilde{C}_{p\times p}=C(2:p+1,2:p+1)$, 其中$C=(X'X)^{-1}$,

$\Delta x_0 = x_0 - \bar{x}$, 其中 $\bar{x} = (\bar{x}_1, \cdots, \bar{x}_p)$, 则

$$\boxed{T = \frac{y_0 - \hat{y}_0}{\hat{\sigma}_e \sqrt{1 + \frac{1}{n} + (\Delta x_0)_{1\times p} \tilde{C}_{p\times p} (\Delta x_0)'_{p\times 1}}} \sim t(n-p-1)}$$

$$\Rightarrow [\hat{y}_0 \pm \delta(x_0)] \tag{4.77}$$

其中, $\delta(x_0) = \hat{\sigma}_e \cdot t_{1-\alpha/2}(n-p-1) \cdot \sqrt{1 + \frac{1}{n} + (\Delta x_0)_{1\times p} \tilde{C}_{p\times p} (\Delta x_0)'_{p\times 1}}$.

特别地, 当n较大而 x_{0j} 接近于 $\bar{x}_j$ 时, y_0 的置信水平为 $1-\alpha$ 的预测区间近似为

$$[\hat{y}_0 - \hat{\sigma}_e \cdot t_{1-\alpha/2}(n-p-1), \hat{y}_0 + \hat{\sigma}_e \cdot t_{1-\alpha/2}(n-p-1)] \tag{4.79}$$

【★例4.7(P_{185})】在例4.5中, 若取 $(x_{01}, x_{02}, x_{03}) = (5,10,50)$, 求 y_0 的置信水平为95%的预测区间.

七、最优回归方程的选择

1. 最优的准则:

最优是指, 一方面方程包含所有对y有显著影响的自变量, 另一方面方程所含自变量的个数尽可能少. 当多个方程都满足上述两个要求时, 以σ^2的无偏估计$\hat{\sigma}^2 = \dfrac{Q_e}{n - p_0 - 1} \triangleq \hat{\sigma}_e^2$最小者为优.

其中Q_e为残差平方和, p_0为当前方程中自变量的个数, n为样本容量.

2. 选择最优方程的方法:

(1) “全部比较”法

Step1: 若所有自变量共有p个, 则分别建立包含一个自变量、二个自变量、…、p个自变量的线性回归方程, 共有2^p-1个;

Step2: 对各方程以及各方程中的系数进行显著性检验, 选出方程与方程中每个系数均显著的方程作为“最优”的备选.

Step3: 对各备选方程, 利用

$$\hat{\sigma}_e^2=\frac{Q_e}{n-p_0-1}$$

计算出$\hat{\sigma}_e^2$, $\hat{\sigma}_e^2$值最小的备选方程即为“最优”线性回归方程.

优缺点: 此方法可找到最优回归方程; 但p稍大时计算量就非常大.

(2) “只出不进”法

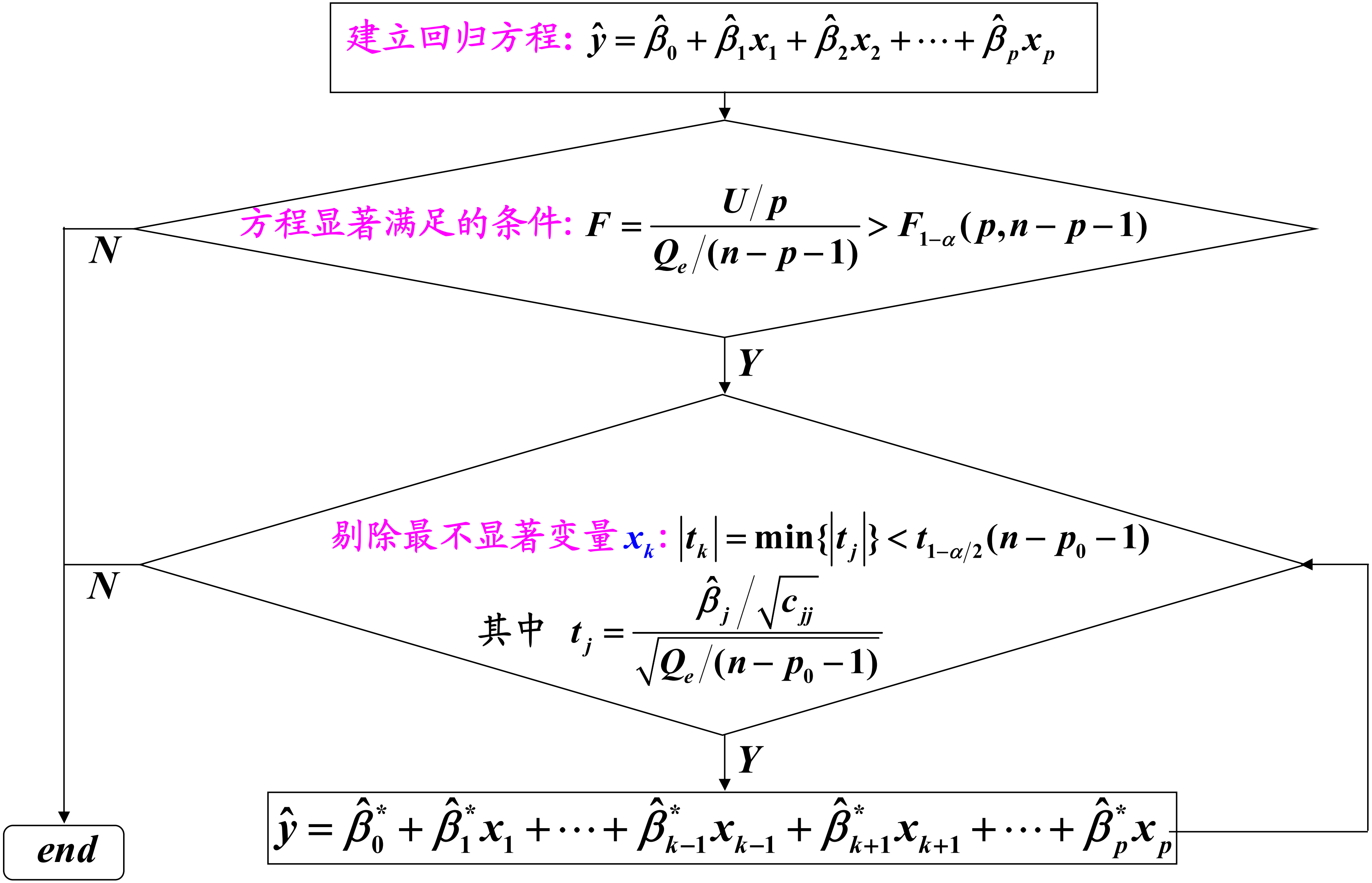
Step1: 若共有p个自变量$x_1, x_2, \cdots, x_p$, 则建立p元线性回归方程, 并对方程作显著性检验, 若方程显著, 则进入第二步;

Step2: 对每个自变量作显著性检验, 剔除最不显著的一个变量x_k;

Step3: 对余下的p-1个变量重新建立回归方程, 重复Step2, 直到回归方程中所有的变量都显著为止.

优缺点: 此方法在自变量不多、特别是不显著变量不多时使用, 其工作量不是太大. 但此法有一个致命的缺点, 就是自变量一旦被剔除就再也不能进入回归方程中, 因此, 此法所得回归方程不一定是真正的最优回归方程.

线性回归分析中“只出不进”法的算法流程图



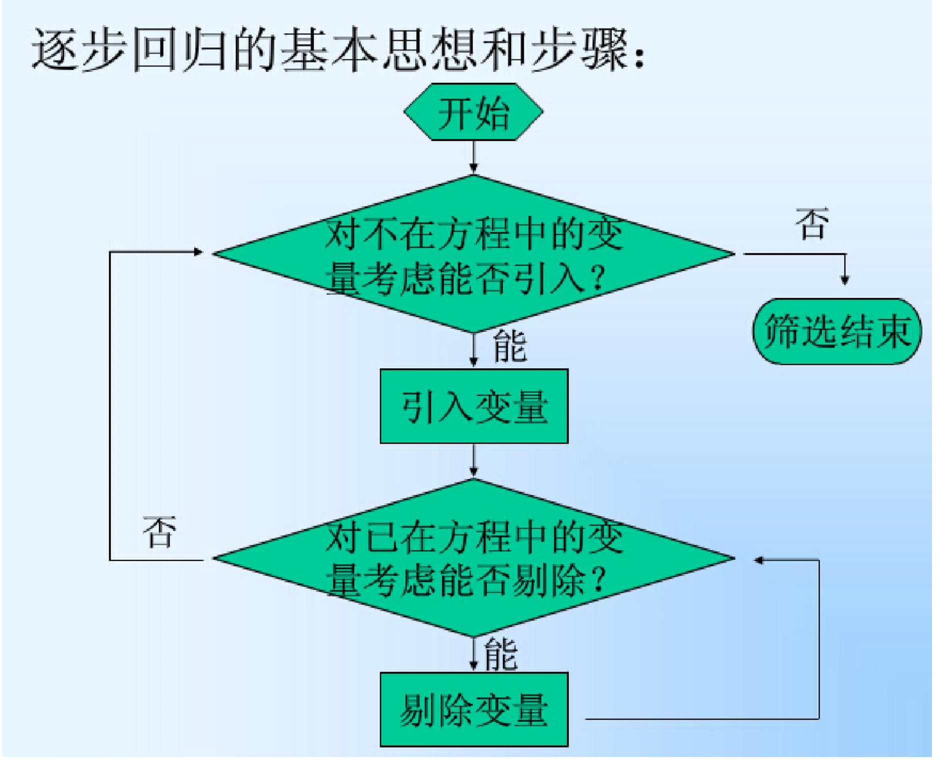
(3) 逐步回归法

Step1: 求y对每个$x_j(j=1,\cdots,p)$的一元线性回归方程, 并对x_j进行显著性检验, 从所有显著变量($|T_j|>t_{1-\alpha/2}(n-2)$)中选出最显著自变量$x_k$, 进入第二步;

Step2: 求y对已引入变量x_k和各未引入的$x_j(j\neq k)$的二元线性回归方程:

①对方程进行显著性检验;

②在显著性方程中对x_j进行显著性检验, 计算$|T_j|$, 选出最显著自变量x_t所在的回归方程, 再利用"只出不进"法剔除掉所有不显著自变量后进入第三步;



Step3: 对未引入的自变量逐一引入, 重复**Step2**中的②, 直到既无法剔除已入选的自变量, 也无法再入选新的自变量为止.

优缺点: 此方法也不能保证最后所得的方程是真正的最优回归方程, 但实际应用中, 用此法得到的回归方程进行预测效果还是比较好的, 加之计算量不太大, 又有较成熟的计算程序可供使用, 故此法是目前用得最多的一种方法.

注意: 使用逐步回归法要恰当地选取显著性水平α, α选取得较大, 将会选入较多的自变量; α选取得较小, 将会导致一些重要的自变量被剔除.