

协方差改进估计中的变量选择 *

尹素菊 王松桂

(北京工业大学应用数理学院, 北京, 100022)

摘 要

本文探讨了如何选择协变量的问题, 提出了协变量选择的典则相关检验法, 给出了检验统计量的近似分布, 并且在经济、生物和医药等方面常见的三种误差协方差阵的假定下进行了计算机模拟, 其结果显示了协变量选择的必要性及典则相关检验法的优良性.

关键词: 线性模型, 协变量选择, 两步估计, 协方差改进.

学科分类号: O212.1, O212.5.

§ 1. 引 言

我们考虑如下线性模型

$$y = X\beta + e, \quad e \sim N_n(0, V), \quad (1.1)$$

这里 y 为 $n \times 1$ 的观测向量, X 为 $n \times p$ 的设计矩阵, $\text{rank}(X) = p$, 若协方差阵 V 已知时, 应用线性模型的一般理论, 我们很容易求得此时未知参数 β 的最佳线性无偏估计 (The Best Linear Unbiased Estimate, BLUE) 为

$$\beta^* = (X'V^{-1}X)^{-1}X'V^{-1}y.$$

但在实际应用中 V 一般是未知的, 此时我们可以选择使用 β 的最小二乘估计 (The Least Squares Estimate, LSE) $\hat{\beta} = (X'X)^{-1}X'y$, 由于其没有利用误差协方差信息, 所以经常会有估计精度上的损失. 另一种经常选用的估计是 β 的两步估计 (Two-stage Estimate, TSE) $\tilde{\beta}$, 即首先计算 V 的估计 \hat{V} , 将其代入 β^* 的计算式中得到的 β 的估计量, 即 $\tilde{\beta} = (X'\hat{V}^{-1}X)^{-1}X'\hat{V}^{-1}y$. 由于两步估计往往是观测向量的很复杂的非线性函数, 因此它的统计性质的研究难度很大, 关于这方面的更多内容可参看文献 [1]. Toyooka 等 [2], 王松桂等 [3,4] 在协方差阵的某些特殊结构下, 研究了两步估计的均方误差阵的上界及与此有关的若干性质. Rao^[5] 发现如果能够从历史数据或其它方面得到模型误差协方差阵的一个独立估计, 则可得到两步估计的协方差阵的精确表达式. 假设从历史数据计算出来, 或是从重复观测中计算得到一个与 y 独立的 V 的估计 S/m , 且

$$S \sim W_n(m, V), \quad (1.2)$$

这里 $W_n(m, V)$ 表示自由度为 m , 参数为 V 的 n 维 wishart 分布. 例如, 若已经从重复观测中得到

$$y_i = X_0\beta_0 + e_i, \quad e_i \sim N_n(0, V), \quad i = 0, 1, 2, \dots, m, \quad (1.3)$$

* 国家自然科学基金 (10271010) 和北京市自然科学基金 (1032001) 基金资助课题.

本文 2004 年 3 月 22 日收到, 2005 年 12 月 28 日收到修改稿.

这里, y_i 为 $n \times 1$ 的观测向量, X_0 为 $n \times p$ 的设计矩阵, β_0 为 $p \times 1$ 的回归参数, e_i 为 $n \times 1$ 随机误差, 则从 (1.3) 可得到形如 (1.2) 的估计 S .

对模型 (1.1), 记 $T_1 = \hat{\beta} = (X'X)^{-1}X'y$, 即 T_1 为 β 的 LSE, 显然它是 β 的一个无偏估计. 设 Z 为 $n \times (n-p)$ 矩阵, 且 $Z'X = 0$, $Z'Z = I_{n-p}$. 记 $T_2 = Z'y$, 则视 T_2 为协变量, Rao^[5] 得到 β^* 就是利用 T_2 而对 $\hat{\beta}$ 的协方差改进估计, 因此我们可以把 $\tilde{\beta}$ 看做是 $\hat{\beta}$ 的协方差改进两步估计. 下面这个引理是 Rao^[5] 得到的, 我们不加证明地给出

引理 1.1 对于线性模型 (1.1), 假设 (1.2) 成立, 则

$$\text{Cov}(\tilde{\beta}) = \frac{m-1}{m-(n-p)-1} \text{Cov}(\beta^*). \quad (1.4)$$

这个定理刻划了由于用 S 代替未知协方差阵 V 后在估计精度上所产生的损失. $n-p$ 是做两步估计时选用的协变量的个数, 从 (1.4) 式我们看到若选用的协变量个数较多, 但其中有些与 $\hat{\beta}$ 相关程度不高, 会导致 $\text{Cov}(\tilde{\beta})$ 的膨胀, 估计的精度会下降. 因此当 V 未知时, 选取与 $\hat{\beta}$ 相关程度较高的适当个数的协变量 (以下简称主协变量) 是十分重要的.

本文探讨了如何选择协变量的问题, 提出了协变量选择的典则相关检验法, 给出了检验的近似分布, 并通过模拟进一步显示了协变量选择的优良性及典则相关检验法的优良性. 在 §2 我们给出了典则相关检验法及检验的近似分布, 在 §3 我们选用了在经济、生物和医药等方面常见的三种误差协方差阵的假定下进行了模拟, 取得了良好的性质.

§2. 协变量选择的典则相关检验法

通过上面的讨论我们看到, 减少所使用的与 T_1 不相关或相关程度很小的协变量 (我们称它为次协变量) 的个数会提高两步估计 $\tilde{\beta}$ 的精度. 因此当 V 未知时, 选取与 $\hat{\beta}$ 相关程度较高的适当个数的主协变量是十分重要的. 在生长曲线模型中, Fujikoshi 和 Rao^[6] 提出了协变量选择的似然比法, 王松桂等^[7] 提出了 Two-way 选择法等等. 而对于一般的线性模型, 变量的选择方法很多, 但我们还没有见到从协变量角度考虑变量的选择问题.

若把 Z 分块为 $(Z_1 : Z_2)$, 其中 Z_1 是 $n \times q$ 阵, Z_2 是 $n \times r$ 阵, 且 $q+r = n-p$. 则 $T_2 = Z'y = (Z_1 : Z_2)'y \triangleq (T_{21}' : T_{22}')'$. 则

$$\begin{aligned} \text{Cov} \begin{pmatrix} \hat{\beta} \\ T_2 \end{pmatrix} &= \text{Cov} \begin{pmatrix} \hat{\beta} \\ T_{21} \\ T_{22} \end{pmatrix} = \begin{pmatrix} \Delta_{11} & \Delta_{12} \\ \Delta_{21} & \Delta_{22} \end{pmatrix} \\ &= \begin{pmatrix} (X'X)^{-1}X'VX(X'X)^{-1} & (X'X)^{-1}X'VZ_1 & (X'X)^{-1}X'VZ_2 \\ Z_1'VX(X'X)^{-1} & Z_1'VZ_1 & Z_1'VZ_2 \\ Z_2'VX(X'X)^{-1} & Z_2'VZ_1 & Z_2'VZ_2 \end{pmatrix} \\ &\triangleq \begin{pmatrix} \Delta_{11} & \Delta_{1t_1} & \Delta_{1t_2} \\ \Delta_{t_1 1} & \Delta_{t_1 t_1} & \Delta_{t_1 t_2} \\ \Delta_{t_2 1} & \Delta_{t_2 t_1} & \Delta_{t_2 t_2} \end{pmatrix}. \end{aligned} \quad (2.1)$$

若 $\Delta_{12} \neq 0$, 如果视 T_2 为协变量, Rao^[5] 得到 β^* 就是利用 T_2 对 $\hat{\beta}$ 的协方差改进估计, 即

$$\beta^* = \hat{\beta} - \Delta_{12}\Delta_{22}^{-1}T_2 = \hat{\beta} - (X'X)^{-1}X'VZ(Z'VZ)^{-1}Z'y, \quad (2.2)$$

因此我们可以把 $\tilde{\beta}$ 看做是 $\hat{\beta}$ 的协方差改进两步估计. 假设用第一部分协变量 T_{21} (我们称其为主协变量) 来改进 LSE, 易得基于 $\hat{\beta}$ 和 T_{21} 的协方差改进估计为

$$\hat{\beta}_A = \hat{\beta} - \Delta_{1t_1}\Delta_{t_1t_1}^{-1}T_{21} = \hat{\beta} - (X'X)^{-1}X'VZ_1(Z_1'VZ_1)^{-1}Z_1'y. \quad (2.3)$$

其协方差改进两步估计为

$$\tilde{\beta}_A = \hat{\beta} - (X'X)^{-1}X'SZ_1(Z_1'SZ_1)^{-1}Z_1'y. \quad (2.4)$$

与 $\hat{\beta}$ 不相关或相关程度很小的第二部分协变量 T_{22} 可以去除的话 (我们称其为次协变量), 则 $\hat{\beta}_A$ 与 β^* 应偏离不大. 经过简单计算可以得到

$$\begin{aligned} \hat{\beta}_A - \beta^* &= \Delta_{12}\Delta_{22}^{-1}T_2 - \Delta_{1t_1}\Delta_{t_1t_1}^{-1}T_{21} \\ &= (X'X)^{-1}X'V \left[(Z_1 : Z_2) \begin{pmatrix} Z_1'VZ_1 & Z_1'VZ_2 \\ Z_2'VZ_1 & Z_2'VZ_2 \end{pmatrix}^{-1} \begin{pmatrix} Z_1' \\ Z_2' \end{pmatrix} - Z_1(Z_1'VZ_1)^{-1}Z_1' \right] y \\ &= (X'X)^{-1}X'V(Z_2 - Z_1(Z_1'VZ_1)^{-1}Z_1'VZ_2)\Delta_{t_2t_2.t_1}^{-1}(Z_2'y - Z_2'VZ_1(Z_1'VZ_1)^{-1}Z_1'y) \\ &= (\Delta_{1t_2} - \Delta_{1t_1}\Delta_{t_1t_1}^{-1}\Delta_{t_1t_2})\Delta_{t_2t_2.t_1}^{-1}(T_{22} - \Delta_{t_2t_1}\Delta_{t_1t_1}^{-1}T_{21}) \\ &= \Delta_{1t_2.t_1}\Delta_{t_2t_2.t_1}^{-1}(T_{22} - \Delta_{t_2t_1}\Delta_{t_1t_1}^{-1}T_{21}). \end{aligned} \quad (2.5)$$

记 $T_{22} - \Delta_{t_2t_1}\Delta_{t_1t_1}^{-1}T_{21} \triangleq M$, 它可以看做是利用主协变量 T_{21} 后对次协变量 T_{22} 的改变, 易验证 M 与 T_{21} 相互独立, 而 $\Delta_{1t_2.t_1} = \Delta_{1t_2} - \Delta_{1t_1}\Delta_{t_1t_1}^{-1}\Delta_{t_1t_2}$ 是 $\hat{\beta}$ 与 M 的协方差阵, $\Delta_{t_2t_2.t_1}$ 是 M 的协方差阵, 即

$$\text{Cov} \begin{pmatrix} \hat{\beta} \\ M \end{pmatrix} = \begin{pmatrix} \Delta_{11} & \Delta_{1t_2.t_1} \\ (\Delta_{1t_2.t_1})' & \Delta_{t_2t_2.t_1} \end{pmatrix}.$$

从 (2.5) 可以看出, 若 $\Delta_{1t_2.t_1}$ 很小接近于 0, 则 $\hat{\beta}_A$ 与 β^* 就非常接近.

从 (2.5) 还可以看到, 考查第二部分协变量 T_{22} 是否可以被去除, 还可以从 M 的大小 (如矩阵范数) 来衡量, 若此指标很小, 接近于 0, 则第二部分协变量 T_{22} 就可以看做次协变量被去除, 否则, 就不能轻易去除. 在影响分析中, Cook^[8] 距离是经常采用的度量工具, 为了衡量估计量 $\hat{\beta}_A$ 与 β^* 的距离, 我们在此也提出一种类似的距离, 我们定义的统计量为

$$\begin{aligned} d_{Z_2} &= E\|\Delta_{11}^{-1/2}(\hat{\beta}_A - \beta^*)\|^2 = E[(\Delta_{11}^{-1/2}(\hat{\beta}_A - \beta^*))'(\Delta_{11}^{-1/2}(\hat{\beta}_A - \beta^*))] \\ &= E\text{tr}[(\Delta_{11}^{-1/2}(\hat{\beta}_A - \beta^*))'(\Delta_{11}^{-1/2}(\hat{\beta}_A - \beta^*))] \\ &= \text{tr}E[\Delta_{11}^{-1/2}\Delta_{1t_2.t_1}\Delta_{t_2t_2.t_1}^{-1}MM'(\Delta_{t_2t_2.t_1}^{-1})'\Delta_{1t_2.t_1}'\Delta_{11}^{-1/2}] \\ &= \text{tr}(\Delta_{11}^{-1/2}\Delta_{1t_2.t_1}\Delta_{t_2t_2.t_1}^{-1}\Delta_{1t_2.t_1}'\Delta_{11}^{-1/2}) = \sum_{i=1}^p \rho_i^2, \end{aligned} \quad (2.6)$$

其中 $\rho_1 \geq \rho_2 \geq \cdots \geq \rho_p$ 是 $\hat{\beta}$ 和 M 的典则相关系数, 即 ρ_i^2 为方程

$$|\Delta_{11}^{-1/2}\Delta_{1t_2.t_1}\Delta_{t_2t_2.t_1}^{-1}(\Delta_{1t_2.t_1})'\Delta_{11}^{-1/2} - \rho_i^2 I| = 0$$

的根, 记 r_i 为对应的样本典则相关系数. 也就是说, ρ_i^2 是排除主协变量 T_{21} 的影响后, $\hat{\beta}$ 与 T_{22} 的条件典则相关系数平方, 它刻画了在排除主协变量 T_{21} 的影响后 $\hat{\beta}$ 与 T_{22} 线性相关的程度, 于是它的值应该比较小并接近于 0. 从 (2.6) 看到, 若所有的 ρ_i 都很小接近于 0, 则第二部分协变量 T_{22} 就可以看做次协变量被去除. Bartlett (1939)^[9] 提出了此假设 $H_0: \Delta_{1t_2, t_1} = 0$ ($r_1 = \cdots = r_p = 0$) 可以用近似 χ^2 分布来检验, 对给定的显著性水平 α , 当下式满足时拒绝 H_0 ,

$$-\left(m - \frac{1}{2}(n+1)\right) \ln \prod_{i=1}^p (1 - r_i^2) > \chi_{p(n-p)}^2(\alpha).$$

其中 $\chi_{p(n-p)}^2(\alpha)$ 是自由度为 $p(n-p)$ 的 χ^2 的上 α 分位点.

注 我们还可以从另一个角度理解这个结论, 与 $\hat{\beta}$ 不相关或相关程度很小的次协变量 T_{22} 可以去除的话, 则 $\hat{\beta}$ 和 T_2 的协方差阵与 $\hat{\beta}$ 和 T_{21} 的协方差阵相差不大, 可以通过检验

$$H_0: \text{tr}(U_{11}^{-1}U_{12}U_{22}^{-1}U_{21}) = \text{tr}(U_{11}^{-1}U_{1t_1}U_{t_1t_1}^{-1}U_{t_11}) \quad (2.7)$$

来衡量, 其中 U_{ij} 就是在 (2.1) 中用 S 代替后对应的形式. 若 H_0 接近成立, 则次协变量 T_{22} 可以去除. 若以很强的根据否定 H_0 , 则不能有充足的理由去除 T_{22} . 经过一系列计算得到 H_0 成立等价于 $U_{1t_2, t_1} = U_{1t_2} - U_{1t_1}U_{t_1t_1}^{-1}U_{t_1t_2} = 0$.

§ 3. 随机模拟

对线性模型 $y = X\beta + e$, 其中 $e \sim N_6(0, V_i)$, $i = 1, 2, 3$, 在经济、生物和医药等方面较常见的三种误差协方差阵为 (参见 [10])

I: V_1 是具有均匀误差协方差阵, 例如 $V_1 = 2I_6 + 5J_6$, 其中, J_n 表示所有元素都是 1 的 n 阶方阵;

II: V_2 具有指数相关协方差阵, 例如 $V_2 = (\sigma_{ij})$, 其中 $\sigma_{ij} = 5 + 2e^{-|i-j|}$;

III: V_3 具有一阶自回归误差协方差阵, 例如我们选取误差协方差阵具有如 V_3 的结构,

$$V_3 = \sigma^2 \begin{pmatrix} 1 & \theta & \theta^2 & \cdots & \theta^5 \\ \theta & 1 & \theta & \cdots & \theta^4 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \theta^5 & \theta^4 & \theta^3 & \cdots & 1 \end{pmatrix},$$

其中, 取 $\theta = 0.9$.

我们在模拟中选取模型设计阵 X 为

$$X = \begin{bmatrix} 4.2 & 108.1 & 15.9 \\ 4.1 & 114.8 & 16.4 \\ 3.1 & 123.2 & 19.0 \\ 3.1 & 126.9 & 19.1 \\ 1.1 & 132.1 & 18.8 \\ 2.2 & 137.7 & 20.4 \end{bmatrix},$$

β 的真值设定为 $(3, 5, 7)'$, 在上面三种已知误差协方差阵下, 模拟研究了所有协变量下的未知参数的两步估计 $\tilde{\beta}$, 主协变量选择下的两步估计 $\tilde{\beta}_A$. 模拟结果进一步说明了在适当选择协变量的条件下得到的协方差改进两步估计 $\tilde{\beta}_A$ 的精度会提高.

对每一种设定的协方差阵 V_i , 模拟重复次数为 1000. 计算这 1000 次模拟中各估计量的均值做为参数的估计值, 如下表.

表 1 参数的各种估计比较

m 值	模型 参数	I			II			III		
		β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3
8	$\tilde{\beta}$	3.0293	5.0142	6.9057	3.0303	5.0130	6.9134	2.9993	5.0026	6.9844
	$\tilde{\beta}_A$	3.0107	5.0103	6.9328	3.0033	5.0077	6.9494	2.9928	5.0016	6.9909
10	$\tilde{\beta}$	2.9965	4.9883	7.0759	2.9970	4.9919	7.0537	3.0054	4.9982	7.0114
	$\tilde{\beta}_A$	2.9902	4.9883	7.0796	2.9902	4.9928	7.0496	3.0008	4.9982	7.0121
20	$\tilde{\beta}$	2.9789	4.9902	7.0728	2.9848	4.9932	7.0516	3.0009	4.9990	7.0084
	$\tilde{\beta}_A$	2.9868	4.9938	7.0460	2.9836	4.9938	7.0448	2.9965	4.9986	7.0106
40	$\tilde{\beta}$	3.0084	5.0084	6.9364	3.0094	5.0051	6.9581	3.0034	5.0008	6.9916
	$\tilde{\beta}_A$	3.0137	5.0074	6.9427	3.0207	5.0055	6.9536	3.0042	5.0008	6.9911

我们看到在三种假定的误差协方差阵下, 两种估计的精度都比较高, 但多数情况下 $\tilde{\beta}_A$ 比 $\tilde{\beta}$ 要好, 为了更清楚地比较这两种估计, 我们分别计算其均方误差 ($1000 \times \text{MSE}$), 如图 1 所示, 图中实线为 $\tilde{\beta}$ 的 MSE, 点线为 $\tilde{\beta}_A$ 的 MSE. 其中均方误差的计算为 (若 $\hat{\beta}$ 为 β 的一种估计)

$$\text{MSE}(\hat{\beta}) = \frac{1}{1000} \sum_{i=1}^{1000} (\hat{\beta}_i - \beta)'(\hat{\beta}_i - \beta).$$

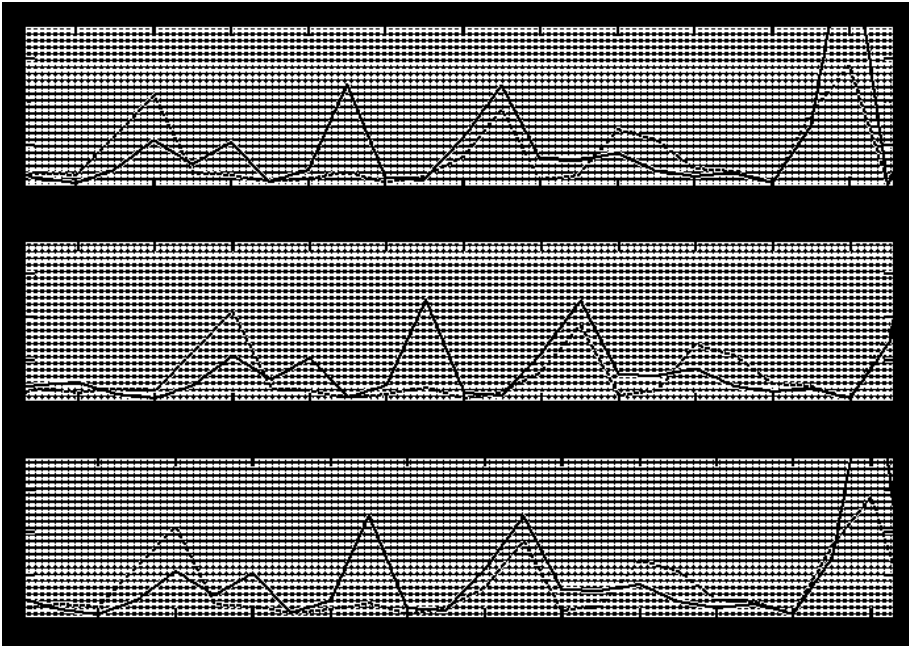


图 1 均方误差图

通过模拟我们看到, 适当选择协变量的得到的协方差改进两步估计 $\tilde{\beta}_A$ 的精度确实会提高. 从均方误差图看到, 在一般情况下, 协方差改进两步估计 $\tilde{\beta}_A$ 的均方误差更小并且与最佳线性无偏估计 β^* 的较接近, 这和先验信息量 m 没有直接的线性关系. 从图 1 中我们还清楚地看到, 即使 m 值不大的情况下, 协方差改进两步估计也有很好性质.

参 考 文 献

- [1] 陈希孺, 王松桂, 线性模型中的最小二乘法, 上海科技出版社, 上海, 2003.
- [2] Toyooka, Y. and Kariya, T., An approach to upper bound problems for risks generalized least squares estimator, *The Ann. Statist.*, **14**(1986), 679–685.
- [3] 王松桂, 刘爱义, 两步估计的效率, *数学学报*, **32**(1989), 42–50.
- [4] 王松桂, 范永辉, Panel 模型中两步估计的优良性, *应用概率统计*, **14**(2)(1998), 177–184.
- [5] Rao, C.R., Least squares theory using an estimated dispersion matrix and its application to measurement of signals, In *Proceedings of the Fifth Berkeley Symposium on Math. Statist & Prob.* (Eds. by Lecam, J. and Neyman, J.), Vol. 1, 1967, 355.
- [6] Fujikoshi, Y. and Rao, C.R., Selection of covariables in the growth curve model, *Biometrika*, **78**(1991), 779–785.
- [7] Wang, S.G. (王松桂), Liski, E.P., Nummi, T., Two-way selection of covariables in multivariate growth curve models, *Linear Algebra and its Applications*, **289**(1999), 333–342.
- [8] Cook, R.D., Detection of influential observation in linear regression, *Technometrics*, **19**(1977), 15–18.
- [9] Johnson, R.A., Wichern, D.W., *Applied Multivariate Statistical Analysis* (5th Ed), China Statistics Press, Beijing, 2003.
- [10] Diggle, P.J., Liang, K.Y. and Zeger, S.L., *Analysis of Longitudinal Data*, Oxford Science, New York, 2000.

The Variable Selection in Covariance Adjusted Estimates

YIN SUJU

WANG SONGGUI

(Beijing University of Technology, Beijing, 100022)

In this paper we study the problem about how to select the covariables. A new test approach by using the canonical correlation is proposed, the approximate distribution of the test statistics is given. A simulation comparison is made under three forms of covariance matrices which are adopted very often in the economy, biology and medicine. Our results show that the necessity of covariable selection and the good property of the canonical correlation test method.