应用概率统计　第十八卷
第三期　2002 年 8 月

Chinese Journal of Applied Probability
and Statistics Vol. 18 No. 3 Aug. 2002

协方差阵估计的比较

陈　茂　学
(山东农业大学农学院，泰安，271018)

摘　要

本文在一般线性回归模型误差异方差情况下，通过计算机模拟对回归系数最小二乘估计的协方差矩阵的估计进行了比较．结果表明，当样本大小大于 50 时，回归系数的最小二乘估计具有较高的估计精度；其协方差矩阵的五种估计以普通最小二乘估计的协方差矩阵为最优．

关 键 词：线性回归模型，协方差矩阵，异方差，模拟．
学 科 分 类 号：O212.1, O212.5.

§1. 引　　言

考虑一般线性回归模型

$$y = X\beta + \varepsilon, \qquad \mathsf{E}(\varepsilon) = 0, \qquad \mathsf{Var}(\varepsilon) = \Phi, \tag{1.1}$$

其中 y 为 $n \times p$ 的随机观测向量，X 为 $n \times p$ 的列满秩设计矩阵，β 为 $p \times 1$ 的未知回归系数，ε 为 $n \times 1$ 的随机误差向量，其均值为 0，协方差矩阵为 $\Phi = \mathrm{diag}(\varphi_{ii})$．众所周知，$\beta$ 的最小二乘估计为

$$\widehat{\beta} = (X'X)^{-1}X'y, \tag{1.2}$$

其协方差矩阵为

$$\mathsf{Var}(\widehat{\beta}) = (X'X)^{-1}X'\Phi X(X'X)^{-1}. \tag{1.3}$$

如果误差是等方差的，即 $\varphi_{ii} = \sigma^2$，$\Phi = \sigma^2 I$，则 $\widehat{\beta}$ 是 β 的最佳线性无偏估计 (见王松桂 (1987))，(1.3) 式变为

$$\mathsf{Var}(\widehat{\beta}) = \sigma^2(X'X)^{-1}. \tag{1.4}$$

在实际应用中，σ^2 往往未知，这时可用样本进行估计．σ^2 通常用 $\widehat{\sigma}^2 = \sum_{i=1}^{n} e_i^2/(n-p)$ 估计，它是无偏估计，这里 $e_i = y_i - x_i\widehat{\beta}$ 为残差，x_i 是 X 的第 i 行，从而得到普通最小二乘估计的协方差矩阵的估计

$$D_1 = \frac{\sum_{i=1}^{n} e_i^2}{n-p}(X'X)^{-1}. \tag{1.5}$$

用 D_1 可对回归系数进行假设检验．但是，当误差方差不相等时，(1.3) 式不等于 (1.4) 式，再用 D_1 对回归系数进行假设检验一般是不合适的．如果 Φ 已知，用 (1.3) 对回归系数进行检验．大多数情况下，这种异方差协方差矩阵的形式未知，许多统计学家提出了不同的估计方法．

一种基本的想法是用 e_i^2 估计 φ_{ii}，即 $\widehat{\varphi}_{ii} = e_i^2$，$\widehat{\Phi} = \mathrm{diag}(e_i^2)$．从而得异方差协方差矩阵 (1.3) 的估计

$$D_2 = (X'X)^{-1}X'\mathrm{diag}(e_i^2)X(X'X)^{-1}, \tag{1.6}$$

D_2 是最普通的一种形式，并称为 White 估计、Eicker 估计或 Huber 估计．White(1980) 证明了 D_2 是 $\mathsf{Var}(\widehat{\beta})$ 的相合估计．

本文 2001 年 12 月 4 日收到，2002 年 1 月 14 日收到修改稿．

为了改善 D_2 的小样本性质， Hinkley(1987) 提出了一种最简单的调整方法，将每个残差 e_i 都乘以与自由度有关的因子 $\sqrt{n/(n-p)}$, 从而得 D_3:

$$D_3 = \frac{n}{n-p}(X'X)^{-1}X'\mathrm{diag}(e_i^2)X(X'X)^{-1} = \frac{n}{n-p}D_2. \tag{1.7}$$

基于奇异值和影响点 (见陈希孺，王松桂 (1987)) 的考虑，定义 $h_{ii} = x_i(X'X)^{-1}x_i'$, 在等方差条件下，则 $\mathsf{Var}(e_i) = \sigma^2(1-h_{ii}) \neq \sigma^2$. 由于 $1/n \leq h_{ii} \leq 1$, $\mathsf{Var}(e_i)$ 低估了 σ^2. 基于这种思想，在异方差条件下， e_i^2 是 σ_i^2 的有偏估计， $e_i^2/(1-h_{ii})$ 可减少这种偏差，因此得 D_4 (Horn(1975)):

$$D_4 = (X'X)^{-1}X'\mathrm{diag}\Big(\frac{e_i^2}{1-h_{ii}}\Big)X(X'X)^{-1}. \tag{1.8}$$

为了进一步增大强影响点的误差方差的估计值，将 e_i^2 除以 $(1-h_{ii})^2$, 从而得 D_5 (MacKinnon(1985)):

$$D_5 = (X'X)^{-1}X'\mathrm{diag}\Big(\frac{e_i^2}{(1-h_{ii})^2}\Big)X(X'X)^{-1}. \tag{1.9}$$

在异方差存在且未知的情况下，以上四种估计都是最小二乘估计协方差矩阵的相合估计．目前大部分是采用 D_2, 但很少有人讨论它的小样本性质．使用 D_3 、 D_4 、 D_5 较少的原因可能是由于对其性质不了解或没有能直接使用的计算软件．究竟在怎样的误差结构下，采用哪一种形式的协方差矩阵，样本量应多大，回归系数的估计精度如何，都是使用者十分关心的问题．

本文的目的是在线性回归模型 (1.1) 下，研究不同误差结构下，最小二乘估计 $\widehat{\beta}$ 对 β 估计精度的影响；协方差矩阵 $\mathsf{Var}(\widehat{\beta})$ 的不同估计 D_1, D_2, D_3, D_4, D_5 对回归系数假设检验的影响，从中找出协方差矩阵 $\mathsf{Var}(\widehat{\beta})$ 的最优估计．我们采用模拟方法进行研究，在第二节中，将确定模型 (1.1) 的具体形式，包括自变量个数，回归系数，样本大小，误差结构，计算内容及计算方法．第三节将模拟计算结果给出直观的图形形式并进行分析．结果显示：在误差等方差条件下，当样本大小 $n \geq 20$ 和在误差异方差条件下 $n \geq 50$, 可得到精度较高的 β 的最小二乘估计，估计精度随误差方差的增大而降低，随样本量的增大而提高；在异方差条件下，专为估计协方差矩阵 $\mathsf{Var}(\widehat{\beta})$ 的四种估计 D_2, D_3, D_4, D_5 其优劣依次为 D_4, D_3, D_5, D_2; 无论是在误差等方差条件下，还是在误差异方差条件下，对 $\mathsf{Var}(\widehat{\beta})$ 的五种估计 D_1 是最优者．所以，当我们应用线性回归模型时，可以放心使用普通最小二乘估计而不必担心是否存在异方差性，这对实际应用带来极大的方便．但这也说明，在异方差条件下，协方差矩阵 $\mathsf{Var}(\widehat{\beta})$ 的四种估计 D_2, D_3, D_4, D_5 还不理想，需要进一步寻求更好的估计．

§2. 模拟方法及步骤

我们选择了含四个自变量的线性回归模型，确定了自变量的取值范围及六种不同的样本大小，选择了五种具有代表性的误差结构形式及两种常见分布正态分布和 t 分布．制定了合理的模拟步骤，对不同的组合确定了 10000 次重复，以提高模拟结果的可靠性．

2.1 模拟计算方法

在回归模型 (1.1) 中，我们取 4 个自变量 x_1, x_2, x_3, x_4, 回归系数 $\beta' = (2,1,2,3,4)$ 得回归模型

$$y_i = 2 + 1x_{i1} + 2x_{i2} + 3x_{i3} + 4x_{i4} + \varepsilon_i, \tag{2.1}$$

x_{ij} $(i = 1,2,\cdots,n;\ j = 1,2,3,4)$ 的取值范围确定在 $0 \sim 5$ 之间的均匀分布．ε_i 选取不同的误差结构，并满足 (1.1) 的假定条件 $\mathsf{E}(\varepsilon_i) = 0$, 选择了 5 种不同的误差结构 (误差函数) 于 (2.2).

$$\begin{cases} \mathrm{ST0}: & \varepsilon_i = \varepsilon_i^*, \\ \mathrm{ST1}: & \varepsilon_i = \sqrt{x_{i1}}\varepsilon_i^*, \\ \mathrm{ST2}: & \varepsilon_i = \sqrt{x_{i3}+1.6}\varepsilon_i^*, \\ \mathrm{ST3}: & \varepsilon_i = \sqrt{x_{i3}(x_{i4}+2.5)}\varepsilon_i^*, \\ \mathrm{ST4}: & \varepsilon_i = \sqrt{x_{i1}(x_{i2}+2.5)x_{i3}}\varepsilon_i^*. \end{cases} \tag{2.2}$$

其中 ε_i^* 是服从标准正态分布 $N(0,1)$ 或服从自由度为 5 的 t 分布 $t(5)$ 的随机变量．误差结构 ST0 是等方差结构，即不存在异方差性．异方差往往是由于自变量的取值不同所引起，因此误差结构 ST1 的异方差仅由自变量 x_1 的不同取值所致．其方差 $\mathrm{Var}(\varepsilon_i) = x_{i1}(\varepsilon_i^* \sim N(0,1))$ 或 $\mathrm{Var}(\varepsilon_i) = (5/3)x_{i1}(\varepsilon_i^* \sim t(5))$. 误差结构 ST2 的异方差是由自变量 x_3 的取值加上一个常数 1.6 所引起．同理， ST3 和 ST4 分别与 2 个自变量 x_3, x_4 和 3 个自变量 x_1, x_2, x_3 的联合作用有关．误差方差最大的是 ST4, 最小的是 ST0. 所以选取的这些误差结构具有一定的代表性 (见 Long(2000)). 对不同的误差结构，不同的样本量 $n = 20, 30, 50, 100, 200$ 和 500 进行模拟．

2.2 模拟步骤

引进记号 $\varepsilon_i^*(d)$, $d = 1, 2$ 分别表示 ε_i^* 服从正态分布 $N(0,1)$ 和 t 分布 $t(5)$; $\mathrm{ST}(s)$, $s = 0, 1, 2, 3, 4$ 分别表示误差结构 ST0, ST1, ST2, ST3 和 ST4; $n(m)$, $m = 1, 2, 3, 4, 5, 6$ 分别表示 $n = 20, 30, 50, 100, 200$ 和 500. 具体模拟步骤如下：

(1) 取 $d = 1$, $s = 1$, $m = 1$;

(2) 由 $\varepsilon_i^*(d)$ 确定 ε_i^* 的分布;

(3) 由 $\mathrm{ST}(s)$ 确定误差结构 ε_i;

(4) 由 $n(m)$ 确定样本量 n;

(5) 对固定的 $\varepsilon_i^*(d)$, $\mathrm{ST}(s)$, $n(m)$, 随机产生 (见高慧璇 (1994))ε_i^* 和 x_{ij}, 并计算 ε_i 和协方差矩阵 $\Phi = \mathrm{diag}(\mathrm{Var}(\varepsilon_i))$, 由 (2.1) 式计算 y_i. 从而形成设计矩阵 X 和随机观测向量 y. 由 (1.2) 式计算回归系数 $\widehat{\beta}$, 由 (1.3) 式计算协方差矩阵 $\mathrm{Var}(\widehat{\beta})$, 由 (1.5)~(1.9) 式计算协方差矩阵 D_1, D_2, D_3, D_4, D_5. 记模型 (2.1) 的系数向量为 $b' = (2, 1, 2, 3, 4)$, 用不同的协方差矩阵 D_1, D_2, D_3, D_4, D_5 分别对回归系数进行假设检验， $H_{01} : \beta_k = b_k$ 和 $H_{02} : \beta_k = 0$, $k = 0, 1, 2, 3, 4$, 其检验统计量为 t 统计量，显著性水平为 0.05.

(6) 重复 (5)10000 次，计算 β 与 $\widehat{\beta}$ 平均值的差，反映了对回归系数的估计精度；计算 10000 次中拒绝 H_{01} 和 H_{02} 的百分率及平均百分率；计算 $\mathrm{Var}(\widehat{\beta})$ 以及 D_1, D_2, D_3, D_4, D_5 的平均值及其迹；计算 $\mathrm{Var}(\widehat{\beta})$ 的平均值与各 D_1, D_2, D_3, D_4, D_5 的平均值之差的平方的迹，它反映了对协方差矩阵的估计精度．

(7) 置 $m = m + 1$, 如果 $m > 6$, 则置 $m = 1$ 并转到 (8); 否则，返回到 (4);

(8) 置 $s = s + 1$, 如果 $s > 5$, 则置 $s = 1$ 并转到 (9); 否则，返回到 (3);

(9) 置 $d = d + 1$, 如果 $d > 2$, 则模拟结束；否则，返回到 (2).

§3. 模拟结果与分析

本节主要对以上模拟结果绘图分析，分析不同误差结构对回归系数估计精度的影响；不同协方差矩阵估计 D_1, D_2, D_3, D_4, D_5 对回归系数检验结果的影响；分析不同协方差矩阵 D_1, D_2, D_3, D_4, D_5 的估计误差，为实际应用选择最优的协方差矩阵提供依据．

3.1 回归系数的估计精度比较

在固定的 ε_i^* 情况下，对不同的误差结构，各计算 10000 个回归系数的平均值与 β 的差，分析在不同条件下回归系数的估计精度．图 1 给出了在 ε_i^* 服从正态分布 $N(0,1)$ 的情况下，误差结构分别为 ST0, ST1, ST2, ST3 的回归系数的估计精度．

由图 1 可以看出，在误差结构为 ST0, 即等方差的情况下，回归系数的估计精度最高；在误差结构为 ST3, 即异方差存在且方差较大的情况下，回归系数的估计精度最低；随着样本量的增大，估计精度随之提高，但提高的速度随着方差的增大而降低．如在误差结构 ST0 和 ST1 下，只要 $n \geq 100$, 则估计误差近于 0; 而在误差结构 ST2 和 ST3 下， $n \geq 500$ 才能使估计误差近于 0, 对 ST4 有类似的结果．总之，在误差方差较小或样本量较大 ($n \geq 50$) 的情况下，回归系数的估计精度都会提高．

3.2 不同协方差矩阵检验 $H_{01} : \beta_k = b_k$ 的差异比较

在不同的误差结构下，用不同的协方差矩阵 D_1, D_2, D_3, D_4, D_5 对回归系数进行检验 $H_{01} : \beta_k = b_k$, $k = 0, 1, 2, 3, 4$, 显著性水平为 0.05, 计算 10000 次重复拒绝 H_{01} 的百分率，这个百分数越接近于 0.05, 对应的协方差矩阵就越好．任取部分结果绘于图 2. 图 2 列出了在 $\varepsilon_i^* \sim N(0,1)$ 的情况下，取误差结构为 ST0, ST1,

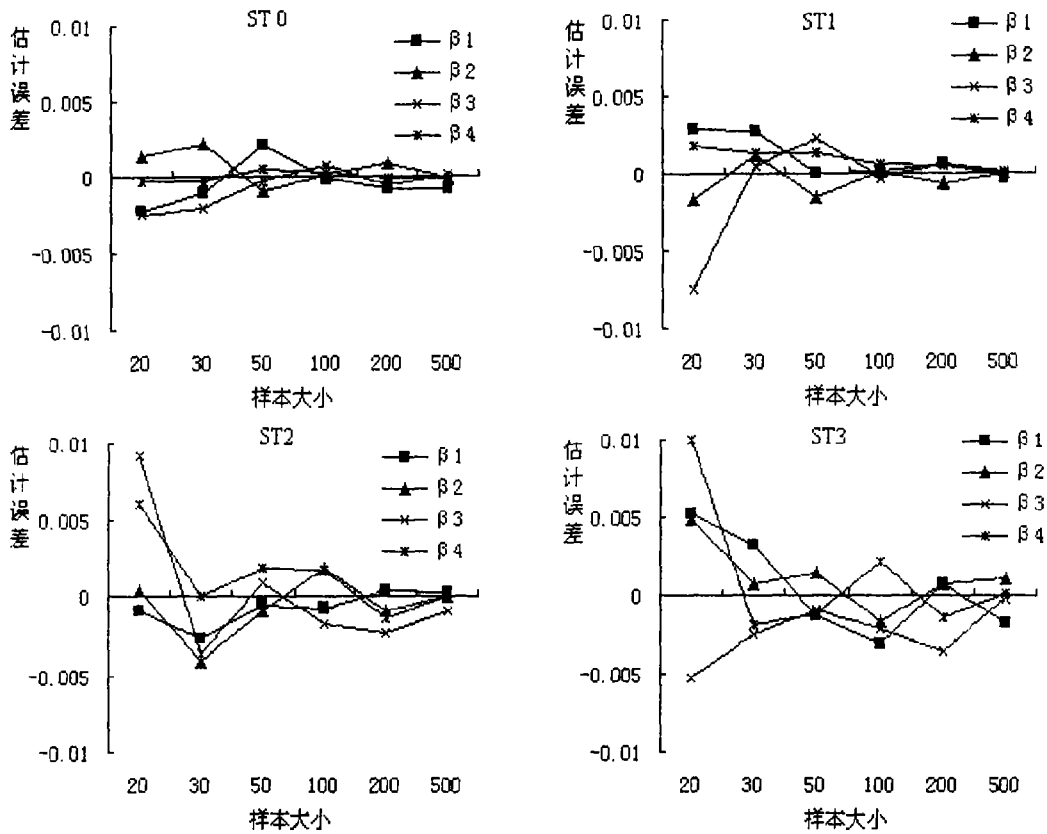

图 1 误差结构为 ST0, ST1, ST2, ST3, 回归系数的估计误差

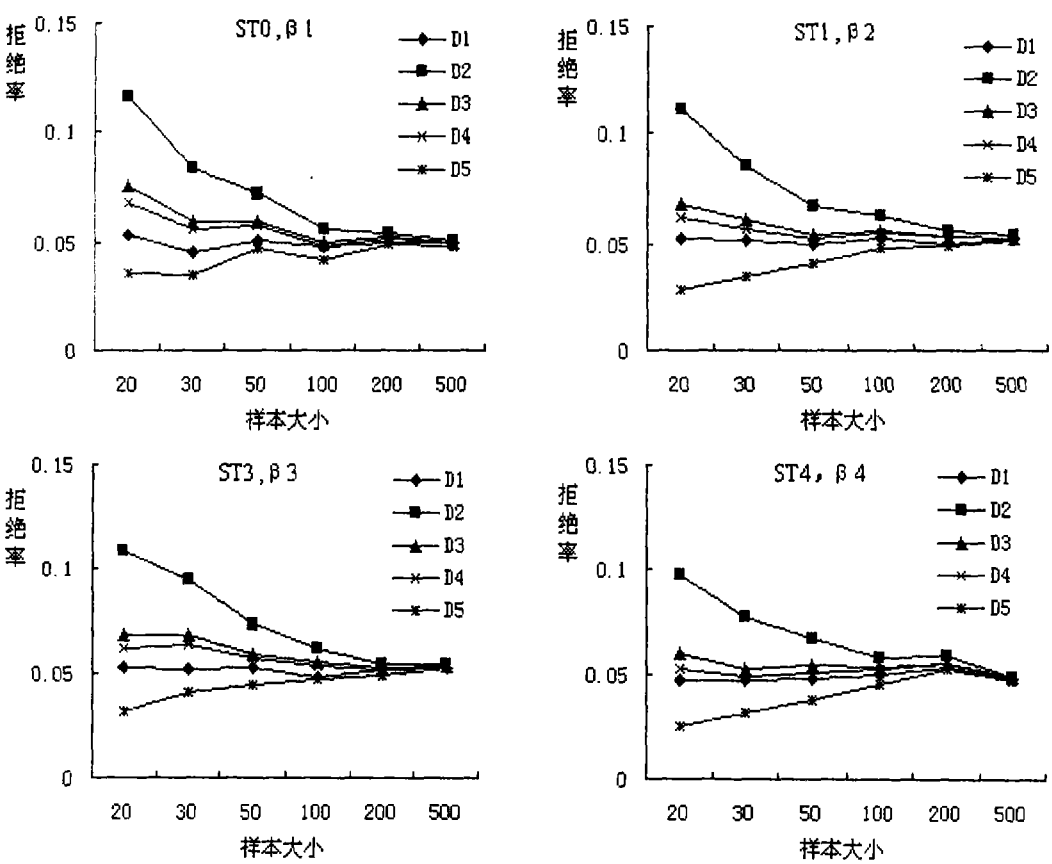

图 2 不同误差结构下拒绝 $H_{01}:\beta_k=b_k$ 的百分率

ST3, ST4, 分别用协方差矩阵 D_1, D_2, D_3, D_4, D_5 检验 $H_{01}:\beta_1=1$, $H_{01}:\beta_2=2$, $H_{01}:\beta_3=3$ 及 $H_{01}:\beta_4=4$ 的结果. 由图 2 看出: 无论是误差等方差 (ST0) 还是误差异方差 (ST1, ST3, ST4) 情况下, D_1 是最优的, D_2 最差; 当样本量 $n\geq 200$ 时, 它们都趋于一致 0.05; 在误差异方差情况下, 五种不同的协方差矩阵的优劣依次为 D_1, D_4, D_3, D_5, D_2; 随着误差方差的增大, 从 ST0 变化到 ST4, D_4 和 D_3 逐渐显示出其优越性, 即趋于 0.05. 图 2 仅给出了少数几种情况下的结果, 对其余情况的结果类似. 图 3 给出了拒绝 $H_{01}:\beta_k=b_k$, $k=0,1,2,3,4$ 的平均百分率, 结果与图 2 类似, 但趋势更明显.

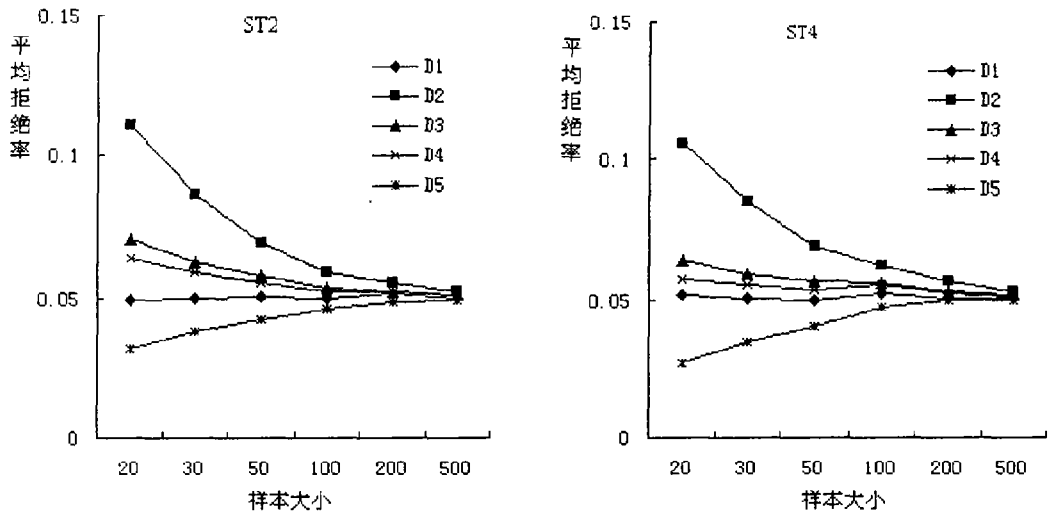

图 3　不同误差结构下拒绝 $H_{01}:\beta_k=b_k$ 的平均百分率

3.3　不同协方差矩阵检验 $H_{02}:\beta_k=0$ 的差异比较

在不同的误差结构下, 用不同的协方差矩阵 D_1, D_2, D_3, D_4, D_5 对回归系数进行检验 $H_{02}:\beta_k=0$, $k=0,1,2,3,4$, 显著性水平为 0.05, 计算 10000 次重复拒绝 H_{02} 的百分率, 即功效 (Power), 功效越大, 对应的协方差矩阵应该说越好. 但是它与 H_{01} 的检验是一对矛盾, 应是协调的关系, 过大或过小都不好. 取部分结果绘于图 4, 其它结果与此类似. 由图 4 看出: 当 $n\geq 100$ 时, 不同协方差矩阵下的检验功效都趋于 1; 当 $n\leq 50$ 时, D_1, D_3, D_4 基本一致, 且界于 D_2 和 D_5 之间, 所以 D_1, D_3, D_4 不失是好估计.

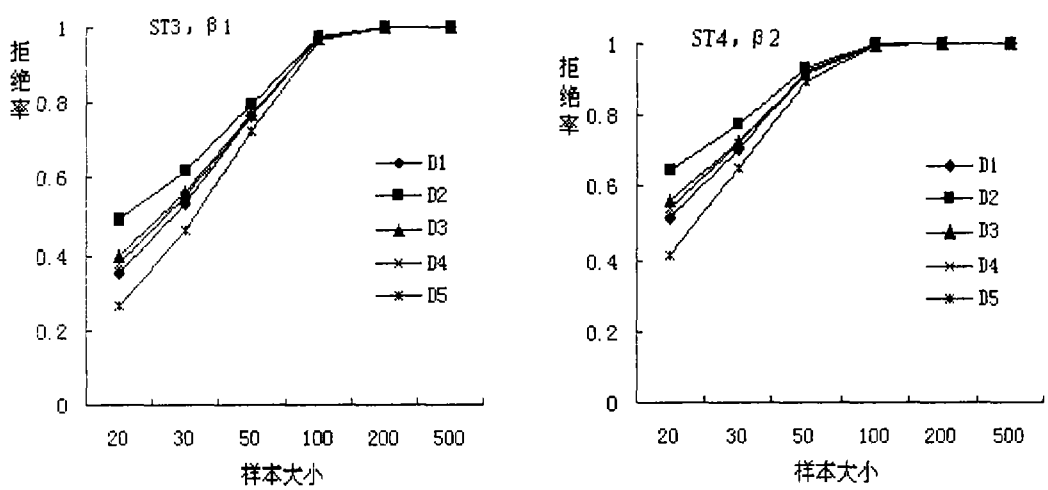

图 4　不同误差结构下拒绝 $H_{02}:\beta_k=0$ 的百分率 (功效)

3.4　不同协方差矩阵的迹及其估计精度比较

由 (1.3), (1.5)~(1.9) 式计算各协方差矩阵 $\mathrm{Var}(\hat{\beta})$, D_1, D_2, D_3, D_4, D_5 的平均值及其迹; 计算 $\mathrm{Var}(\hat{\beta})$ 的平均值与各 D_1, D_2, D_3, D_4, D_5 的平均值之差的平方的迹, 它反映了协方差矩阵的估计精度. 图 5 给出了不

同误差结构下D_1, D_2, D_3, D_4, D_5 的平均值的迹.

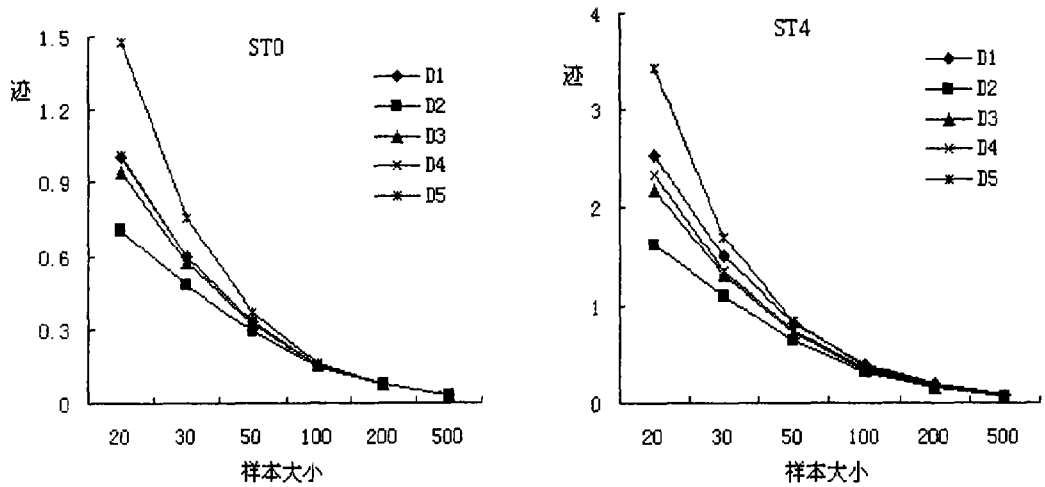

图 5 不同协方差矩阵 D_1, D_2, D_3, D_4, D_5 平均值的迹

图 5 显示：各协方差矩阵迹的大小依次为 D_5, D_1, D_4, D_3, D_2. 其中 D_1, D_4, D_3 三者基本相同，且与 D_2, D_5 差异较大，这与检验 H_{01} 的结果是一致的，方差越大，拒绝 H_{01} 的百分率越小. 当样本量 $n \geq 100$ 时，各协方差矩阵的迹趋于一致，且逐渐减小.

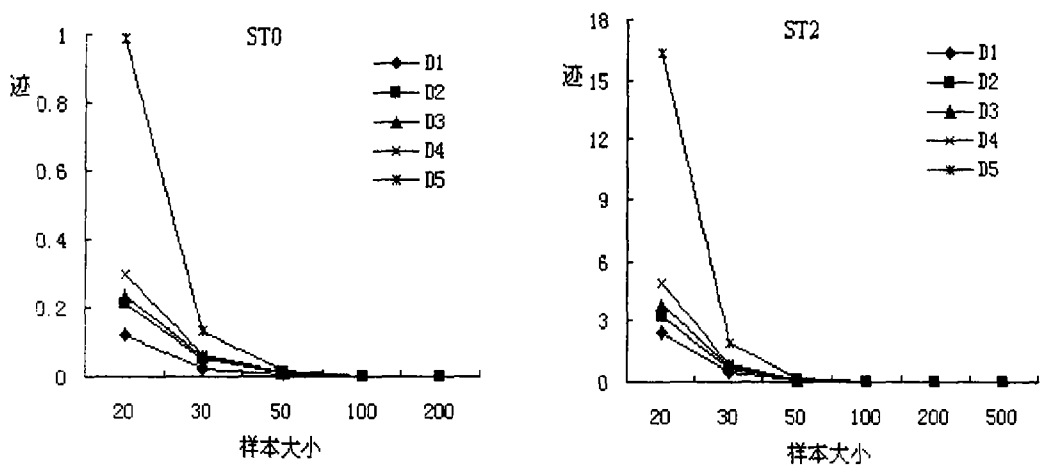

图 6 $\mathrm{Var}(\widehat{\beta})$ 的平均值与 D_1, D_2, D_3, D_4, D_5 的平均值之差的平方的迹

由图 6 看出：无论在等方差或异方差下， D_1 的估计精度最高， D_5 的精度最低. 估计误差的大小随着误差方差的增大而增大，当 $n \geq 50$ 时，它们的估计误差都趋于 0.

§4. 讨论与结论

以上是在 $\varepsilon_i^* \sim N(0,1)$ 条件下的模拟结果，对 $\varepsilon_i^* \sim t(5)$ 的模拟结果与其完全类似. 在模型 (2.1) 中，回归系数分别为 2,1,2,3,4, 所以自变量 x_j, $j = 1,2,3,4$ 对 y 的作用是不均衡的，因此再取均衡模型

$$y_i = 1 + 1x_{i1} + 1x_{i2} + 1x_{i3} + 1x_{i4} + \varepsilon_i \tag{4.1}$$

进行模拟. 其模拟结果与对 (2.1) 的模拟结果也完全类似. 考虑到 x_{ij} 取值范围的不同，是否对模拟结果有影响，再选取 $x_{ij} \in [0,2]$ 进行模拟，其结果与 $x_{ij} \in [0,5]$ 的模拟结果也完全类似. 对两种不同的分布，两种不

同的模型，两种不同的取值范围和五种不同的误差结构，共 40 种组合，各进行了 10000 次模拟，所以模拟结果具有一定的可靠性．

这种模拟对理论研究也具有一定的指导作用．如在误差异方差条件下，对协方差矩阵 $\mathrm{Var}(\widehat{\beta})$ 的估计有 D_2、D_3、D_4、D_5．其中 D_3，D_4，D_5 都是在 D_2 的基础上改进获得的．事实上，D_3，D_4，D_5 是将 D_2 中 e_i^2 分别乘以一个膨胀因子 $n/(n-k)$，$1/(1-h_{ii})$ 及 $1/(1-h_{ii})^2$ 得到的．从模拟结果看出，膨胀因子 $n/(n-k)$，$1/(1-h_{ii})$ 偏小，而 $1/(1-h_{ii})^2$ 则偏大．所以，对寻找界于 $1/(1-h_{ii})$ 与 $1/(1-h_{ii})^2$ 之间的一个膨胀因子，得到 $\mathrm{Var}(\widehat{\beta})$ 更好的估计具有指导意义．

从我们的模拟结果可以得出如下主要结论：

1. 在等方差条件下当样本大小 $n \geq 20$ 和在异方差条件下样本大小 $n \geq 50$ 时，可得到精度较高的 β 的最小二乘估计．且估计误差随样本量的增大而降低，随误差方差的增大而上升．

2. 在异方差条件下，专为估计协方差矩阵 $\mathrm{Var}(\widehat{\beta})$ 的四种估计 D_2，D_3，D_4，D_5 中，最优者为 D_4．

3. 无论是在等方差条件下，还是在异方差条件下，对 $\mathrm{Var}(\widehat{\beta})$ 的五种估计 D_1 是最优者．因此，当我们应用线性回归模型时，可以放心使用普通最小二乘估计而不必担心是否存在异方差性．

参 考 文 献

[1] 王松桂，线性模型的理论及应用，合肥：安徽教育出版社，1987.

[2] White, H., A heteroskedastic-consistent covariance matrix estimator and a direct test of heteroskedasticity, *Econometrica*, **48**(1980), 817–838.

[3] Hinkley, D.N., Jackknifing in unbalanced situations, *Technometrics*, **19**(1977), 285–292.

[4] 陈希孺，王松桂，近代回归分析，合肥：安徽教育出版社，1987.

[5] Horn, S.D., Horn, R.A., and Duncan, D.B., Estimating heteroscedastic variances in linear model, *Journal of the American Statistical Association*, **70**(1975), 380–385.

[6] MacKinnon, J.G., and White, H., Some heteroskedasticity consistent covariance matrix estimators with improved finite sample properties, *Journal of Econometrics*, **29**(1985), 53–57.

[7] Long, J.S. and Ervin, L.H., Using heteroscedasticity consistent standard errors in the linear regression model, *The American Statistician*, **54**(2000), 217–224.

[8] 高慧璇，统计计算，北京：北京大学出版社，1994.

Comparison of Estimates on Covariance Matrix

CHEN MAOXUE

(*College of Agronomy, Shandong Agricultural University, Taian, 271018*)

For linear regression models with heteroscedastical errors, this paper compares the estimates of covariance matrix on the least square estimate of regression coefficiencies by computer simulation. Our research shows that when sample size is greater than 50, the least square estimate of regression coefficiencies has high estimate precision. Among five estimates of its covariance, that of the ordinary least squares estimate is the best.