DOI: 10.13656/j.cnki.gxkx.2009.01.013
广西科学 Guangxi Sciences 2009, 16(1): 46~47, 54

同模型下数据缺失时线性回归模型反应变量均值的经验似然置信区间*

Empirical Likelihood Confidence Intervals of the Mean of the Response Variables for the Same Linear Regression Model with Missing Data

庞伟才,韦程东

PANG Wei-cai, WEI Cheng-dong

(广西师范学院数学与计算机科学系,广西南宁 530023)

(Department of Mathematics and Computer Science, Guangxi Teachers Education University, Nanning, Guangxi, 530023, China)

摘要:在两独立总体具有相同的线性回归模型下,当第一总体的样本为完全样本,第二总体的反应变量完全缺失时,利用第一总体的样本信息,得到第二总体反应变量均值的经验似然置信区间.

关键词:线性回归 经验似然 置信区间

中图法分类号: O212.1 文献标识码: A 文章编号: 1005-9164(2009)01-0046-02

Abstract This paper supposes that two independent population have the same linear regession model, when the samples of the first population are complete and the response variables of the second population are missing. By using the sample information of the first population, we obtain mean's empirical likelihood confidence intervals of the second population.

Key words linear regression, empirical likelihood, confidence intervals

考虑两个独立总体 (X_1, Y_1), (X_2, Y_2), (X_i, Y_i), $i = 1, 2$ 为 $R^d \times R^1$ 上的随机向量,假设它们都满足相同的线性回归模型

$$Y = X^{\mathrm{T}}U + \varepsilon, \tag{1}$$

其中 U 为 d 维未知常数向量,ε 为随机误差,且 $E\varepsilon = 0, 0 < \sigma^2 = E\varepsilon^2 < \infty$,$\varepsilon$ 与 X 独立.某次试验要从两总体中抽样,得到总体 (X_1, Y_1) 的样本为完全样本 $(X_{11}, Y_{11}), \cdots, (X_{n_1 1}, Y_{n_1 1})$,总体 (X_2, Y_2) 的样本为不完全样本 $(X_{12}, *), (X_{n_2 2}, *)$,其中 $*$ 表示 Y_{i2} 全部缺失,$i = 1, 2, \cdots, n_2$.本文在一些假设条件下,利用总体 (X_1, Y_1) 的完全样本的有关信息,得到 $\theta = EY_2$ 的经验似然置信区间.

1 相关引理

利用总体 (X_1, Y_1) 的完全样本构造 U 的最小二乘估计

$$\hat{U} = \left(\sum_{j=1}^{n_1} X_{j1}X_{j1}^{\mathrm{T}}\right)^{-1}\sum_{j=1}^{n_1} X_{j1}Y_{j1}, \tag{2}$$

因此,可以用 $\hat{Y}_{i2} = X_{i2}^{\mathrm{T}}\hat{U}$ 补充 Y_{i2}.

类似于文献[1],可得 $\theta = EY_2$ 的对数经验似然比统计量

$$\hat{l}_{n_2}(\theta) = 2\sum_{i=1}^{n_2} \log\left\{1 + \lambda_{n_2}(\hat{Y}_{i2} - \theta)\right\}, \tag{3}$$

其中 $\lambda_{n_2} = \lambda_{n_2}(\theta)$ 满足方程

$$\sum_{i=1}^{n_2} \frac{\hat{Y}_{i2} - \theta}{1 + \lambda_{n_2}(\hat{Y}_{i2} - \theta)} = 0. \tag{4}$$

引理 **1** 假设 $E\|X\|^2 < \infty$, $0 < \sigma^2 = E\varepsilon^2 < \infty$, $n_2/n_1 \to \lambda$, $0 < \lambda < \infty$,如果 θ 为真参数,则有

$$n_2^{-1/2}\sum_{i=1}^{n_2}(\hat{Y}_{i2} - \theta) \xrightarrow{L} N(0, V(\theta)),$$

其中 $V(\theta) = U^{\mathrm{T}}EX_2X_2^{\mathrm{T}}U - 2(EX_2)^{\mathrm{T}}U\theta + \theta^2 + \lambda(EX_2)^{\mathrm{T}}(EX_1X_1^{\mathrm{T}})^{-1}EX_2\sigma^2$.

证明 由于

$$n_2^{-1/2}\sum_{i=1}^{n_2}(\hat{Y}_{i2} - \theta) = n_2^{-1/2}\sum_{i=1}^{n_2} X_{i2}^{\mathrm{T}}(\hat{U} -$$

收稿日期: 2008-01-25

作者简介:庞伟才(1979-),男,硕士,主要从事概率论与数理统计的教学与研究.

* 广西自然科学基金项目(0575051),广西师范学院青年科研基金项目(0811B001)资助.

$$\hat{U}) + n_2^{-1/2}\sum_{i=1}^{n_2}(X_{i2}^{\mathrm{f}}U - \theta) = R_1 + R_2. \quad (5)$$

由 $\hat{U}$的定义有

$$R_1 = n_2^{-1/2}\sum_{i=1}^{n_2} X_{i2}^{\mathrm{f}}(\hat{U} - U) =$$

$$n_2^{-1/2}\sum_{i=1}^{n_2} X_{i2}^{\mathrm{f}}\left\{\left(\sum_{j=1}^{n_1} X_{j1}X_{j1}^{\mathrm{f}}\right)^{-1}\sum_{j=1}^{n_1} X_{j1}Y_{j1}\right\} =$$

$$\lambda^{1/2}EX_2^{\mathrm{f}}(EX_1X_1^{\mathrm{f}})^{-1}n_1^{-1/2}\sum_{j=1}^{n_1} X_{j1}Y_{j1} + o_p(1) \xrightarrow{L} N(0,$$

$$\lambda(EX_2)^{\mathrm{f}}(EX_1X_1^{\mathrm{f}})^{-1}EX_2\sigma^2). \quad (6)$$

由中心极限定理[2]知

$$R_2 \xrightarrow{L} N(0, U^{\mathrm{f}}EX_2X_2^{\mathrm{f}}U - 2(EX_2)^{\mathrm{f}}U\theta + \theta^2). \quad (7)$$

直接计算得 $\mathrm{cov}(R_1, R_2) = 0$,因此,由(5)~(7)式知引理 1成立.

引理 **2** 在引理 1的条件下,如果 θ为真参数,则有

$$n_2^{-1}\sum_{i=1}^{n_2}(\hat{Y}_{i2} - \theta)^2 \xrightarrow{p} V_1(\theta),$$

其中 $V_1(\theta) = U^{\mathrm{f}}EX_2X_2^{\mathrm{f}}U - 2(EX_2)^{\mathrm{f}}U\theta + \theta^2$.

证明 容易看到

$$n_2^{-1}\sum_{i=1}^{n_2}(\hat{Y}_{i2} - \theta)^2 = I_1 + I_2 + I_3, \quad (8)$$

其中 $I_1 = n_2^{-1}\sum_{i=1}^{n_2}(X_{i2}^{\mathrm{f}}(\hat{U} - U))^2$, $I_2 = n_2^{-1}\sum_{i=1}^{n_2}(X_{i2}^{\mathrm{f}}U - \theta)^2$, $I_3 = 2n_2^{-1}\sum_{i=1}^{n_2} X_{i2}^{\mathrm{f}}(\hat{U} - U)(X_{i2}^{\mathrm{f}}U - \theta)$.

由大数定律及 $\hat{U} \xrightarrow{p} U$知

$$I_1 = o_p(1), I_3 = o_p(1), \quad (9)$$

$$I_2 = U^{\mathrm{f}}EX_2X_2^{\mathrm{f}}U - 2(EX_2)^{\mathrm{f}}U\theta + \theta^2 + o_p(1). \quad (10)$$

因此,由(8)~(10)式知引理 2成立.

引理 **3** 令 $\hat{Y}_{(n_2)} = \max_{1\le i\le n_2}|\hat{Y}_{i2}|$,在引理 1的条件下有 $\hat{Y}_{(n_2)} = o_p(n_2^{1/2})$.

证明 $\hat{Y}_{(n_2)} = \max_{1\le i\le n_2}|\hat{Y}_{i2}| = \max_{1\le i\le n_2}|X_{i2}^{\mathrm{f}}\hat{U}| \le \max_{1\le i\le n_2}\|X_{i2}\|\,\|\hat{U}\|$,由假设条件 $E\|X\|^2 < \infty$,根据文献[3]有 $\max_{1\le i\le n_2}\|X_{i2}\| = o_p(n_2^{1/2})$,再由 $\hat{U} = O_p(1)$,引理 3得证.

引理 **4** 若引理 1的条件成立,则有 $\lambda_{n_2} = O_p(n_2^{-1/2})$.

证明 由引理 1易得 $n_2^{-1}\sum_{i=1}^{n_2}(\hat{Y}_{i2} - \theta) = O_p(n_2^{-1/2})$,再根据引理 2和引理 3,类似于文献[1]的方法可得引理 4成立.

2 主要结果

定理 **1** 假设 $E\|X\|^2 < \infty$, $0 < \sigma^2 = EY^2 < \infty$, $n_2/n_1 \to \lambda$, $0 < \lambda < \infty$,如果 θ为真参数,则有

$$\hat{l}_{n_2}(\theta) \xrightarrow{L} \frac{V(\theta)}{V_1(\theta)}\chi_1^2,$$

其中 $V(\theta) = U^{\mathrm{f}}EX_2X_2^{\mathrm{f}}U - 2(EX_2)^{\mathrm{f}}U\theta + \theta^2 + \lambda(EX_2)^{\mathrm{f}}(EX_1X_1^{\mathrm{f}})^{-1}EX_2\sigma^2$, $V_1(\theta) = U^{\mathrm{f}}EX_2X_2^{\mathrm{f}}U - 2(EX_2)^{\mathrm{f}}U\theta + \theta^2$, χ_1^2是自由度为 1的 χ^2变量, $\xrightarrow{L}$表示依分布收敛.

证明 对(3)式进行泰勒展开可得

$$\hat{l}_{n_2}(\theta) = 2\sum_{i=1}^{n_2}\left[\lambda_{n_2}(\hat{Y}_{i2} - \theta) - \frac{1}{2}(\lambda_{n_2}(\hat{Y}_{i2} - \theta))^2\right] + V_{n_2}, \quad (11)$$

其中 $|V_{n_2}| \le c\sum_{i=1}^{n_2}|\lambda_{n_2}(\hat{Y}_{i2} - \theta)|^3$, a. s.. 根据引理 2~4,有

$$|V_{n_2}| \le c|\lambda_{n_2}|^3 \max_{1\le i\le n_2}|\hat{Y}_{i2} - \theta|\sum_{i=1}^{n_2}(\hat{Y}_{i2} - \theta)^2 = o_p(1). \quad (12)$$

注意到

$$n_2^{-1}\sum_{i=1}^{n_2}\frac{\hat{Y}_{i2} - \theta}{1 + \lambda_{n_2}(\hat{Y}_{i2} - \theta)} = n_2^{-1}\sum_{i=1}^{n_2}(\hat{Y}_{i2} - \theta) -$$

$$\left[n_2^{-1}\sum_{i=1}^{n_2}(\hat{Y}_{i2} - \theta)^2\right]\lambda_{n_2} + n_2^{-1}\sum_{i=1}^{n_2}\frac{\lambda_{n_2}^2(\hat{Y}_{i2} - \theta)^3}{1 + \lambda_{n_2}(\hat{Y}_{i2} - \theta)}. \quad (13)$$

由(4)式,(13)式及引理 2~4,得到

$$\lambda_{n_2} = \left(\sum_{i=1}^{n_2}(\hat{Y}_{i2} - \theta)^2\right)^{-1}\sum_{i=1}^{n_2}(\hat{Y}_{i2} - \theta) + o_p(n_2^{-1/2}). \quad (14)$$

再利用(4)式,有

$$0 = \sum_{i=1}^{n_2}\frac{\hat{Y}_{i2} - \theta}{1 + \lambda_{n_2}(\hat{Y}_{i2} - \theta)} =$$

$$\sum_{i=1}^{n_2}[\lambda_{n_2}(\hat{Y}_{i2} - \theta)] - \sum_{i=1}^{n_2}[\lambda_{n_2}(\hat{Y}_{i2} - \theta)]^2 +$$

$$\sum_{i=1}^{n_2}\frac{[\lambda_{n_2}(\hat{Y}_{i2} - \theta)]^3}{1 + \lambda_{n_2}(\hat{Y}_{i2} - \theta)}. \quad (15)$$

由(4)式及引理 3和 4又可以得到

$$\sum_{i=1}^{n_2}\frac{[\lambda_{n_2}(\hat{Y}_{i2} - \theta)]^3}{1 + \lambda_{n_2}(\hat{Y}_{i2} - \theta)} = o_p(1). \quad (16)$$

由(15)式和(16)式有

$$\sum_{i=1}^{n_2}\lambda_{n_2}(\hat{Y}_{i2} - \theta) = \sum_{i=1}^{n_2}[\lambda_{n_2}(\hat{Y}_{i2} - \theta)]^2 + o_p(1). \quad (17)$$

(下转 第 54页 Continue on page 54)

$|f(x^*)| < 10^{-14}$.计算结果见表 2.

测试函数 $f(x)$分别取为(a) x^3+x^2-10, $x=1.8674600246063$, (b) $(x-1)^6-1$, $x=2$, (c) $(x-1)^3(x+2)^4$, $x_1=2$, $x_2=-2$, (d) $\sin(x-1)+x-1$, $x=1$.

从表 2结果可以看出,新方法适合函数类的范围比文献[4]和文献[6]都要宽.

表 2 数值实验结果

Table 2 Numerical results

$f(x)$	x_0	迭代次数 N Iteration times N				
		HN[4]	MN[4]	SN[6]	GN[6]	新方法 New algorithm
(a)	-0.5	42	10	10	4	8
	1	3	3	3	3	8
	2	3	3	3	3	4
(b)	1.5	7	58	195	12	13
	2.5	4	5	5	5	6
	3	5	6	6	5	7
(c)	1.4	41	49	49	46	49
	-3	60	70	72	67	70
(d)	1.5	3	3	2	3	4
	3	14643	4	3	7	63
	-1	18592	4	3	7	13

5 结束语

新方法在较广的函数类中至少是三阶收敛,比文献[1]的指数迭代法高一阶.同时,新方法也保持了文献[2]的一个重要的优点:去掉了强加给 $f(x)$的单调性即要求 $f'(x)\neq 0$.

总之,不同的方法有不同的特点,从以上的理论分析和数值实验可知,新方法是一种较优的求解非线性方程的方法,在理论上和实用上都有一定的价值.

参考文献:

[1] 吴新元. 解非线性方程的二阶敛速指数迭代法[J]. 计算数学, 1998, 20(4): 367-370.

[2] 吴新元.对牛顿迭代法的一个重要修改[J]. 应用数学与力学, 1999, 20(8): 863-866.

[3] 吴忠麟, 吴新元.解非线性方程的一个非线性迭代法[J]. 高等学校计算数学学报, 1995, 17(4): 318-322.

[4] ZBAN A Y O. Some variants of Newton's methods [J]. Applied Mathematics Letters, 2004, 17: 677-682.

[5] 郑权,黄松奇. 解非线性方程的 Newton类方法及其变形[J]. 清华大学学报: 自然科学版, 2004, 44(3): 372-375.

[6] 王霞,赵玲玲,李飞敏. 牛顿方法的两个新格式[J]. 数学的实践与认识, 2007, 37(1): 72-76.

(责任编辑: 尹 闯)

(上接第 47页 Continue from page 47)

再由(11)式,(12)式,(14)式和(17)式得到

$$\hat{l}_{n_2}(\theta)=\left[n_2^{-1}\sum_{i=1}^{n_2}\left(\hat{Y}_{i2}-\theta\right)^2\right]^{-1}\left[n_2^{-1/2}\sum_{i=1}^{n_2}\left(\hat{Y}_{i2}-\theta\right)\right]^2+o_p(1). \quad (18)$$

最后由引理 1和 2知,定理 1成立.

由于非标准的 χ^2分布不能对 θ作区间估计,故需引入调整的对数经验似然比 $\hat{l}_{n_2,ad}(\theta)$,首先构造 $V(\theta)$的相合估计,令

$$\bar{X}_2=n_2^{-1}\sum_{i=1}^{n_2}X_{i2},\ \hat{e}^2=n_1^{-1}\sum_{i=1}^{n_1}(Y_{i1}-X_{i1}^{\mathrm{T}}\hat{U})^2,\ \overline{X_1X_1^{\mathrm{T}}}=n_1^{-1}\sum_{i=1}^{n_1}X_{i1}X_{i1}^{\mathrm{T}},\ \overline{X_2X_2^{\mathrm{T}}}=n_2^{-1}\sum_{i=1}^{n_2}X_{i2}X_{i2}^{\mathrm{T}},$$

则 $\hat{V}(\theta)=\hat{U}^{\mathrm{T}}\overline{X_2X_2^{\mathrm{T}}}\hat{U}-2(\bar{X}_2)^{\mathrm{T}}\hat{U}\theta+\theta^2+\frac{n_2}{n_1}(\bar{X}_2)^{\mathrm{T}}(\overline{X_1X_1^{\mathrm{T}}})^{-1}\bar{X}_2\hat{e}^2$为 $V(\theta)$的相合估计.由引理 2知 $\hat{V}_1(\theta)=n_2^{-1}\sum_{i=1}^{n_2}(\hat{Y}_{i2}-\theta)^2$为 $V_1(\theta)$的相合估计,进一步令 $r(\theta)=\frac{\hat{V}_1(\theta)}{\hat{V}(\theta)}$, $\hat{l}_{n_2,ad}(\theta)=r(\theta)\hat{l}_{n_2}(\theta)$.

定理 **2** 在定理 1的假设下,如果 θ为真参数,则 $\hat{l}_{n_2,ad}(\theta)$渐近 χ^2_1分布,即 $P(\hat{l}_{n_2,ad}(\theta)\leq c_T)=1-T+o(1)$,其中 $P(\chi^2_1\leq c_T)=1-T$.

证明 由 $\hat{l}_{n_2,ad}(\theta)$的定义和(18)式知

$$\hat{l}_{n_2,ad}(\theta)=\left(n_2^{-1/2}\sum_{i=1}^{n_2}\frac{\hat{Y}_{i2}-\theta}{\hat{V}^{1/2}(\theta)}\right)^2+o_p(1). \quad (19)$$

由于 $\hat{V}(\theta)$为 $V(\theta)$的相合估计,由引理 1和(19)式,定理 2得证.

注 由定理 2可以构造 θ的置信区间,令 $I_{n_2,T}=\{\theta': \hat{l}_{n_2,ad}(\theta')\leq c_T\}$,则 $P(\theta\in I_{n_2,T})=1-T+o(1)$.

参考文献:

[1] Owen A. Empirical likelihood ratio confidence intervals [J]. Biometrika, 1988, 75(2): 237-249.

[2] 苏淳.概率论[M].北京: 科学出版社, 2004.

[3] Owen A. Empirical likelihood ratio confidence regions [J]. Annals of statistics, 1990, 18(1): 90-120.

(责任编辑: 尹 闯)