第五章 自变量的选择与逐步回归

5.1 自变量选择对估计和预测的影响

5.2 所有子集回归

5.3 逐步回归

5.4 本章小结与评注

§第5章 自变量选择与逐步回归

从20世纪60年代开始，关于回归自变量的选择成为统计学中研究的热点问题。统计学家们提出了许多回归选元的准则，并提出了许多行之有效的选元方法。

本章从回归选元对回归参数估计和预测的影响开始，介绍自变量选择常用的几个准则；扼要介绍所有子集回归选元的几个方法；详细讨论逐步回归方法及其应用。

§5.1 自变量选择对估计和预测的影响

一、全模型和选模型

设研究某一实际问题涉及到对因变量有影响的因素共有m个，回归模型为：

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\ldots+\beta_mx_m+\varepsilon \qquad (5.1)$$

称为全回归模型。

如果我们从所有可供选择的m个变量中挑选出p个，记为$x_1, x_2, \ldots, x_p$, 构成的回归模型为：

$$y=\beta_{0p}+\beta_{1p}x_1+\beta_{2p}x_2+\ldots+\beta_{pp}x_p+\varepsilon_p \qquad (5.2)$$

称模型（5.2）式为选模型。

§5.1 自变量选择对估计和预测的影响

一、全模型和选模型

模型选择不当会给参数估计和预测带来什么影响?下面我们将分别给予讨论。

为了方便，我们把模型（5.1）式的参数估计向量$\hat{\boldsymbol{\beta}}$和σ^2的估计记为：

$$\hat{\boldsymbol{\beta}}_m = (\mathbf{X}_m' \mathbf{X}_m)^{-1} \mathbf{X}_m' \mathbf{y} \qquad \hat{\sigma}_m^2 = \frac{1}{n-m-1} SSE_m$$

把模型（5.2）式的参数估计向量记为

$$\hat{\boldsymbol{\beta}}_p = (\mathbf{X}_p' \mathbf{X}_p)^{-1} \mathbf{X}_p' \mathbf{y} \qquad \hat{\sigma}_p^2 = \frac{1}{n-p-1} SSE_p$$

§5.1 自变量选择对估计和预测的影响

二、自变量选择对预测的影响

关于自变量选择对预测的影响可以分成两种情况：

第一种情况是全模型正确而误用了选模型；

第二种情况是选模型正确而误用了全模型式。

§5.1 自变量选择对估计和预测的影响

（一）全模型正确而误用选模型的情况

性质 1. 在 x_j 与 $x_{p+1}, \cdots, x_m$ 的相关系数不全为 0 时，选模型回归系数的最小二乘估计是全模型相应参数的有偏估计,即

$$E(\hat{\beta}_{jp}) = \beta_{jp} \neq \beta_j \quad (j=1,2, \quad \cdots,p)。$$

§5.1 自变量选择对估计和预测的影响

（一）全模型正确而误用选模型的情况

性质 2. 选模型的的预测是有偏的。

给定新自变量值 $\mathbf{x}_{0p} = (x_{01}, x_{02}, \cdots, x_{0m})'$，因变量新值为

$$y_0 = \beta_0 + \beta_1 x_{01} + \beta_2 x_{02} + \cdots + \beta_m x_{0m} + \varepsilon_0$$

用选模型的预测值为

$$\hat{y}_{0p} = \hat{\beta}_{0p} + \hat{\beta}_{1p} x_{01} + \hat{\beta}_{2p} x_{02} \cdots + \hat{\beta}_{pp} x_{0p}$$

作为 y_0 的预测值是有偏的，即 $E(\hat{y}_{0p} - y_0) \neq 0$ 。

§5.1 自变量选择对估计和预测的影响

（一）全模型正确而误用选模型的情况

性质 3. 选模型的参数估计有较小的方差

选模型的最小二乘参数估计为 $\hat{\boldsymbol{\beta}}_p = (\hat{\beta}_{0p}, \hat{\beta}_{1p}, \cdots, \hat{\beta}_{pp})'$

全模型的最小二乘参数估计为 $\hat{\boldsymbol{\beta}}_m = (\hat{\beta}_{0m}, \hat{\beta}_{1m}, \cdots, \hat{\beta}_{mm})'$

这条性质说明 $D(\hat{\beta}_{jp}) \le D(\hat{\beta}_{jm}), j = 0,1,\cdots,p$ 。

§5.1 自变量选择对估计和预测的影响

（一）全模型正确而误用选模型的情况

性质 4. 选模型的预测残差有较小的方差。

选模型的预测残差为 $e_{0p} = \hat{y}_{0p} - y_0$

全模型的预测残差为 $e_{0m} = \hat{y}_{0m} - y_0$

其中 $y_0 = \beta_0 + \beta_1 x_{01} + \beta_2 x_{02} + \cdots + \beta_m x_{0m} + \varepsilon$

则有 $D(e_{0p}) \leqslant D(e_{0m})$。

§5.1 自变量选择对估计和预测的影响

（一）全模型正确而误用选模型的情况

性质 5. 记 $\boldsymbol{\beta}_{m-p}=(\beta_{p+1},\cdots,\beta_m)'$

用全模型对 $\boldsymbol{\beta}_{m-p}$ 的最小二乘估计为 $\hat{\boldsymbol{\beta}}_{m-p}=(\hat{\beta}_{p+1},\cdots,\hat{\beta}_m)'$

则在 $D(\hat{\boldsymbol{\beta}}_{m-p})\geq\boldsymbol{\beta}_{m-p}\boldsymbol{\beta}'_{m-p}$ 的条件下

$E(e_{0p})^2=D(e_{0p})+(E(e_{0p}))^2\leqslant D(e_{0m})$

即选模型预测的均方误差比全模型预测的方差更小。

§5.1 自变量选择对估计和预测的影响

（二）选模型正确而误用全模型的情况

如果选模型正确，从无偏性的角度看，

选模型的预测值 $\hat{y}_{0p} = \hat{\beta}_{0p} + \hat{\beta}_{1p}x_{01} + \hat{\beta}_{2p}x_{02} \cdots + \hat{\beta}_{pp}x_{0p}$

是因变量新值 $y_0 = \beta_0 + \beta_1 x_{01} + \beta_2 x_{02} + \cdots + \beta_p x_{0p} + \varepsilon_0$

的无偏估计，此时全模型的预测值

$\hat{y}_{0m} = \hat{\beta}_0 + \hat{\beta}_1 x_{01} + \hat{\beta}_2 x_{02} \cdots + \hat{\beta}_m x_{0m}$ 是 y_0 的有偏估计。

§5.1 自变量选择对估计和预测的影响

（二）选模型正确而误用全模型的情况

从预测方差的角度看，根据性质 4，选模型的预测方差 $D(\hat{y}_{0p})$ 小于全模型的预测方差 $D(\hat{y}_{0m})$

从均方预测误差的角度看，全模型的均方预测误差

$$E(\hat{y}_{0m}-y_0)^2=D(\hat{y}_{0m})+[E(\hat{y}_{0m})-E(y_0)]^2$$

包含预测方差与预测偏差的平方两部分

而选模型的均方预测误差 $E(\hat{y}_{0p}-y_0)^2=D(\hat{y}_{0p})$

仅包含预测方差这一项，并且 $D(\hat{y}_{0p})\leqslant D(\hat{y}_{0m})$

因而从均方预测误差的角度看，全模型的预测误差将更大。

§5.1 自变量选择对估计和预测的影响

（二）选模型正确而误用全模型的情况

上述结论告诉我们，一个好的回归模型，并不是考虑的自变量越多越好。在建立回归模型时，选择自变量的基本指导思想是“少而精”。哪怕我们丢掉了一些对因变量y还有些影响的自变量，由选模型估计的保留变量的回归系数的方差，要比由全模型所估计的相应变量的回归系数的方差小。而且，对于所预测的因变量的方差来说也是如此。丢掉了一些对因变量y有影响的自变量后，所付出的代价是估计量产生了有偏性。然而，尽管估计量是有偏的，但预测偏差的方差会下降。另外，如果保留下来的自变量有些对因变量无关紧要，那么，方程中包括这些变量会导致参数估计和预测的有偏性和精度降低。

§5.2 所有子集回归

一、所有子集的数目

有m个可供选择的变量$x_1,x_2,\ldots,x_m$,由于每个自变量都有入选和不入选两种情况，这样y关于这些自变量的所有可能的回归方程就有2^m-1个。

从另一个角度看

$$C_m^0+C_m^1+\cdots+C_m^m=2^m$$

§5.2 所有子集回归

二、关于自变量选择的几个准则

从数据与模型拟合优劣的直观考虑出发，认为残差平方和SSE最小的回归方程就是最好的。还曾用复相关系数R来衡量回归拟合的好坏。然而这两种方法都有明显的不足，这是因为：

$$SSE_{p+1} \leqslant SSE_p$$

$$R^2_{p+1} \geqslant R^2_p$$

§5.2 所有子集回归

准则1 自由度调整复相关系数达到最大

$$R_a^2 = 1 - \frac{n-1}{n-p-1}(1-R^2)$$

显然有 $R_a^2 \leqslant R^2$， R_a^2 随着自变量的增加并不一定增大。

从拟合优度的角度追求“最优”，则所有回归子集中 R_a^2 最大者对应的回归方程就是“最优”方程。

§5.2 所有子集回归

准则1 自由度调整复相关系数达到最大

从另外一个角度考虑回归的拟合效果，回归误差项方差σ²的无偏估计为：

$$\hat{\sigma}^2 = \frac{1}{n-p-1} SSE$$

此无偏估计式中也加入了惩罚因子n-p-1

§5.2 所有子集回归

准则1 自由度调整复相关系数达到最大

由以上分析，用平均残差平方和$\hat{\sigma}^2$作为自变量选元准则是合理的，那末它和调整的复判定系数R_a^2准则有什么关系哪？实际上，这两个准则是等价的，容易证明以下关系式成立

$$R_a^2 = 1 - \frac{n-1}{SST}\hat{\sigma}^2$$

由于SST是与回归无关的固定值，因而R_a^2与$\hat{\sigma}^2$是等价的

§5.2　所有子集回归

准则2　赤池信息量AIC达到最小

AIC准则是日本统计学家赤池(Akaike)1974年根据极大似然估计原理提出的一种较为一般的模型选择准则，人们称它为Akaike信息量准则 (Akaike Information Criterion, 简记为AIC)。AIC准则既可用来作回归方程自变量的选择，又可用于时间序列分析中自回归模型的定阶上。由于该方法的广泛应用，使得赤池乃至日本统计学家在世界的声誉大增。

§5.2　所有子集回归

准则2　赤池信息量AIC达到最小

设回归模型的似然函数为$L(\boldsymbol{\theta},\mathbf{x})$, θ的维数为p，$\mathbf{x}$为样本，在回归分析中样本为$\mathbf{y}=(y_1, y_2, \ldots y_n)'$，则AIC定义为：

$$AIC=-2\ln L(\hat{\boldsymbol{\theta}}_L,\mathbf{x})+2p$$

其中$\hat{\boldsymbol{\theta}}_L$是 θ 的极大似然估计，p 是未知参数的个数。

§5.2 所有子集回归

准则2 赤池信息量AIC达到最小

假定回归模型的随机误差项ε遵从正态分布，即

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

对数似然函数为

$$\ln L_{\max} = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\hat{\sigma}_L^2) - \frac{1}{2\hat{\sigma}_L^2} SSE$$

将 $\hat{\sigma}_L^2 = \frac{1}{n}SSE$ 代入得

$$\ln L_{\max} = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\frac{SSE}{n}) - \frac{n}{2}$$

§5.2　所有子集回归

准则2　赤池信息量AIC达到最小

带入公式 AIC=-2lnL($\hat{\theta}_L$,**x**)+2p 中

这里似然函数中的未知参数个数为p+2，略去与p无关的常数，得回归模型的AIC公式为

$$AIC=n\ln(SSE)+2p$$

对每一个回归子集计算AIC，其中AIC最小者所对应的模型是“最优”回归模型

§5.2 所有子集回归

准则4 C_p统计量达到最小

1964年马勒斯（Mallows)从预测的角度提出一个可以用来选择自变量的统计量———C_p统计量。根据性质5，即使全模型正确，但仍有可能选模型有更小的预测误差。C_p正是根据这一原理提出来的。

§5.2 所有子集回归

准则4 C_p统计量达到最小

考虑在n个样本点上，用选模型（5.2）式作回报预测时，预测值与期望值的相对偏差平方和为：

$$J_p = \frac{1}{\sigma^2}\sum_{i=1}^{n}(\hat{y}_{ip} - E(y_i))^2$$

$$= \frac{1}{\sigma^2}\sum_{i=1}^{n}(\hat{\beta}_{0p} + \hat{\beta}_{1p}x_{i1} + \cdots + \hat{\beta}_{pp}x_{ip} - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_m x_{im}))^2$$

§5.2 所有子集回归

准则4 C_p统计量达到最小

可以证明，J_p的期望值是

$$E(J_p)=\frac{E(SSE_p)}{\sigma^2}-n+2(p+1)$$

略去无关的常数2，据此构造出C_p统计量为

$$C_p=\frac{SSE_p}{\hat{\sigma}^2}-n+2p=(n-m-1)\frac{SSE_p}{SSE_m}-n+2p$$

§5.2 所有子集回归

准则4 C_p统计量达到最小

其中$\hat{\sigma}^2 = \dfrac{1}{n-m-1} SSE_m$是全模型中$\sigma^2$的无偏估计。

这样我们得到一个选择变量的C_p准则：

选择使C_p最小的自变量子集，这个自变量子集对应的回归方程就是“最优”回归 方程。

§5.2 所有子集回归

例5.1 y表示某种消费品的销售额，

x_1表示居民可支配收入，

x_2表示该类消费品的价格指数，

x_3表示其他消费品平均价格指数。

表5.1给出了某地区18年某种消费品销售情况资料，试建立该地区该消费品销售额预测方程。

§5.2 所有子集回归

表5.1

序号	x_1（元）	x_2(%)	x_3(%)	（百万元）
1	81.2	85.0	87.0	7.8
2	82.9	92.0	94.0	8.4
3	83.2	91.5	95.0	8.7
4	85.9	92.9	95.5	9.0
5	88.0	93.0	96.0	9.6
6	99.9	96.0	97.0	10.3
7	102.0	95.0	97.5	10.6
8	105.3	95.6	97.0	10.9
9	117.7	98.9	98.0	11.3
10	126.4	101.5	101.2	12.3
11	131.2	102.0	102.5	13.5
12	148.0	105.0	104.0	14.2
13	153.0	106.0	105.9	14.9
14	161.0	109.0	109.5	15.9
15	170.0	112.0	111.0	18.5
16	174.0	112.5	112.0	19.5
17	185.0	113.0	112.3	19.9
18	189.0	114.0	113.0	20.5

§5.2 所有子集回归

这个例子中，n=18,m=3，所有的自变量子集有$2^m-1=7$个，即有7个回归子集。

表5.2

自变量子集	R^2	R_a^2	AIC	Cp
x_1	0.9728	0.9711	40.06	4.134
x_2	0.9566	0.9539	48.48	16.151
x_3	0.9508	0.9477	50.74	20.452
x_1,x_2	0.9747	0.9714	40.76	4.734
x_1,x_3	0.9784	0.9755	37.93	2.005
x_2,x_3	0.9576	0.9519	50.09	17.461
x_1,x_2,x_3	0.9811	0.9771	37.52	2.000

§5.2　所有子集回归

由表5.2的3项指标均可看到x_1，x_2，x_3是“最优”子集，x_1，x_3是“次优”子集。回归方程分别为

$$\hat{y} = -10.1489 + 0.1008x_1 - 0.3104x_2 + 0.4110x_3$$

$$\hat{y} = -14.049 + 0.07641x_1 + 0.1178x_3$$

§5.2　所有子集回归

三、用SAS软件寻找最优子集

例5.2　对例 3.1 的数据，用调整的复判定系数R_a^2准则选择最优子集回归模型。

SAS软件共有三个基本窗口，分别为：

（1）程序编辑窗（PROGRAM EDITOR），用来编辑程序。

（2）日志窗（LOG），显示已执行的语句和系统信息，包括错误信息。

（3）输出窗（OUTPUT）显示程序运行结果。

用主菜单的Window命令可以实现在三个窗口间的转换。

§5.2 所有子集回归

```
data data1;
input x1-x12 y;
cards;
1.94  4.5  154.45  207.33 246.87  277.64  135.79  30.58  110.67  80.83
51.83  14.09  2384
0.33  6.49  133.16  127.29  120.17  114.88  81.21  14.05  35.7  16  27.1
2.93   202
…
;
proc reg;
model y=x1-x12/selection=adjrsq;
run;
```

§5.2 所有子集回归

以下是部分输出结果：

Adjusted R-square	R-square	In	Variables in Model
0.82985517	0.86388414	6	X3 X5 X8 X9 X10 X11
0.82692850	0.86731185	7	X3 X5 X6 X8 X9 X10 X11
0.82487399	0.85989919	6	X3 X6 X8 X9 X10 X11
0.82366778	0.86481197	7	X3 X4 X5 X8 X9 X10 X11
0.82343275	0.86463178	7	X3 X5 X8 X9 X10 X11 X12
0.82311828	0.86439068	7	X3 X5 X7 X8 X9 X10 X11

…

§5.3 逐步回归

一、问题的提出及逐步回归的思想

自变量的所有可能子集构成2^m-1个回归方程，当可供选择的自变量不太多时，用前边的方法可以求出一切可能的回归方程，然后用几个选元准则去挑出“最好”的方程，但是当自变量的个数较多时，要求出所有可能的回归方程是非常困难的。为此，人们提出了一些较为简便、实用、快速的选择“最优”方程的方法。人们所给出的方法各有优缺点，至今还没有绝对最优的方法，目前常用的方法有“前进法”、“后退法”、“逐步回归法”，而逐步回归法最受推崇。

§5.3 逐步回归

一、问题的提出及逐步回归的思想

在后边的讨论中，无论我们从回归方程中剔除某个自变量，还是给回归方程增加某个自变量都要利用（3.42）式的偏F检验，这个偏F检验与（3.40）式的t检验是等价的，F检验的定义式的统计意义更为明了，并且容易推广到对多个自变量的显著性检验，因而采用F检验。

$$F_j = \frac{\Delta SSR_{(j)}/1}{SSE/(n-p-1)} \qquad t_j = \frac{\hat{\beta}_j}{\sqrt{c_{jj}}\,\hat{\sigma}}$$

§5.3 逐步回归

一、前进法

前进法的思想是变量由少到多，每次增加一个，直至没有可引入的变量为止。首先分别对因变量 y 建立 m 个一元线性回归方程，并分别计算这 m 个一元回归方程的 m 个回归系数的 F 检验值，记为$\{F_1^1, F_2^1, \cdots, F_m^1\}$，选其最大者记为：

$$F_j^1 = \max\{F_1^1, F_2^1, \cdots, F_m^1\}$$

给定显著性水平 α，若$F_j^1 \geqslant F_\alpha(1, n-2)$，则首先将 x_j 引入回归方程，为方便，设 x_j 就是 x_1。

§5.3 逐步回归

一、问题的提出及逐步回归的思想

接下来因变量 y 分别与 (x_1, x_2)，(x_1, x_3)，…，(x_1, x_m)建立 m-1 个二元线性回归方程，对这 m-1 个回归方程中 $x_2, x_3, \cdots, x_m$ 的回归系数进行 F 检验，计算 F 值，记为$\{F_2^2, F_3^2, \cdots, F_m^2\}$，选其最大的记为：

$$F_j^2 = \max\{F_2^2, F_3^2, \cdots, F_m^2\}$$

若 $F_j^2 \geqslant F_\alpha(1, n-3)$，则接着将 x_j 引入回归方程。

§5.3 逐步回归

一、问题的提出及逐步回归的思想

依上述方法接着做下去。直至所有未被引入方程的自变量的F值均小于$F_{\alpha}(1, n-p-1)$时为止。这时，得到的回归方程就是最终确定的方程。

每步检验中的临界值$F_{\alpha}(1,n-p-1)$与自变量数目p有关，在用软件计算时，我们实际使用的是显著性P值（或记为sig）做检验。

§5.3 逐步回归

一、问题的提出及逐步回归的思想

例5.4 对例3.1国际旅游外汇收入y对第三产业的12个变量做回归的数据，用前进法做变量选择，取显著性水平$\alpha_{进}$=0.05。

首先进入线性回归对话框，将y与x1至x12分别选入各自的变量框，然后在Method对话框中点选前进法Forward,点选Options选项看到默认的显著性水平$\alpha_{进}$正是0.05。部分运行结果如下：

§5.3 逐步回归

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-209.535	124.469		-1.683	.103
	X7	6.907	1.163	.741	5.938	.000
2	(Constant)	-96.142	108.300		-.888	.382
	X7	13.791	2.101	1.479	6.564	.000
	X4	-2.520	.682	-.832	-3.695	.001
3	(Constant)	-174.886	108.984		-1.605	.120
	X7	11.152	2.351	1.196	4.744	.000
	X4	-2.034	.685	-.672	-2.970	.006
	X10	10.761	5.139	.260	2.094	.046
4	(Constant)	-228.815	104.015		-2.200	.037
	X7	8.786	2.417	.942	3.635	.001
	X4	-3.261	.832	-1.077	-3.919	.001
	X10	13.864	4.965	.335	2.792	.010
	X3	2.849	1.244	.647	2.290	.030
5	(Constant)	-140.625	102.304		-1.375	.181
	X7	3.910	3.003	.419	1.302	.205
	X4	-1.997	.927	-.660	-2.154	.041
	X10	18.431	4.939	.446	3.732	.001
	X3	5.090	1.473	1.157	3.455	.002
	X11	-7.442	3.086	-.551	-2.411	.024

§5.3 逐步回归

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.741	.549	.533	455.9279
2	.835	.697	.675	380.4405
3	.860	.739	.710	359.3347
4	.885	.783	.749	334.0439
5	.908	.824	.789	306.8386

§5.3 逐步回归

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	7329802.2	1	7329802.2	35.261	.000
	Residual	6028236.5	29	207870.22		
	Total	13358039	30			
2	Regression	9305460.3	2	4652730.1	32.147	.000
	Residual	4052578.4	28	144734.94		
	Total	13358039	30			
3	Regression	9871760.2	3	3290586.7	25.484	.000
	Residual	3486278.6	27	129121.43		
	Total	13358039	30			
4	Regression	10456820	4	2614204.9	23.428	.000
	Residual	2901218.9	26	111585.34		
	Total	13358039	30			
5	Regression	11004290	5	2200858.1	23.376	.000
	Residual	2353748.2	25	94149.928		
	Total	13358039	30			

§5.3 逐步回归

一、问题的提出及逐步回归的思想

前进法依次引入了变量 $x_7, x_4, x_{10}, x_3, x_{11}$，最优回归模型为

$$\hat{y} = -140.625 + 5.090x_3 - 1.997x_4 + 3.910x_7 + 18.431x_{10} - 7.442x_{11}$$

复判定系数 $R^2=0.824$，调整的复判定系数为 $R_a^2=0.789$，

而全模型的复判定系数 $R^2=0.875$，调整的复判定系数为 $R_a^2=0.791$。

§5.3 逐步回归

二、后退法

后退法与前进法相反，首先用全部 m 个变量建立一个回归方程，然后在这 m 个变量中选择一个最不重要的变量，将它从方程中剔除。设对 m 个回归系数进行 F 检验，记求得的 F 值为$\{F_1^m, F_2^m, \cdots, F_m^m\}$，选其最小者记为：

$$F_j^m = \min\{F_1^m, F_2^m, \cdots, F_m^m\}$$

给定显著性水平 α，若$F_j^m \leqslant F_\alpha(1, n-m-1)$，则首先将 x_j 从回归方程中剔除，为方便，设 x_j 就是 x_m。

§5.3 逐步回归

二、后退法

接着对剩下的 m-1 个自变量重新建立回归方程，进行回归系数的显著性检验，像上面那样计算出 F_j^{m-1}，如果又有 $F_j^{m-1} \leqslant F_\alpha$（1,n-（m-1）-1），则剔除 x_j，重新建立 y 关于 m-2 个自变量的回归方程，依此下去，直至回归方程中所剩余的 p 个自变量的 F 检验值均大于临界值 $F_\alpha(1,n-p-1)$，没有可剔除的自变量为止。这时，得到的回归方程就是最终确定的方程。

§5.3 逐步回归

二、后退法

续例5.4 对例3.1国际旅游外汇收入y对第三产业的12个变量做回归的数据，用后退法做变量选择，取显著性水平$\alpha_{出}$=0.10。

首先进入线性回归对话框，将y与x1至x12分别选入各自的变量框，然后在Method对话框中点选后退法Backward,点选Options选项看到默认的显著性水平$\alpha_{出}$正是0.10。部分运行结果见表5.4：

§5.3 逐步回归

二、后退法

Coefficients

		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
Model		B	Std. Error	Beta		
8	(Constant)	-184.690	98.357		-1.878	0.0721
	X3	4.325	0.873	0.9825	4.955	0.0000
	X8	-20.188	7.089	-0.6813	-2.848	0.0087
	X9	17.334	7.102	1.0377	2.441	0.0221
	X10	11.644	6.450	0.2815	1.805	0.0831
	X11	-12.998	3.558	-0.9625	-3.653	0.0012

§5.3 逐步回归

二、后退法

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.935	.875	.791	304.8038
2	.935	.875	.802	296.7067
3	.935	.875	.812	289.3330
4	.935	.874	.820	282.8410
5	.933	.870	.823	281.0489
6	.931	.867	.827	277.6026
7	.929	.864	.830	275.2454
8	.923	.851	.822	281.7979

§5.3 逐步回归

二、后退法

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	11685742	12	973811.87	10.482	.000
	Residual	1672296.2	18	92905.347		
	Total	13358039	30			
2	Regression	11685377	11	1062307.0	12.067	.000
	Residual	1672662.2	19	88034.853		
	Total	13358039	30			
3	Regression	11683766	10	1168376.6	13.957	.000
	Residual	1674272.2	20	83713.612		
	Total	13358039	30			
4	Regression	11678059	9	1297562.1	16.220	.000
	Residual	1679979.8	21	79999.039		
	Total	13358039	30			
5	Regression	11620291	8	1452536.4	18.389	.000
	Residual	1737747.2	22	78988.510		
	Total	13358039	30			
6	Regression	11585585	7	1655083.6	21.477	.000
	Residual	1772453.4	23	77063.193		
	Total	13358039	30			
7	Regression	11539798	6	1923299.6	25.387	.000
	Residual	1818241.0	24	75760.040		
	Total	13358039	30			
8	Regression	11372787	5	2274557.4	28.643	.000
	Residual	1985251.8	25	79410.074		
	Total	13358039	30			

§5.3 逐步回归

三、逐步回归法

逐步回归的基本思想是“有进有出”。具体做法是将变量一个一个引入，当每引入一个自变量后，对已选入的变量要进行逐个检验，当原引入的变量由于后面变量的引入而变得不再显著时，要将其剔除。这个过程反复进行，直到既无显著的自变量选入回归方程，也无不显著自变量从回归方程中剔除为止。这样就避免了前进法和后退法各自的缺陷，保证了最后所得的回归子集是“最优”回归子集。

§5.3 逐步回归

三、逐步回归法

在逐步回归中需要注意的一个问题是引入自变量和剔除自变量的显著性水平α值是不相同的，要求

$$\alpha_{进}<\alpha_{出}$$

否则可能产生“死循环”。也就是当$\alpha_{进}\geq\alpha_{出}$时，如果某个自变量的显著性P值在$\alpha_{进}$与$\alpha_{出}$之间，那末这个自变量将被引入、剔除、再引入、再剔除、…，循环往复，以至无穷。

§5.3 逐步回归

三、逐步回归法

续例5.4 对例3.1国际旅游外汇收入y对第三产业的12个变量做回归的数据，用逐步回归法做变量选择，取显著性水平α进=0.05，α出=0.10。

首先进入线性回归对话框，将y与x1至x12分别选入各自的变量框，然后在Method对话框中点选逐步回归法Stepwise,点选Options选项看到默认的显著性水平正是α进=0.05，α出=0.10。部分运行结果见表5.5：

§5.3 逐步回归

三、逐步回归法

从表5.5看到，逐步回归的最优回归子集为模型7，回归方程为：

$$\hat{y} = -117.497 + 4.975x_3 + 21.479x_{10} - 11.264x_{11}$$

逐步回归的选元过程为第一步引入 x_7；第二步引入 x_4；第三步引入 x_{10}，第四步引入 x_3；第五步引入 x_{11}；第六步剔除 x_7；第七步剔除 x_4。

§5.3 逐步回归

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.741[a]	.549	.533	455.9279
2	.835[b]	.697	.675	380.4405
3	.860[c]	.739	.710	359.3347
4	.885[d]	.783	.749	334.0439
5	.908[e]	.824	.789	306.8386
6	.901[f]	.812	.783	310.9102
7	.889[g]	.791	.768	321.5075

a. Predictors: (Constant), X7

b. Predictors: (Constant), X7, X4

c. Predictors: (Constant), X7, X4, X10

d. Predictors: (Constant), X7, X4, X10, X3

e. Predictors: (Constant), X7, X4, X10, X3, X11

f. Predictors: (Constant), X4, X10, X3, X11

g. Predictors: (Constant), X10, X3, X11

§5.4 本章小结与评注

一、逐步回归实例分析

例5.5 为了研究香港股市的变化规律，此例以恒生指数为例，建立回归方程，分析影响股票价格趋势变动的因素。这里我们选了6个影响股票价格指数的经济变量：

x_1(百万$) —成交额，　　x_2—九九金价（$/两),

x_3—港汇指数，　　x_4—人均生产总值(现价$),

x_5—建筑业总开支(现价百万$),

x_6—房地产买卖金额(百万$),

x_7—优惠利率(最低%)。

y为恒生指数。

§5.3 逐步回归

年份	y	x1	x2	x3	x4	x5	x6	x7
1974	172.9	11246	681	105.9	10183	4110	11242	9
1975	352.94	10335	791	107.4	10414	3996	12693	6.5
1976	447.67	13156	607	114.4	13134	4689	16681	6
1977	404.02	6127	714	110.8	15033	6876	22131	4.75
1978	409.51	27419	911	99.4	17389	8636	31353	4.75
1979	619.71	25633	1231	91.4	21715	12339	43528	9.5
1980	1121.17	95684	2760	90.8	27075	16623	70752	10
1981	1506.94	105987	2651	86.3	31827	19937	125989	16
1982	1105.79	46230	2105	125.3	35393	24787	99468	10.5
1983	933.03	37165	3030	107.4	38823	25112	82478	10.5
1984	1008.54	48787	2810	106.6	46079	24414	54936	8.5
1985	1567.56	75808	2649	115.7	47871	22970	87135	6
1986	1960.06	123128	3031	110.1	54372	24403	129884	6.5
1987	2884.88	371406	3644	105.8	65602	30531	153044	5
1988	2556.72	198569	3690	101.6	74917	37861	215033	5.25

§5.3 逐步回归

	Y	X1	X2	X3	X4	X5	X6	X7
Y	1.0000	0.9171	0.8841	-0.0425	0.9382	0.8786	0.9372	-0.0955
X1	0.9171	1.0000	0.7375	-0.1293	0.7842	0.6973	0.7817	-0.1732
X2	0.8841	0.7375	1.0000	-0.1083	0.9195	0.9477	0.8747	0.1517
X3	-0.0425	-0.1293	-0.1083	1.0000	0.0725	0.0469	-0.0952	-0.4164
X4	0.9382	0.7842	0.9195	0.0725	1.0000	0.9601	0.9137	-0.1409
X5	0.8786	0.6973	0.9477	0.0469	0.9601	1.0000	0.9167	0.0666
X6	0.9372	0.7817	0.8747	-0.0952	0.9137	0.9167	1.0000	0.0617
X7	-0.0955	-0.1732	0.1517	-0.4164	-0.1409	0.0666	0.0617	1.0000

§5.3 逐步回归

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.938[a]	.880	.871	295.57599
2	.983[b]	.966	.960	164.57982
3	.991[c]	.981	.976	126.49374

a. Predictors: (Constant), x4

b. Predictors: (Constant), x4, x1

c. Predictors: (Constant), x4, x1, x6

§5.3 逐步回归

ANOVA[d]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	8347442	1	8347442.122	95.547	.000[a]
	Residual	1135747	13	87365.165		
	Total	9483189	14			
2	Regression	9158151	2	4579075.528	169.054	.000[b]
	Residual	325038.2	12	27086.517		
	Total	9483189	14			
3	Regression	9307182	3	3102393.974	193.892	.000[c]
	Residual	176007.3	11	16000.667		
	Total	9483189	14			

a. Predictors: (Constant), x4

b. Predictors: (Constant), x4, x1

c. Predictors: (Constant), x4, x1, x6

d. Dependent Variable: y

§5.3 逐步回归

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-147.176	151.913		-.969	.350
	x4	.038	.004	.938	9.775	.000
2	(Constant)	38.563	91.146		.423	.680
	x4	.023	.003	.569	6.604	.000
	x1	.004	.001	.471	5.471	.000
3	(Constant)	75.807	71.109		1.066	.309
	x4	.013	.004	.319	3.038	.011
	x1	.004	.001	.417	6.086	.000
	x6	.004	.001	.319	3.052	.011

a. Dependent Variable: y