国防科技大学学报
JOURNAL OF NATIONAL UNIVERSITY OF DEFENSE TECHNOLOGY
第12卷第3期 1990年9月 Vol.12 No.3

一种新的自变量选择准则

王正明
(系统工程与应用数学系)

摘 要 本文提出了一种着眼于提高回归系数估值精度的自变量选取准则，同时给出了相应的算法。

关键词 估计理论，回归分析，自变量，回归系数
分类号 O212.1

考虑线性回归模型：

$$Y=X\beta+e \qquad e\sim N(0,\ \sigma^2 I) \tag{1}$$

其中：$X=X_{nt}$是秩为 t 的矩阵，$n\gg t$。

记：$X=(X_P, X_R)$；$X_P=(x_1, x_2, \cdots x_p)$；$X_R=(x_{p+1}, x_{p+2}, \cdots, x_t)$；

$Z=(Z_P, Z_R)$；$Z_P=X_P$；$T=(X_P'X_P)^{-1}X_P'X_R$；

$X_{PR}=X_PT$；$Z_R=X_R-X_{PR}$；$D=X_P'X_P$；$B=Z_R'Z_R$

为下文讨论问题方便，我们先引入如下引理，见[1]。

引理 1 设$X'X$为 t 阶非奇异矩阵，则

$$(X'X)^{-1}=\begin{pmatrix} D^{-1}+TB^{-1}T' & -TB^{-1} \\ -B^{-1}T' & B^{-1} \end{pmatrix} \tag{2}$$

下面我们来分析自变量选择对参数估计所起的作用。以下总以$\beta=\begin{pmatrix}\beta_P\\ \beta_R\end{pmatrix}$表示模型(1)式中回归系数的真值，以$\beta_{LS}$表示由模型(1)式得到的$\beta$的最小二乘估计，以$\beta_{PLS}$表示由选模型

$$Y=X_P\beta_P+\hat{e} \tag{3}$$

得到的β_P的最小二乘估计，亦即

$$\beta_{LS}=\begin{pmatrix}\beta_{LS}^P\\ \beta_{LS}^R\end{pmatrix}=(X'X)^{-1}X'Y \tag{4}$$

$$\beta_{PLS}=(X_P'X_P)^{-1}X_P'Y \tag{5}$$

综合[1]中的有关结论，可得：

定理 1 设β_{LS}^P和β_{PLS}分别是由(4)、(5)式给出的β_P的估计，则

$$\mathrm{E}(\|\beta_{LS}^P-\beta_P\|^2)=\sigma^2\mathrm{tr}(D^{-1})+\sigma^2\mathrm{tr}(TB^{-1}T') \tag{6}$$

1988年11月12日收稿

$$E(\|\beta_{PLS}-\beta_P\|^2)=\sigma^2\mathrm{tr}(D^{-1})+\beta_R^{\prime}T^{\prime}T\beta_R \tag{7}$$

为方便，记

$$\Delta_P=\sigma^2\mathrm{tr}(TB^{-1}T^{\prime})-\beta_R^{\prime}T^{\prime}T\beta_R \tag{8}$$

由定理1可知，当$\Delta_P>0$时，

$$E(\|\beta_{PLS}-\beta_P\|^2)<E(\|\beta_{LS}^P-\beta_P\|^2) \tag{9}$$

显然，此时用β_{PLS}作β_P的估计比用β_{LS}^P要好。

从估计参数的意义上讲，用Δ_P愈大愈好的原则来选择最优回归模型是合理的，最优选模型中的参数都能得到比较理想的估计，关于这一点读者还可从文末的例子中看得更清楚。该原则还有两个显著优点：一是能找到较好的近似Δ_P的统计量Δ_P^*，二是计算容易实现。

现在我们来构造近似Δ_P的统计量Δ_P^*．由于$n\gg t$，故由[2]知

$$\hat{\sigma}^2=\frac{\|Y-X\beta_{LS}\|^2}{n-t}$$

给出了σ^2的一个较准确的估计，$\hat{\sigma}^2\approx\sigma^2$．下面我们还需构造一个近似$\|T\beta_R\|^2$的统计量，注意到(2)、(4)两式，故有

$$\beta_{LS}^R=B^{-1}Z_R^{\prime}Y \tag{10}$$

于是

$$E(\|T\beta_{LS}^R\|^2)=\|T\beta_R\|^2+\sigma^2\mathrm{tr}(TB^{-1}T^{\prime}) \tag{11}$$

从(11)式看，用$\|T\beta_{LS}^R\|^2$作为$\|T\beta_R\|^2$的估计一般是偏大了；是否可以用$\|T\beta_{LS}^R\|^2-\sigma^2\mathrm{tr}(TB^{-1}T^{\prime})$作为$\|T\beta_R\|^2$的估计呢？当$\|T\beta_{LS}^R\|^2<\sigma^2\mathrm{tr}(TB^{-1}T^{\prime})$时，该设想显然不妥。理想的办法是对统计量$\|T\beta_{LS}^R\|^2$作压缩。

令

$$g(k)\triangleq E(k\|T\beta_{LS}^R\|^2-\|T\beta_R\|^2)^2=\min \tag{12}$$

应用[1]中引理2.1，类似[1](P184)可推得：

$$k=\frac{\|T\beta_R\|^4+\sigma^2\|T\beta_R\|^2\mathrm{tr}(TB^{-1}T^{\prime})}{\|T\beta_R\|^4+6\sigma^2\|T\beta_R\|^2\mathrm{tr}(TB^{-1}T^{\prime})+2\sigma^4\mathrm{tr}(TB^{-1}T^{\prime})^2} \tag{13}$$

记

$$k^*=\frac{\|T\beta_R\|^2+\sigma^2\mathrm{tr}(TB^{-1}T^{\prime})}{\|T\beta_R\|^2+6\sigma^2\mathrm{tr}(TB^{-1}T^{\prime})} \tag{14}$$

则因$g(k)$是二次函数及

$$k<k^*<1,\quad g(k)=\min$$

知

$$g(k)<g(k^*)<g(1)$$

但由于(14)式中包含了未知项σ^2、$\|T\beta_R\|^2$，故k^*仍不能求得，从(11)式看，用以下统计量作为k^*的近似是合适的。

$$k^{**}=\frac{\|T\beta_{LS}^R\|^2}{\|T\beta_{LS}^R\|^2+5\hat{\sigma}^2\mathrm{tr}(TB^{-1}T^{\prime})} \tag{15}$$

定义统计量

$$\Delta_P^*=\hat{\sigma}^2\mathrm{tr}(TB^{-1}T^{\prime})-k^{**}\|T\beta_{LS}^R\|^2 \tag{16}$$

本文把Δ_P^*愈大愈好作为自变量选择的准则。

下面我们给出计算Δ_P^*的方法。由[1]之定理2.6知

$$S_1S_2\cdots S_P\begin{pmatrix}X_P'X_P & X_P'X_R & X_P'Y\\ X_R'X_P & X_R'X_R & X_R'Y\\ Y'X_P & Y'X_R & Y'Y\end{pmatrix}=\begin{pmatrix}D^{-1} & T & \beta_{PLS}\\ * & * & *\\ * & * & \|Y-X_P\beta_{PLS}\|^2\end{pmatrix} \tag{17}$$

这里$S_i(i=1,2,\cdots,p)$是[1](P117)中定义的扫描运算。

一般说来，最优子集X_P并不是由X的前p列组成的，不妨设X_P由X的第i_1，i_2，…，i_p列组成，定义矩阵

$$C=S_{i_1}S_{i_2}\cdots S_{i_p}\begin{pmatrix}X'X & X'Y\\ Y'X & Y'Y\end{pmatrix} \tag{18}$$

记v_i，$\tau_i(i=1,2,\cdots,t)$分别为矩阵$(X'X)^{-1}$，C的第i个对角元；V是由C的第i_1，i_2，…，i_p行、前t列的元素按C中顺序排成的矩阵；a是由下式定义的t维向量。

$$a(j)=\begin{cases}0 & j=i_r\\ \beta_{LS}(j) & j\neq i_r\end{cases}\quad(r=1,2,\cdots,p)$$

类似[1](P116)的推导，我们可证得：

$$\mathrm{tr}(TB^{-1}T')=\sum_{r=1}^{p}(\gamma_{i_r}-\tau_{i_r}) \tag{19}$$

$$T\beta_{LS}^R=Va \tag{20}$$

有了以上准备，我们就可以利用[1]中的扫描程序编出计算Δ_P^*和β_{PLS}的程序了。显然，使用本文的准则并不比使用基于RSS_P的各类准则增加多少计算量，而比PRESS准则的计算量要小得多。

为说明本文方法，我们给出一个模拟计算的例子。该例中有6个自变量，自变量和回归系数的真值为：

$$X=\begin{pmatrix}1 & 1 & 0 & 5 & 1 & 0\\ 3 & 2 & 0 & 8 & 8 & 0\\ 9 & 3 & 0 & 6 & 4 & 8\\ 0 & 4 & 0 & 2 & 0 & 31\\ 8 & 5 & 0 & 17 & 2 & 2\\ 7 & 6 & 0 & 5.1 & 71 & 5\\ 6 & 7 & 6 & 44 & 55 & 7\\ 11 & 8 & 5 & 60 & 0 & 9\\ 2 & 9 & 4 & 0 & 4 & 2\\ 6 & 10 & 3 & 2 & 6 & 1\\ 7.8 & 11 & 2 & 1 & 7 & 0\\ 4 & 12 & 81 & 3 & 33 & 5\end{pmatrix},\quad \beta=\begin{pmatrix}11.05\\ 0.01\\ 0.2\\ -0.002\\ 0.3\\ 10.0\end{pmatrix} \tag{21}$$

因变量的产生过程与[1]中的例[3.8]相同。我们共算了15组随机数，这里列出有代表性的一组，结果如表1.

表 1　几类准则的估值效果

参数	准则			
	$\min C_P$ $\min S_P$ $\min \mathrm{RMS}_P$	$\max \Delta^*_P$	$\max \Delta_P$	β_{LS}
	0.1682D−1		0.1696D−1	
β_1	0.11069D+2	0.11066D+2	0.11065D+2	0.11068D+2
β_2	—	—	0.28521D−2	0.86065D−3
β_3	0.38350D+0	0.25857D+0	0.25812D+0	0.36544D+0
β_4	−0.89727D−1	—	—	−0.86620D−1
β_5	—	—	—	0.11855D−1
β_6	0.10012D+2	0.10017D+2	0.10017D+2	0.10013D+2

我们通过该例说明：Δ^*_P是Δ_P的一个较准确的估计；本文的准则比基于RSS_P的各准则有一个明显的优势，即最优选模型中各自变量对应的参数能得到更准确的估计。

本文承沙钰教授指导，谨表感谢！

参 考 文 献

[1] 陈希孺，王松桂．近代实用回归分析．广西人民出版社，1984
[2] G A F 塞伯．线性回归分析．科学出版社，1987

A New Rule to Select the Best Regression Equations

Wang Zhengming
(Department of Applied Mathematics and System Engineering)

Abstract

In this paper, we derive a rule endeavouring to increase the accuracy of the estimate of parameters. Corresponding compu-tational methods are given.

Key words: estimation theory, regression analysis, independent variable regression coefficient