DOI:10.13656/j.cnki.gxkx.1995.04.005

广西科学 Guangxi Sciences 1995，2（4）：17～18，68

# 基于 $Cp$ 统计量的自变量选择原则

# The Independent Variable Selection Rule Based on the $Cp$ Statistics

周婉枝
Zhou Wanzi

（广西大学数学与信息科学系，
南宁市西乡塘路 530004）
(Dept. of Math. & Inf's Sci. Guangxi University, 10 Xixiangtang Road, Nanning, Guangxi, 530004)

陈宇丹
Chen Yudan

（广西医科大学基础部
南宁市滨湖路 6 号 530021）
(Basics Dept., Guangxi Medical University, 6 Binhu Road, Nanning, 530021)

**摘要** 对回归模型 $y=\beta_0+\beta_1x_1+\cdots+\beta_tx_t$（其中 $y$ 是 $q\times1$ 随机向量，$\beta i$ 为 $q\times1$ 维参数向量），提出在 $q\geqslant1$ 的情况下，基于 $Cp$ 统计量的自变量的选择原则：选择自变量是子集 $P$，使其相对应的 $Cp$ 值满足条件 $Cp\leqslant\gamma_a(t,n-t-1,q)$.

**关键词** 选择原则 $Cp$ 统计量 置信椭球

**Abstract** For the regression model $y=\beta o+\beta_1x_1+\cdots+\beta_tx_1$, where $y$ is $q\times1$ random vector and $\beta_i$ is $q\times1$ parameter vecter, the independent variable selection rule based on the $Cp$ statistics under the condition of $q\geqslant1$: The selection of the independent variable subset $P$ can be done through the calculation of $Cp$ corresponding to $P$, which satisfied $Cp\leqslant\gamma_a(t,n-t-1,q)$.

**Key words** selection rule, $Cp$ statistics, confidence ellipsoid

## 1 $Cp$ 统计量

对回归模型 $y=\beta_0+\beta_1x_1+\cdots+\beta_tx_t$，1964 年 C. L Mallows 在 $q=1$ 的条件下提出了 $Cp$ 统计量及基于 $Cp$ 统计量的自变量选择原则[1]，本文讨论在 $q\geqslant1$ 的条件下自变量的选择原则.

### 1.1 模型

$$\text{原模型}\quad\begin{cases}Y=XB+e\\e=(e_{(1)}\cdots e_{(n)}')'\\e_{(1)}\cdots e_{(n)}\quad\text{相互独立且同分布 }Ng(0.\Sigma)\end{cases}\qquad(1)$$

这里 $Y$ 为 $n\times q$ 随机观察阵，$X$ 为 $n\times(t+1)$ 设计矩阵，$B$ 为 $(t+1)\times q$ 参数矩阵，$e$ 为 $n\times q$ 误差阵选择自变量子集 $P$，相应于 $P$ 的资料阵 $X$.

$$\text{选模型}\quad\begin{cases}Y=X_*B_*+e\\e=(e_{(1)}'\cdots e_{(n)}')'\\e_{(1)}\cdots e_{(n)}\quad\text{相互独立同分布 }Ng(0.\Sigma)\end{cases}\qquad(2)$$

### 1.2 $Cp$ 统计量

在模型(1)真下，选用模型(2)，文献[2]在损失函数

$$Jp=\mathrm{tr}[\Sigma^{-1}(Y_*-XB)'(Y_*-XB)]$$

导出了统计量

$$Cp=(2p-t-1)q+(n-t-q-2)\mathrm{tr}[(RSS)^{-1}(RSS_*-RSS)]$$

1995-06-23 收稿.

## 2 主要结果

定义：$P$ 是自变量子集，$X$ 分剖为 $(X_1,X_R)$，其中 $X_1$ 相应于子集 $P$ 的资料阵. $B$ 相应地剖成 $B_1,B_R$，若 $B_R=0$，则称 $P$ 为合宜的.

利用此定义，若 $P$ 合宜，选用模型(2)是正确的.

### 2.1 模型中心化

对模型(1)，把常数项单独列出改写为

$$\begin{cases}Y=1\beta_0'+XB_t+e\qquad 1\text{ 为元素均为 1 的向量}\\e_{(1)}\cdots e_{(n)}\text{独立同分布 }Ng\ (0.\ \Sigma)\end{cases}$$

令 $\tilde{X}=X-(1/n)\ 11'X=(I-(1/n)\ 11_n')\ X$

令 $\tilde{\beta}_0'=\beta_0'+(1/n)\ 1'XB_t$，则 (1) 改写为

$$\begin{cases}Y=1\tilde{\beta}_0'+\tilde{X}B_t+e\\e_{(1)}\cdots e_{(n)}\text{独立同分布 }Ng\ (0.\ \Sigma)\end{cases}$$

此时，由于 $1'\tilde{X}=0$，计算得

$$\begin{pmatrix}\hat{\beta}_0'\\ \hat{B}_t\end{pmatrix}=\begin{pmatrix}(1/n)\ \ 1'Y\\ (\tilde{X}'\tilde{X})^{-1}\tilde{X}'Y\end{pmatrix}.$$

$\hat{B}_t=(\tilde{X}'\tilde{X})^{-1}\tilde{X}'Y$ 相当于从 $Y=\tilde{X}B_t+e$ 进行最小二乘估计而得.

## 2.2 回归系数 $B_t$ 的置信椭球

设 $A_n\sim Wq(n,I)$，$D\sim Wq(m,I)$　$A_n$ 与 $D$ 相互独立

$(n-q-1)\ \mathrm{tr}A_n^{-1}D$ 的分布为 $r(m,n,q)$

由文献[2]，可得 $r(m,n,q)$ 的极限分布为 $x^2_{mq}$ $(n\to\infty)$.

**定理**：$\hat{B}_t$ 是模型(1)下 $B_t$ 的最小二乘估计，$RSS$ 为相互残差阵，令 $Q=\tilde{X}'\tilde{X}$，则有

$$(n-t-q-2)\ \mathrm{tr}\left[(RSS)^{-1}(\hat{B}_t-B_t)Q(\hat{B}_t-B_t)\right]\sim r(t,n-t-1,q)$$

**证**：由 2.1 的结论，

$$\hat{B}_t=(\tilde{X}'\tilde{X})^{-1}\tilde{X}'Y$$
$$=B_t+(\tilde{X}'\tilde{X})^{-1}\tilde{X}'e,\mathrm{rank}(\tilde{X})=t$$

故 $(\hat{B}_t-B_t)'Q(\hat{B}_o-B_t)=e'P_{\tilde{X}}e\sim Wq(t,\Sigma)$

而 $RSS\sim Wq(n-t-1,\Sigma)$ 且因为 $(I-(1/n)11'-P_{\tilde{X}})P_{\tilde{X}}=0$

有 $RSS$ 与 $(\hat{B}_t-B_t)'Q(\hat{B}_t-B_t)$ 独立

从而 $(n-t-q-2)\ \mathrm{tr}\left[(RSS)^{-1}(\hat{B}_t-B_t)'Q(\hat{B}_t-B_t)\right]\sim r(t,n-t-1,q)$

$B_t$ 的置信椭球

定义：令 $v=n-t-q-2$

$$S_a=\{B_t:v\cdot\mathrm{tr}[(RSS)^{-1}(\hat{B}_t-B_t)'Q(\hat{B}_t-B_t)\leqslant r_a(t,n-t-1,q)\}$$

$$P\{r(t,n-t-1,q)\leqslant r_a(t,n-t-1,q)\}=1-\alpha$$

由 $r(t,n-t-1,q)$ 的渐近分布，$n$ 足够大，$r_a(t,n-t-1,q)$ 可由 $x^2_{iq}$ 的百分位点得到.

## 2.3 自变量子集 $P$ 合宜时与 $Cp$ 的关系

**定理**：以下三点等价　对子集 $P$

(i) $P$ 合宜且 $B_t\in S_a$

(ii) $v\cdot\mathrm{tr}(RSS)^{-1}(RSS_*-RSS)\leqslant r_a(t,n-t-1,q)$

(iii) $Cp\leqslant(2p-t-1)q+r_a(t,n-t-1,q)$

先证

(a) $RSS_*-RSS=(B_*^{(6)}-\hat{B}t)'Q(B_*^{(0)}-\hat{B}_t)$，

其中 $B_*^{(0)}=\begin{pmatrix}(\tilde{B}_{*(-0)})\\ 0\end{pmatrix}$，$\tilde{B}_{*(-0)}$ 是指在选模型下 $\tilde{B}_*$ 的最小二乘估计且除去常数项，对模型中心化后所求的 $\tilde{B}_*$ 正是没有中心化前的 $\tilde{B}_{*(-0)}$，为方便起见，用中心化后求得的 $\tilde{B}_*$ 代替 $\tilde{B}_{*(-0)}$，中心化后 $\tilde{X}$ 的仍用 $\tilde{X}$ 表示. 故 $Q=X'X$

相应于子集 $P$ 记 $Q=\begin{pmatrix}B & C\\ C' & D\end{pmatrix}$　$Q^{-1}=\begin{pmatrix}B_1 & C_1\\ C'_1 & D_1\end{pmatrix}$

由于 $RSS_*-RSS=Y'(P_X-P_{X1})Y$

$$=\hat{B}'_R(D-C'B^{-1}C)\hat{B}_R$$

因 $B_*^{(0)}=(\tilde{B}'_*,0')'$

$$(B_*^{(0)}-\hat{B}_t)'Q(B_*^{(0)}-\hat{B}_t)$$
$$=\tilde{B}'_*B\tilde{B}_*+\hat{B}'_tQ\hat{B}_t-\tilde{B}'_*B\hat{B}_*-\tilde{B}'_*C\hat{B}_R$$
$$-\hat{B}'_*B\tilde{B}_*-\hat{B}'_RC'\tilde{B}_*.$$

由正规方程　$X=(X_1,X_R)$

$$\begin{cases}B\hat{B}_*+C\hat{B}_R=X_1Y\\ C'\hat{B}_*+D\hat{B}_R=X_RY\\ B\tilde{B}_*=X_1Y\end{cases}\Rightarrow$$

$$\begin{cases}\tilde{B}_*=B^{-1}X_1Y\\ \hat{B}_*=B^{-1}\left[X'_1-(D-C'B^{-1}C)^{-1}(X'_R-C'B^{-1}X_1)\right]Y\\ \hat{B}_R=(D-C'B^{-1}C)^{-1}(X'_R-C'B^{-1}X_1)Y\end{cases}$$

代入 $(B_*^{(0)}-\hat{B}_t)'Q(B_*^{(0)}-\hat{B}_t)$ 各项得

$$(B_*^{(0)}-\hat{B}_t)'Q(B_*^{(0)}-\hat{B}_t)=\hat{B}'_R(D-C'B^{-1}C)\hat{B}_R$$

从而 $(a)$ 成立

**定理的证明**：$(ii)\to(i)$，由 $(a)$，

$$RSS_*-RSS=(B_*^{(0)}-\hat{B}_t)Q(B_*^{(0)}-\hat{B}_t)$$

可知 $B_*^{(0)}\in S_a$，且 $B_*^{(0)}=(\tilde{B}'_*,0')'$，故 $P$ 合宜，故 $(i)$ 成立.

$(i)\Rightarrow(ii)$，若 $(i)$ 成立，则存在 $H$，$H=(H_*'\ H_R')'$，$H_R=0$

且 $v\cdot(n-t-q-2)\ \mathrm{tr}\left[(RSS)^{-1}(H-\hat{B}_t)'Q(H-\hat{B}_t)\right]\leqslant r_a(t,n-t-1,q)$

因 $B_*^{(0)}=(\tilde{B}'_*,0')'$，故

$$\hat{B}'_tX'XB_*^{(0)}=Y'(X_1\ \ X_R)\begin{pmatrix}\tilde{B}_*\\ 0\end{pmatrix})=Y'X_1\tilde{B}_*$$

$$B_*^{(0)'}X'XB_*^{(0)}$$
$$=(Y'X_1B^{-1},0)\begin{pmatrix}B & C\\ C' & D\end{pmatrix}\begin{pmatrix}B^{-1}X_1'Y\\ 0\end{pmatrix}$$
$$=Y'X_1\tilde{B}_*$$

所以 $\hat{B}'_tX'XB_*^{(0)}=B_*^{(0)'}X'XB_*^{(0)}$

$$\hat{B}_t'X'XH=Y'X_1H_*$$

$$B_*^{(0)'}X'XH=Y'X_1H_*$$

从而 $(\hat{B}_t-B_*^{(0)})'Q(B_*^{(0)}-H)=0$

$$(\hat{B}_t-H)'Q(B_t-H)$$
$$=(\hat{B}_t-B_*^{(0)})'Q(\hat{B}_t-B_*^{(0)})$$
$$+(B_*^{(0)}-H)'Q(B_*^{(0)}-H)$$

(下转第 68 页 Continue on page 68)

表 6 龙州站洪峰作业预报登记表

Table 6 Record of flood peak forecast from Longzhou station

| 发布时间 Broadcast time | 预报值 Forecast | | 实测值 Measure | | 预见期内水位变幅 Stage fluctuation during forecast period (m) | 洪水总涨幅 Total fluctuation of flood (m) | 发预报时水位 Stage at broad cast time (m) | 预报准确率 Forecast accuracy (%) | 有效预见期 Useful forecast period (h) | 起涨 Start to rise | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 出现时间 Peak time d-h | 水位 Stage (m) | 出现时间 Peak time d-h : min | 水位 Stage (m) | | | | | | 时间 time d-h | 水位 Stage (m) |
| 1971-07-23-20 | 24-23 | 119.50 | 25-07 : 00 | 120.32 | 11.08 | 13.42 | 108.42 | 82.6 | 35.0 | 23-09 | 106.9 |
| 1971-08-29-20 | 30-12 | 116.00 | 30-16 : 0 | 116.68 | 6.68 | 9.20 | 109.32 | 73.9 | 20.0 | 29-08 | 107.4 |
| 1975-09-31-20 | 02-02 | 115.80 | 02-01 : 01 | 116.24 | 6.58 | 8.82 | 109.28 | 74.6 | 39.0 | 31-06 | 107.4 |
| 1978-05-17-05 | 18-08 | 116.00 | 18-16 : 00 | 116.83 | 7.24 | 9.41 | 108.76 | 76.9 | 35.0 | 16-11 | 106.0 |
| 1980-07-24-05 | 25-08 | 120.00 | 25-12 : 00 | 120.21 | 7.24 | 9.91 | 112.76 | 73.1 | 31.0 | 23-20 | 110.30 |
| 1985-09-11-22 | 12-20 | 117.50 | 13-03 : 00 | 118.17 | 5.25 | 7.88 | 112.25 | 66.6 | 29.0 | 11-4 | 110.2 |
| 1986-07-23-05 | 24-12 | 122.00 | 25-01 : 30 | 125.89 | 11.23 | 16.93 | 110.77 | 66.3 | 44.5 | 22-17 | 108.9 |
| 1986-07-23-20 | 24-13 | 125.00 | 25-01 : 30 | 125.89 | 5.55 | 6.44 | 119.45 | 86.2 | 29.5 | 22-17 | 108.9 |

的预报难题。以上预报方案的创立，开创了探讨上游无情报提供的洪水预报科研的良好开端，特别是中小流域，除了可借用毗邻流域的水文，雨量资料进行本河段预报外，尚可以直接应用本站的实测水位、流量、雨量资料进行本站洪峰的预报尝试，解决上游无情报的困难。

《上游无情报的水文预报技术》被江苏场州水利专科学校李慧珑教授称为“教科书没有的、行之有效的方法”

## 参考文献


1 扬振怀等．中国水利百科全书．北京：水利电力出版社，1990．12．

2 华东水利学院水文系编．水文预报．中国工业出版社，1962．8．


（责任编辑：邓大玉　蒋汉明）

（上接第 18 页 Continue from page 18）

$v \cdot \mathrm{tr}\ [\ (RSS)^{-1}\ (RSS_* - RSS)]$

$= v \cdot \mathrm{tr}\ (RSS)^{-1}\ (\hat{B}_t - B_*^{(0)})' Q\ (\hat{B}_t - B_*^{(0)})$

$\leqslant v \cdot \mathrm{tr}\ (RSS)^{-1}\ (\hat{B}_0 - B_*^{(0)})' Q\ (\hat{B}_t - B_*^{(0)})$

$+\ (B_*^{(0)} - H)' Q\ (B_* - H)]$

$= v \cdot \mathrm{tr}\ (RSS)^{-1}\ (\hat{B}_t - H)' Q\ (\hat{B}_t - H)$

$\leqslant r_a\ (t,\ n-t-1,\ q)$

故 $(ii)$ 成立

$(ii) \Leftrightarrow (iii)$ 由 $Cp$ 的定义直接得到 $P$ 合宜与 $Cp$ 的关系

定理的意义在于 $(i) \Leftrightarrow (ii)$，因为 $(i)$ 的含义在于“在置信椭球 $S_a$ 中包含了一个形如 $(B_*'\ \ 0')'$ 的矩阵”，表明从参数 $Bt$ 的区域估计的观点看，“弃后 $r$ 个自变量”与数据是相容（当然在所给的置信系数下），$(i)$ 与 $(iii)$ 的等价性正是把这个事实归于 $Cp$ 的计算，因此，选择 $P$，首先应注意满足 $Cp \leqslant (2p-t-1)\ q + r_a\ (t,\ n-t-1,\ q)$ 的子集．

## 参考文献


1 Mallows C L. Some Commerts on $Cp$. Thehnometrics, 1964, 15: 661～675.

2 周婉枝．$Cp$ 统计量．广西大学学报，1995，20 (3)：291～294.


（责任编辑：莫鼎新　邓大玉）