**26. 逐步回归**

**一、原理**

多元线性回归模型中，并不是所有的自变量都与因变量有显著关系，有时有些自变量的作用可以忽略。这就产生了怎样从大量可能有关的自变量中挑选出对因变量有显著影响的部分自变量的问题。

逐步回归的基本思想是，将变量一个一个引入，引入变量的条件是偏回归平方和经检验是显著的，同时每引入一个新变量后，对已选入的变量要进行逐个检验，将不显著变量剔除，这样保证最后所得的变量子集中的所有变量都是显著的。这样经若干步以后便得“最优”变量子集。

逐步回归每一步只有一个变量单独的回归因子引进或从当前的回归模型中剔除，当没有回归因子能够引进模型时，该过程停止。

**二、步骤**

记引入变量*F*检验的临界值为*Fin*，剔除变量*F*检验的临界值为*Fout*，一般取*Fin* ≥*Fout*. 其确定原则：对*k*个自变量中的*m*个(*m*≤*k*)，则记*F\**=F(1,N-m-1)，取*Fin*=*Fout*= *F\**.

一般也可以直接取*Fin*=*Fout*=2.0或2.5. 要想回归方程多进入一些自变量，甚至也可以取为1.0或1.5.

1. 前进法（变量增加法）

首先对全部*k*个自变量，分别对因变量*Y*建立一元回归方程，并分别计算*k*个回归系数的*F*检验值，记为{}，记为 = max{}, 若有≥*Fin*，则将*X*i引入回归方程，记为X1. 接着考虑*X*1分别与其它自变量与因变量*Y*二元回归方程，选出最大F值若大于*Fin,* 则将该变量引入回归方程，记为*X*2, 依次类推，直到所有末被引入回归方程的自变量的*F*值均小于*Fin*时为止。

增加法的缺点是，它不能反映后来变化的情况，已引入的变量因为后引入的某个自变量的影响，它可能由显著变为不显著，但并没有及时从回归方程中将它剔除掉。

2. 后退法（变量减少法）

首先建立全部自变量*X*1,*X2,*...,*Xk*对因变变量*Y*的回归方程，然后分别求出*k*个回归系数的*F*值为{}，记=min{},若有≤*Fout*，则将自变量*Xi*从回归方程中剔除掉。以此类推，直到在回归方程中的自变量的*F*值均大于*Fout*为止。

减少法的缺点是，一开始把全部变量都引入回归方程，这样计算量比较大。

3. 逐步筛选法（变量增减法）

在实际的数据中，自变量*X*1,*X*2,...,*Xk*之间往往并不是独立的，而是有一定的相关性存在的，这就会使得随着回归方程中变量的增加和减少，某些自变量对回归方程的贡献也会发生变化。

因此将前两种方法综合起来，也就是对每一个自变量，随着其对回归方程贡献的变化，它随时可能被引入回归方程或被剔除出去，最终的回归模型是在回归方程中的自变量均为显著的，不在回归方程中的自变量均不显著。

**三、需要注意的问题**

1. 实际中，往往不使用每次计算出的Fin和Fout, 而是使用固定的F 值（避免出现Fin< Fout, 造成死循环）；

2. 自变量进入模型的顺序并不反映它们的重要程度；

3. 逐步回归方法的一个局限是，预先假定有单一的最优X变量子集，也可能还有其它最优选择；

4. 逐步回归方法的另一个缺点是，当X变量高度相关时，有时会得到不合理的“最优”子集。

**四、PROC REG过程步**

基本语法：

PROC REG data = 数据集;

MODEL *因变量 = 自变量列表 </* *SELECTION=选项>*;

SELECTION=选项

用于逐步回归中选择合适的建立模型方法：

none——全部变量都选

forward——前进法

backward——后退法

stepwise——逐步筛选法

maxr——最大R2增量法

minr——最小R2增量法

rsquare——R2选择法

cp——Mallow's Cp选择法

adjrsq——修正的R2选择法

注：在前进法、后退法或逐步回归的变量选择过程中，都有一个判断是否可进入或剔除的显著水平，分别由model语句的选项slentry=和slstay＝设定的：



start= s指定先从s个自变量的回归模型开始；

stop=s指定找到s个自变量的回归模型结束；

Mallows的*Cp*统计量——是由Mallows提出的作为选择模型的判别式的变量。*Cp*是一个误差平方总和的量度：



其中，*P*是模型中包括截距项的参数个数，*MSE*是满模型时均方误差，*ESSp*是具有*P*个自变量（包括截距项）回归模型的误差平方和。作*Cp*与*P*的曲线图，Matlows建议取*Cp* 首次接近*P*的地方的模型。

常用的输出可选项：

STB——输出标准化偏回归系数矩阵

CORRB——输出参数估计矩阵

COLLINOINT——对自变量进行共线性分析

P——输出个体观测值、预测值及残差 （R/CLM/CLI包含P）

R——输出每个个体观测值、残差及标准误差

CLM——输出因变量均值95%的置信界限的上下限

CLI——对各预测值输出95%的置信界限的上下限

MSE——要求输出随机扰动项方差𝜎2的估计

与残差分析有关的可选项

 VIF——输出变量间相关性的方差膨胀系数，VIF越大，说明由于共线性存在，使方差变大；

COLLIN——输出条件数，它表示最大的特征值与每个自变量特征值之比的平方根。一般情况下，条件数越大越可能存在共线性；

TOL——表示共线性水平的容许值，TOL越小说明其可用别的自变量解释的部分多，自然可能与别的自变量存在共线性关系；

DW——输出Durbin-Watson统计量；

influence——对异常点进行诊断，对每一观测点输出统计量（Cook’s D > 50%, defits/debetas > 2说明该点影响较大）。

**2. 绘制回归分析的图形**

在PROC REG过程步加入绘图选项语句即可。

基本语法：

PROC REG data = 数据集PLOTS *=* (*图形类型*);

可选的绘图类型：

FITPLOT——带回归线、置信预测带的散点图；

RESIDUALS——自变量的残差图；

DIAGNOSTICS——诊断图（包括下面各图）；

COOKSD——Cook's D统计量图；

OBSERVEDBYPREDICTED——根据预测值的因变量图；

QQPLOT ——检验残差正态性的QQ图；

RESIDUALBYPREDICTED——根据预测值的残差图；

RESIDUALHISTOGRAM——残差的直方图；

RFPLOT——残差拟合图；

RSTUDENTBYLEVERAGE——杠杆比率的学生化残差图；

RSTUDENTBYPREDICTED——预测值的学生化残差图；

注：残差图(RESIDUALS)和诊断图(DIAGNOSTICS)是自动生成的，根据模型也有其它默认的图形输出；若只绘制指定的图形需要加上ONLY:

**PROC** **REG** data = 数据集 PLOTS(ONLY) = (图形类型);

**例1**用逐步回归模型，来研究耗氧量的是如何依赖其它变量的。

31位成年人心肺功能的调查数据（见下表）

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| age | weight | Oxygen  耗氧量 | runtime  跑15英哩的时间（分） | rstpulse  休息时每分钟心跳次数 | runpulse  跑步时每分钟心跳次数 | Maxpulse  每分钟心跳次数最大值 |
| 44 | 89.47 | 44.609 | 11.37 | 62 | 178 | 182 |
| 40 | 75.07 | 45.313 | 10.07 | 62 | 185 | 185 |
| 44 | 85.84 | 54.297 | 8.65 | 45 | 156 | 168 |
| 42 | 68.15 | 59.571 | 8.17 | 40 | 166 | 172 |
| 38 | 89.02 | 49.874 | 9.22 | 55 | 178 | 180 |
| 47 | 77.45 | 44.811 | 11.63 | 58 | 176 | 176 |
| 40 | 75.98 | 45.681 | 11.95 | 70 | 176 | 180 |
| 43 | 81.19 | 49.091 | 10.85 | 64 | 162 | 170 |
| 44 | 81.42 | 39.442 | 13.08 | 63 | 174 | 176 |
| 38 | 81.87 | 60.055 | 8.63 | 48 | 170 | 186 |
| 44 | 73.03 | 50.541 | 10.13 | 45 | 168 | 168 |
| 45 | 87.66 | 37.388 | 14.03 | 56 | 186 | 192 |
| 45 | 66.45 | 44.754 | 11.12 | 51 | 176 | 176 |
| 47 | 79.15 | 47.273 | 10.60 | 47 | 162 | 164 |
| 54 | 83.12 | 51.855 | 10.33 | 50 | 166 | 170 |
| 49 | 81.42 | 49.156 | 8.95 | 44 | 180 | 185 |
| 51 | 69.63 | 40.836 | 10.95 | 57 | 168 | 172 |
| 51 | 77.91 | 46.672 | 10.00 | 48 | 162 | 168 |
| 48 | 91.63 | 46.774 | 10.25 | 48 | 162 | 164 |
| 49 | 73.37 | 50.388 | 10.08 | 76 | 168 | 168 |
| 57 | 73.37 | 39.407 | 12.63 | 58 | 174 | 176 |
| 54 | 79.38 | 46.080 | 11.17 | 62 | 156 | 165 |
| 52 | 76.32 | 45.441 | 9.63 | 48 | 164 | 166 |
| 50 | 70.87 | 54.625 | 8.92 | 48 | 146 | 155 |
| 51 | 67.25 | 45.118 | 11.08 | 48 | 172 | 172 |
| 54 | 91.63 | 39.203 | 12.88 | 44 | 168 | 172 |
| 51 | 73.71 | 45.790 | 10.47 | 59 | 186 | 188 |
| 57 | 59.08 | 50.545 | 9.93 | 49 | 148 | 155 |
| 49 | 76.32 | 48.673 | 9.40 | 56 | 186 | 188 |
| 48 | 61.24 | 47.920 | 11.50 | 52 | 170 | 176 |
| 52 | 82.78 | 47.467 | 10.50 | 53 | 170 | 172 |

代码：

**data** fitness ;

input age weight oxygen runtime rstpulse runpulse maxpulse;

datalines;

44 89.47 44.609 11.37 62 178 182

40 75.07 45.313 10.07 62 185 185

44 85.84 54.297 8.65 45 156 168

42 68.15 59.571 8.17 40 166 172

38 89.02 49.874 9.22 55 178 180

47 77.45 44.811 11.63 58 176 176

40 75.98 45.681 11.95 70 176 180

43 81.19 49.091 10.85 64 162 170

44 81.42 39.442 13.08 63 174 176

38 81.87 60.055 8.63 48 170 186

44 73.03 50.541 10.13 45 168 168

45 87.66 37.388 14.03 56 186 192

45 66.45 44.754 11.12 51 176 176

47 79.15 47.273 10.60 47 162 164

54 83.12 51.855 10.33 50 166 170

49 81.42 49.156 8.95 44 180 185

51 69.63 40.836 10.95 57 168 172

51 77.91 46.672 10.00 48 162 168

48 91.63 46.774 10.25 48 162 164

49 73.37 50.388 10.08 76 168 168

57 73.37 39.407 12.63 58 174 176

54 79.38 46.080 11.17 62 156 165

52 76.32 45.441 9.63 48 164 166

50 70.87 54.625 8.92 48 146 155

51 67.25 45.118 11.08 48 172 172

54 91.63 39.203 12.88 44 168 172

51 73.71 45.790 10.47 59 186 188

57 59.08 50.545 9.93 49 148 155

49 76.32 48.673 9.40 56 186 188

48 61.24 47.920 11.50 52 170 176

52 82.78 47.467 10.50 53 170 172

;

**run**;

**proc** **reg** data= fitness;

model oxygen = age weight rstpulse maxpulse runpulse runtime /selection=stepwise ;

**run**;

**运行结果：**

逐步选择: 第 1 步

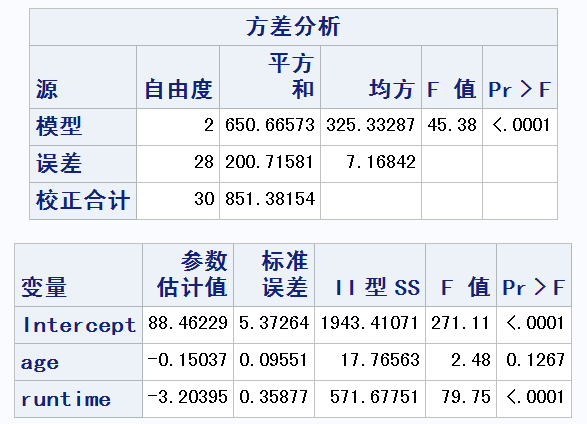
变量 runtime 已输入: R 方 = 0.7434 和 C(p) = 13.5198



条件数字的边界: 1, 1

逐步选择: 第 2 步

变量 age 已输入: R 方 = 0.7642 和 C(p) = 12.2249



条件数字的边界: 1.0369, 4.1478

逐步选择: 第 3 步

变量 runpulse 已输入: R 方 = 0.8111 和 C(p) = 6.8278



条件数字的边界: 1.3548, 11.597

逐步选择: 第 4 步

变量 maxpulse 已输入: R 方 = 0.8368 和 C(p) = 4.7661

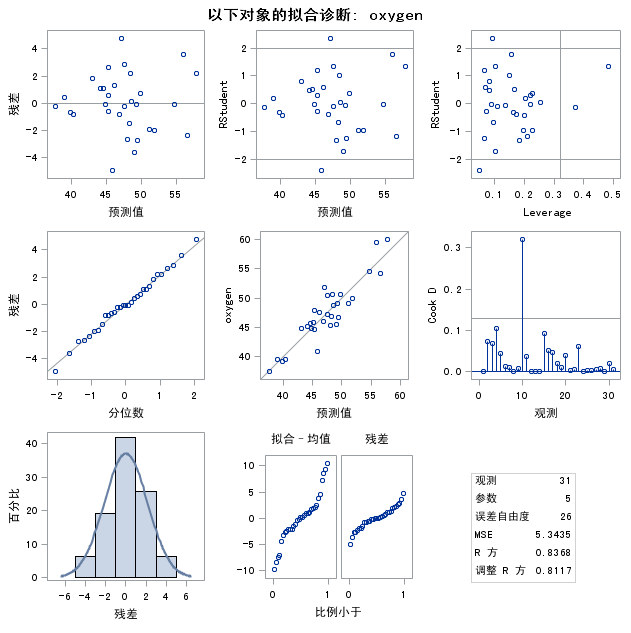


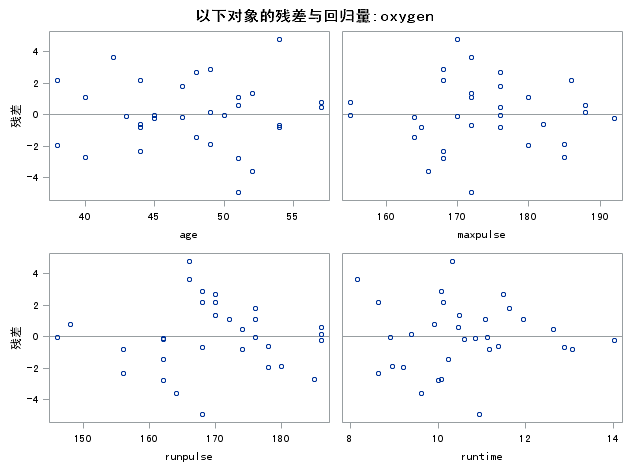
条件数字的边界: 8.4182, 76.851

留在模型中的所有变量的显著性水平都为 0.1500。

没有其他变量满足 0.1500 显著性水平，无法输入该模型。







**结果说明：**

（1）在显著水平015下，用逐步回归法挑选出四个自变量按runtime, age, runpulse, maxpulse 先后被选入回归模型。其它变量在0.15水平下是不显著的。

（2）同时还输出了每个回归模型变化时的R2值增加值、R2值、CP值、相应的F统计量、P值。另外，在每步还列出了条件指数的最小值最大值，以及每一个回归变量的第Ⅱ类平方和。

（3）age变量进入模型后，R2值的增加值（称为偏R2）为

(650.66573-632.90010)/851.38154=0.020867 =0.7642-0.7434

如果按CP值选择最优子集，随着进入回归模型中的自变量个数P从2到5个（包括截距），相应CP值从大到小为

13.5198、12.2249、6.8278、4.7661

按照Mallows提出的回归模型最优自变量个数的选择准则，CP=4.7661是最接近自变量个数P=5的模型。当P=5时，

CP=138.93001792/5.39197－(31－2×5)= 4.76608569

因此，用逐步回归方法及CP值确认的拟合回归模型为：

oxygen = 98.14789-0.19773age+0.27051maxpulse

-0.34811runpulse-2.76758runtime

（4）条件指数为最大特征值和每个特征值之比的平方根。当模型进入第四个自变量maxpulse时，最大的条件指数从较小11.597变成了较大76.851，说明存在一定程度的共线性，根据前面上篇**例1**的分析，这个共线性方程可能为runpulse－maxpulse=0.

若引入新变量代替线性组合关系的变量：

maxrun=maxpulse-runpulse

重新进行逐步回归：

**data** fitness2;

set fitness;

maxrun=maxpulse-runpulse;

**run**;

**proc** **reg** data= fitness2;

model oxygen = age weight rstpulse maxrun runtime /selection=stepwise ;

**run**;

得到回归方程为：（其它输出略）

oxygen = 89.48742 -0.16297 age -0.08063weight

+0.38826 maxrun-2.80555 runtime

**例2 同例1，**用**R2选择法逐步回归**给出全部可能回归的程序，并且以R2值的大到小排序输出。

**代码：**

**proc** **reg** data= fitness;

model oxygen = age weight rstpulse maxpulse runpulse runtime /selection= rsquare b best=2;

\*选项 b 输出各回归系数;

**run**;

**运行结果：**



例如，含2个自变量按R2值第二大（0.7614）的回归模型为：

oxygen= 93.08877－0.07351runpulse－3.14019runtime

**例3 例2**中的各个回归模型到底选哪个呢？方法是先用Mallows的Cp方法确定变量个数；对所有可能的回归按Cp由小到大进行排序并保留其前5种，并绘制Cp图。

**代码：**

goptions reset=global gunit=pct cback=white border

htitle=**6** htext=**3** ftext=swissb colors=(back);

title 'Cp plot with Reference Lines';

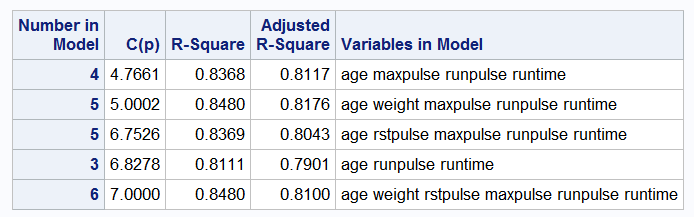
**proc** **reg** data= fitness;

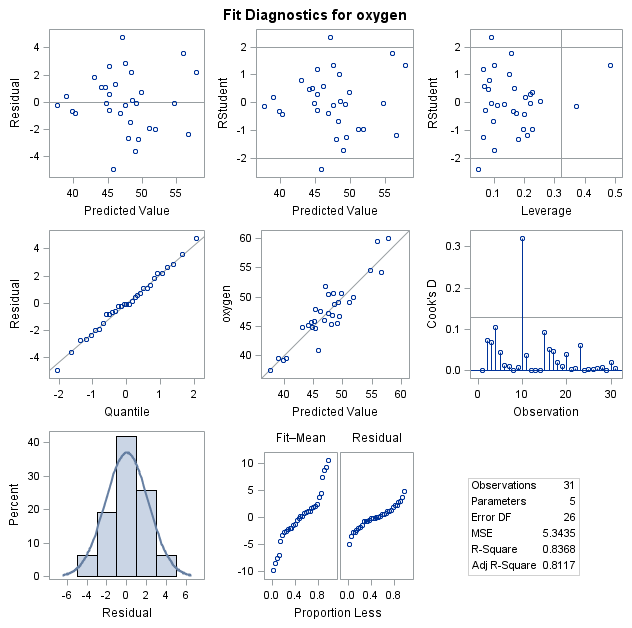
model oxygen = age weight rstpulse maxpulse runpulse runtime /selection=cp adjrsq best=**5** ;

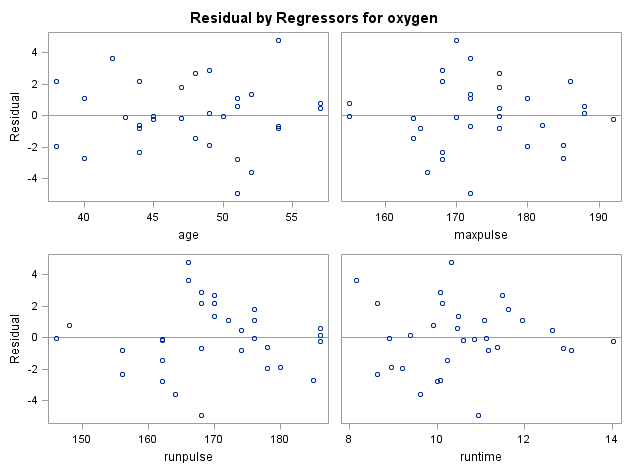
plot cp. \* np. /chocking=red cmallows=blue vaxis=**0** to **15** by **2** haxis=**0** to **8** by **1**;

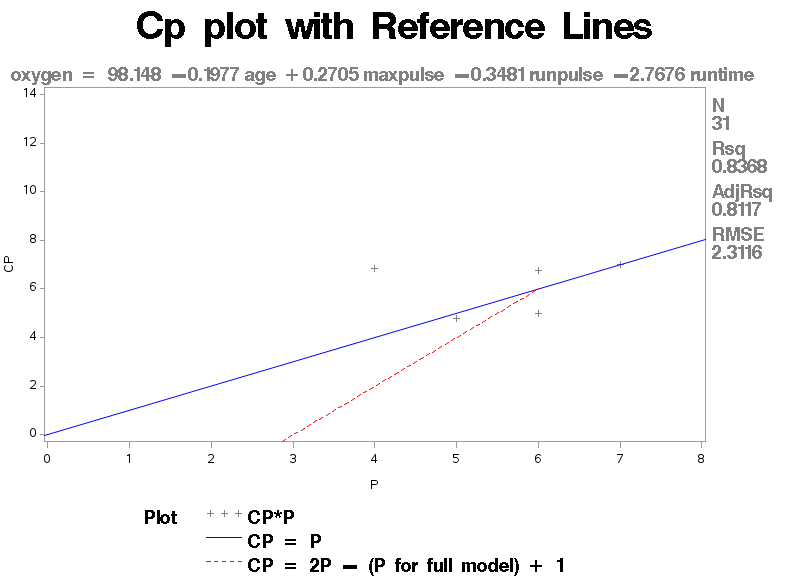
**run**;

**运行结果：**









**程序说明：**

（1） 选项selection=cp 指定用Mallow's Cp选择法。选项*adjrsq*表示要输出每种回归模型的统计量Adj-*R*2. 选项best=5表示保留*Cp*值最小的前5种回归模型；

（2）plot语句中的cp. \* np.表达式（**注意统计量关键字母后的小圆点**）表示Y轴为*Cp*值X轴为*P*值（*P*值包括截距项）；

（3）Hocking（1976）建议选择满足Cp≤2P－Pfull 且Cp≤P的模型。plot语句的选项chocking=red，表示画*Cp*=2*P*－*Pfull* 红色参考虚线，其中*P*是子模型中含截距的参数个数，*Pfull*是全模型中不含截距的参数个数；

（4）Mallows（1973）建议考虑所有满足*Cp*较小且接近*P*的模型。plot语句的选项cmallows=blue，表示画*Cp*=*P*蓝色参考实线，其中*P*是子模型中含截距的参数个数；

（5）从输出结果看，以Mallows的建议为标准，*age*, *maxpulse*, *runpulse*和*runtime*四个变量进入回归模型时*Cp*最小(4.76609)，且与*P*=4+1=5最接近，因为5－4.76609<6－5.00021，所以应该选择*Cp*=4.76609的模型。以Hocking的建议为标准，*Cp*=4.76609的模型不满足要求，因为4.76609>2×5－6=4。而*Cp*=5.00021的模型满足要求，因为5.00021<2×6－6=6且5.00021<6。而*Cp*=6.75259的模型，无论以Mallows的建议还是以Hocking的建议都不满足要求。

其实我们不必具体计算，只要看输出图形，表示*Cp*值大小位置的“+”号，在蓝实线附近即满足Mallows的建议的标准，在红虚线下方即满足Hocking的建议的标准。但从Adj-*R*2看，*Cp*=4.76609的模型将变量*weight*加入后，*R*2更大(0.84800181>0.83681815). 不同的标准提供不同的选择结果，这是常有的情况。