

极大似然估计和拟极大似然估计模拟之比较

何 春¹, 方积乾²

(1. 广东工业大学 应用数学学院, 广东 广州 510090;

2. 中山大学 公共卫生学院流行病与卫生统计系, 广东 广州 510080)

摘要: 通过对单个协变量的带有测量误差的一维结构回归模型中总体平均处理效应的极大似然估计和拟极大似然估计的随机模拟结果进行比较, 发现这两个公式都不受测量误差的影响, 并且可以互换使用. 当其它误差较小时用两个公式计算结果虽然相差不大, 但相比较而言用拟极大似然估计较好, 反之, 当其它误差较大时用极大似然估计较好.

关键词: 总体平均处理效应; 极大似然估计; 拟极大似然估计; 随机抽样; 测量误差; 随机模拟; 比较

中图分类号: O212.1

文献标识码: A

文章编号: 1007-7162(2006)01-0114-03

引言

可交换条件下单个协变量的带有测量误差的多维结构回归模型为

$$\begin{cases} Y_j = \alpha_t + x_{ij}\beta_t + \varepsilon_j, \\ X_j = x_{ij} + u_{ij}, j = 1, \dots, n_t, t = 0, 1, \\ (x_{ij}, u_{ij}, \varepsilon_j)^\tau \sim N[(\mu_x, 0, 0)^\tau, \text{diag}(\sigma_x, \sigma_u, \sigma_\varepsilon I_n)], t = 0, 1 \text{ 且相互独立,} \end{cases}$$

其中 $\sim N$ 表示“服从正态分布且相互独立”, $\text{diag}(\alpha, \alpha_u, \alpha_\varepsilon I_n)$ 是对角矩阵, $(x_{ij}, u_{ij}, \varepsilon_j)^\tau$ 表示 $(x_{ij}, u_{ij}, \varepsilon_j)$ 的转置, Y_j 是第 t 种处理下取得的响应变量的第 j 个观察值, x_{ij} 是对应于 Y_j 的在协变量 x 上的度量, X_j 是协变量 x_{ij} 的观察值, u_{ij} 为测量误差. $Y_j, \alpha_t, \varepsilon_j$ 均为 n 维行向量, β_t 为 n 维行向量, 假定 Y_j 的每一个分量与 x_{ij} 之间的真实关系是线性的, 线性回归系数构成 β_t, ε_j 是随机误差向量 x_{ij}, u_{ij} 的方差 σ_x, σ_u 都大于零, 且 $\sigma_\varepsilon > 0$. 模型中假定, 协变量 x 为随机变量, 服从正态分布 $N(\mu_x, \sigma_x)$, 而 $\{x_{ij}, j = 1, \dots, n_t\}, t = 0, 1$ 为来自同一正态母体 $N(\mu_x, \sigma_x)$ 的两个独立子样^[1-2].

如果对于给定的协变量 x , 在可交换条件下处理 $S = 1$ (相对于 $S = 0$)的平均处理效应是

$$ATE = (\alpha_1 - \alpha_0) + \mu_x(\beta_1 - \beta_0), \quad (1)$$

当 $\alpha_1, \alpha_0, \mu_x, \beta_1, \beta_0$ 的值给定时就可以计算出总体平均处理效应的真值.

文献[3]给出了可交换条件下单个协变量的带有测量误差的一维结构回归模型中总体平均处理效应当暴露组和对照组的测量误差分布相同时的极大似然估计 $E = (\bar{Y}_1 - \bar{Y}_0) - (\sum_{i=0}^1$

收稿日期: 2004-10-20

基金项目: 国家自然科学基金重点课题资助项目(39930160)

作者简介: 何春(1967), 女, 讲师, 主要研究方向为概率统计、卫生统计.

$\frac{n_{1-t}}{n_1+n_0} \frac{M_{X_t Y_t}}{M_{XX}})(\bar{X}_1 - \bar{X}_0), t = 0, 1$, 即模型中 $n = 1$ 时总体平均处理效应的估计. 本文作者在文献 [4] 中给出了可交换条件下单个协变量的带有测量误差的多维结构回归模型中总体平均处理效应暴露组和对照组的测量误差分布相同时的拟极大似然估计 $E_{QL} = (\bar{Y}_1 - \bar{Y}_0) - (\sum_{i=0}^1 \frac{n_{1-t}}{n_1+n_0}$

$\frac{M_{X_t Y_t}}{M_{XX}})(\bar{X}_1 - \bar{X}_0)$, 其中 $M_{X_t X_t} = \frac{1}{n_t} \sum_{j=1}^{n_t} (X_{jt} - \bar{X}_t)^2, M_{XX} = \frac{1}{n_1+n_0} \sum_{t=0}^1 n_t M_{X_t X_t}, M_{X_t Y_t} = \frac{1}{n_t} \sum_{j=1}^{n_t} (X_{jt} - \bar{X}_t)(Y_{jt} - \bar{Y}_t), t = 0, 1$. 由于在多维 ($n \geq 2$) 时无法求出总体平均处理效应的极大似然估计, 即无法求出模型中某些参数的极大似然估计, 只好用其矩估计代替, 最后把求出的估计称为拟极大似然估计 (quasi-maximum like lihood estimators). 事实上, 在求拟极大似然估计时并没有限制 $n \geq 2$, 即 n 可以取 1. 那么对于可交换条件下单个协变量的带有测量误差的一维结构回归模型中总体平均处理效应, 既可以用极大似然估计也可以用拟极大似然估计去估计它, 从随机模拟的结果来看, 这两个公式各有优点.

1 随机模拟结果之比较

设参数值为: $\alpha_1 = 3.5, \alpha_0 = 3.5, \mu = 15, \beta_1 = 4.6, \beta_0 = 4.5, n_1 = 100, n_0 = 100$. 所以有 $ATE = 1.5$. 利用计算机进行 100 次随机抽样 (方法是通过调用 SAS 函数 `rannor(seed, z)`, 产生正态随机变量), 得到 100 份样本, 每份样本中暴露组与对照组各含 100 个个体. 分别用极大似然估计公式和拟极大似然估计公式对样本进行计算, 求出总体平均处理效应的估计值 E_L 和 E_{QL} , 然后通过对误差的变化的控制, 不断重复以上过程, 最后分析结果. 表 1~7 是模拟结果分析 (当暴露组和对照组的测量误差相同时, σ_u 记为 σ_u , 即 $\sigma_{u1} = \sigma_{u0} = \sigma_u$).

表 1 当 $\sigma_x = 10, \sigma_t = 10, t = 0, 1, \sigma_u = 1$ 时, 由 E_L 及 E_{QL} 计算结果之数据分析					
变量	均值	标准差	标准误	t 值	P 值
E_L	1.376 8	1.497 9	0.149 8	9.190	< 0.000 1
E_{QL}	1.993 8	6.036 0	0.603 6	3.30	0.001 3

表 3 当 $\sigma_x = 10, \sigma_t = 10, t = 0, 1, \sigma_u = 100$ 时, 由 E_L 及 E_{QL} 计算结果之数据分析					
变量	均值	标准差	标准误	t 值	P 值
E_L	2.007 0	6.070 1	0.607 0	3.31	0.001 3
E_{QL}	1.997 0	6.062 9	0.606 3	3.29	0.001 4

表 5 当 $\sigma_x = 1, \sigma_t = 1, t = 0, 1, \sigma_u = 10$ 时, 由 E_L 及 E_{QL} 计算结果之数据分析					
变量	均值	标准差	标准误	t 值	P 值
E_L	1.550 7	0.607 0	0.067 0	25.55	< 0.000 1
E_{QL}	1.549 7	0.606 3	0.060 6	25.56	< 0.000 1

2 小结

由以上 7 表可以看出:

1) 由两个公式计算出来的总体平均处理

表 2 当 $\sigma_x = 10, \sigma_t = 10, t = 0, 1, \sigma_u = 10$ 时, 由 E_L 及 E_{QL} 计算结果之数据分析					
变量	均值	标准差	标准误	t 值	P 值
E_L	1.770 0	4.733 3	0.473 3	3.74	0.000 3
E_{QL}	1.995 7	6.049 9	0.605 0	3.30	0.001 3

表 4 当 $\sigma_x = 1, \sigma_t = 1, t = 0, 1, \sigma_u = 1$ 时, 由 E_L 及 E_{QL} 计算结果之数据分析					
变量	均值	标准差	标准误	t 值	P 值
E_L	1.527 0	0.473 3	0.047 3	32.26	< 0.000 1
E_{QL}	1.549 6	0.605 0	0.065 0	25.61	< 0.000 1

表 6 当 $\sigma_x = 1, \sigma_t = 1, t = 0, 1, \sigma_u = 100$ 时, 由 E_L 及 E_{QL} 计算结果之数据分析					
变量	均值	标准差	标准误	t 值	P 值
E_L	1.549 6	0.605 0	0.060 5	25.61	< 0.000 1
E_{QL}	1.549 7	0.606 3	0.060 6	25.56	< 0.000 1

表 7 当 $\sigma_x = 5, \sigma_t = 5, t = 0, 1, \sigma_u = 10$ 时, 由 E_L 及 E_{QL} 计算结果之数据分析					
变量	均值	标准差	标准误	t 值	P 值
E_L	1.720 4	2.852 5	0.285 3	6.03	< 0.000 1
E_{QL}	1.748 3	3.029 0	0.302 9	5.77	< 0.000 1

效应的估计值基本不受测量误差的影响;

2) 两个公式可以互换使用;

3) 当其它的误差(不同于测量误差的误差)较小时用两个公式计算的结果虽然相差不大,但相比较而言用拟极大似然估计好些,反之,当其它的误差较大时用极大似然估计好些;

4) 在实际计算中应根据非测量误差的大小来选用合适的公式.

参考文献:

[1] Fuller W A. Measurement Error Models[M]. New York: Wiley, 1987.

[2] Greenland S, Robins J M, Pearl J. Confounding and Collapsibility in Causal Inference[J]. Statistical Science, 1999, 14(1): 29-46.

[3] 金华. 可交换条件下平均处理效应的估计及应用[D]. 广州: 中山医科大学公共卫生学院, 2000.

[4] 何春. 可交换条件下多维结构模型中总体平均处理效应的估计[D]. 广州: 中山大学公共卫生学院, 2004.

The Comparison between the Random Simulation of the Maximum Likelihood Estimators and the Quasi-maximum Likelihood Estimators

HE Chun¹, FANG Ji-qian²

(1. Faculty of Applied Mathematics, Guangdong University of Technology, Guangzhou 510090, China;

2. Dept. of Medical Statistics, School of Public Health, Sun Yat-sen University, Guangzhou 510080, China)

Abstract: In this paper, the comparison is given between the random simulation of the maximum likelihood estimators and the quasi-maximum likelihood estimators of the population-averaged treatment effects in the one-dimensional structural regression models with one covariate and the measurement errors. It is found that one of the formulae can be replaced by the other, and they are not influenced by the measurement errors. When the other errors(that is, they are different from the measurement errors) are little, the results are similar when using the two formulae, but the use of the quasi-maximum likelihood estimators is better than the use of the maximum likelihood estimators. On the contrary, when the other errors are large, the use of the maximum likelihood estimators is better than that of the quasi-maximum likelihood estimators.

Key words: population-averaged treatment effects; maximum likelihood estimators; quasi-maximum likelihood estimators; random sampling; measurement error; random simulation; comparison