应用概率统计 第 33 卷
第 3 期 2017 年 6 月

Chinese Journal of Applied Probability and Statistics
Jun., 2017, Vol. 33, No. 3, pp. 221-231
doi: 10.3969/j.issn.1001-4268.2017.03.001

污染数据线性回归模型的最小一乘估计 *

叶 鹏 周秀轻*

(南京师范大学数学科学学院、金融与统计研究所, 南京, 210023)

摘 要: 对于线性回归模型, 在因变量受到另一与之独立的随机变量序列的污染时, 基于最小一乘的方法给出模型参数的估计. 在一定条件下, 证明了估计量的相合性和渐近正态性, 并使用模拟对估计方法的小样本性质进行了分析. 模拟结果显示, 本文所提方法在小样本情况下表现良好.

关键词: 污染数据; 最小一乘估计; 相合性; 渐近正态性

中图分类号: O212.1

英文引用格式: Ye P, Zhou X Q. LAD estimation for linear regression models with contaminated data [J]. Chinese J. Appl. Probab. Statist., 2017, 33(3): 221–231. (in Chinese)

§1. 引 言

在实际工作中经常会遇到无法观测到兴趣变量本身, 而只能观测到被污染后的数据的情况. 污染数据模型也被称为混合模型, 在许多领域都有广泛的应用, 比如计量经济学(见文献[1]和[2])、生物和流行病学(见文献[3]和[4])等. 1952年, Davis[5]首次提出污染数据和污染系数的概念. 在此之后, Huber[6]于1964年提出了一类"被污染的正态分布族", 即$F_{\mathrm{N},v}=\{f: f=(1-v)\mathrm{N}(0,1)+vg, g\in F_s\}$, 其中$F_s$ 为一切关于原点对称的一维概率函数的集合. 1991 年, Yu[7]利用矩方法、贝叶斯方法和极大似然方法分别对污染参数进行了估计. Copas[8]于1988年提出了污染数据的二元变量回归模型, 即因变量y的取值为1和0, 且取1的概率为τ, 取0的概率为$(1-\tau)$, 而实际记录y的取值概率为τ^*, 且满足$\tau^*=(1-\gamma)\tau+\gamma(1-\tau)$, 并对模型的异常值和稳健性进行了讨论.

1996年, 郑祖康等[9]提出了污染数据的线性回归模型:

$$y_i=x_i^{\mathrm{T}}\beta+\epsilon_i,\qquad i=1,2,\ldots,n, \tag{1}$$

其中x_i是p维自变量, β是p维待估参数向量, ϵ_i为相互独立的随机误差, 因变量$\{y_i\}$会受到另一串与之独立的随机变量序列$\{t_i\}$的干扰而无法观测到. 真实的观测数据为$\{(y_i^*,x_i),\ i=1,2,\ldots,n\}$, 其中$y_i^*$满足

$$y_i^*=(1-\alpha)y_i+\alpha t_i,\qquad i=1,2,\ldots,n, \tag{2}$$

*国家自然科学基金项目(批准号: 11201235、11471252)和江苏省高校自然科学基金项目(批准号: 12KJB110010)资助.

*通讯作者, E-mail: zhouxiuqing@njnu.edu.cn.

本文2015年1月9日收到, 2016年8月17日收到修改稿.

这里t_i相互独立, $0 \leqslant \alpha < 1$. 常数α称为污染系数或污染比例.

针对上述污染数据的线性回归模型, 郑祖康等[9]在ϵ_i和t_i均服从正态分布的假设条件下, 利用最小二乘方法给出了模型参数的估计, 并在一定条件下证明了参数估计的强相合性. 任哲和陈明华[10]在同样的正态分布假设下, 构造了回归参数的最小一乘估计量, 并在一定条件下证明了估计量的相合性和渐近正态性. 上述工作中都要求ϵ_i和t_i均服从正态假设且方差已知, 而在实际中, ϵ_i和t_i的分布不一定都适合使用正态分布近似, 尤其当ϵ_i和t_i存在异常值或其分布为重尾时, 使用正态假设显然是不合适的. 针对上述污染数据的线性回归模型, 陈明华[11]在不假设ϵ_i和t_i服从正态分布, 但要求其方差均已知的情况下, 给出了参数的最小二乘估计, 并证明了参数估计的强相合性. 本文假设$\epsilon_i \sim \text{Laplace}(0, \sigma_1)$, $t_i \sim \text{Laplace}(0, \sigma_2)$, 在$\sigma_1$和$\sigma_2$均已知或有一个已知, 或者$\sigma_2/\sigma_1 = l$已知但$\sigma_1$和$\sigma_2$均未知三种情况下, 对污染数据线性回归模型进行考虑. 我们使用最小一乘估计构造参数的估计量, 并在一定条件下证明它们的相合性和渐近正态性; 最后为了考查本文所提的方法的小样本性质, 我们对本文提出的方法进行了模拟, 从模拟的结果看, 本文所提方法在小样本情况下表现良好.

§2. 参数估计方法和主要结果

由(1)、(2)可得一个新模型

$$y_i^* = x_i^{\mathrm{T}}\beta^* + \eta_i, \qquad i = 1, 2, \ldots, n, \tag{3}$$

其中

$$\beta^* = (1-\alpha)\beta, \qquad \eta_i = (1-\alpha)\epsilon_i + \alpha t_i. \tag{4}$$

在新的回归模型(3)中, β^*为p维未知回归参数向量, $\{\eta_i\}$为相互独立的随机误差项序列, 并且假设$\epsilon_i \sim \text{Laplace}(0, \sigma_1)$, $t_i \sim \text{Laplace}(0, \sigma_2)$. 在此, 我们假设条件

$$(1-\alpha)\sigma_1 > \alpha\sigma_2 \tag{5}$$

成立. 直观上看, 条件(5)可以保证污染造成的影响不会超过系统本身, 这在实际应用中是个合理的要求, 文献[9–11]中也对所研究的模型有类似的约束. 在此条件下, 可得η_i的密度函数

$$f(\eta) = \frac{1}{2((1-\alpha)^2\sigma_1^2 - \alpha^2\sigma_2^2)}\left[(1-\alpha)\sigma_1 \mathrm{e}^{-|\eta|/[(1-\alpha)\sigma_1]} - \alpha\sigma_2 \mathrm{e}^{-|\eta|/(\alpha\sigma_2)}\right].$$

以下, 我们均在条件(5)下进行考虑.

显然η_i的中位数依然为0. 下面首先根据最小一乘的方法估计模型(3)的回归参数β^*, 然后使用矩方法得到污染系数的估计, 再结合(4)式得到模型(1)的回归参数β的估计$\widehat{\beta}$, 并给出估计量$\widehat{\beta}$的相合性和渐近正态性.

记$\widehat{\beta}^*$为β^*的最小一乘估计, 则$\widehat{\beta}^*$满足

$$\sum_{i=1}^{n}|y_i^*-x_i^{\mathrm{T}}\widehat{\beta}^*|=\min_{\beta^*\in R^p}\Big\{\sum_{i=1}^{n}|y_i^*-x_i^{\mathrm{T}}\beta^*|\Big\}. \tag{6}$$

结合(4)式, 若α有估计$\widehat{\alpha}$, 则β的最小一乘估计定义为

$$\widehat{\beta}=\frac{\widehat{\beta}^*}{1-\widehat{\alpha}}. \tag{7}$$

记$\widehat{\eta}_i=y_i^*-x_i^{\mathrm{T}}\widehat{\beta}^*$, $R=n^{-1}\sum\limits_{i=1}^{n}\widehat{\eta}_i^2$, $G=n^{-1}\sum\limits_{i=1}^{n}\widehat{\eta}_i^4$. 下面利用矩估计的方法得到污染系数$\alpha$的估计$\widehat{\alpha}$.

首先, 在σ_1,σ_2均已知时, 由η_i的密度函数式可得

$$\mathsf{E}(\eta_i^2)=2[(1-\alpha)^2\sigma_1^2+\alpha^2\sigma_2^2].$$

故可建立下列方程

$$R=2[(1-\widehat{\alpha})^2\sigma_1^2+\widehat{\alpha}^2\sigma_2^2].$$

解得

$$\begin{cases}\widehat{\alpha}=\dfrac{\sigma_1^2-\sqrt{0.5R(\sigma_1^2+\sigma_2^2)-\sigma_1^2\sigma_2^2}}{\sigma_1^2+\sigma_2^2}, & \text{若 } (1-\alpha)\sigma_1^2>\alpha\sigma_2^2;\\ \widehat{\alpha}=\dfrac{\sigma_1^2+\sqrt{0.5R(\sigma_1^2+\sigma_2^2)-\sigma_1^2\sigma_2^2}}{\sigma_1^2+\sigma_2^2}, & \text{若 } (1-\alpha)\sigma_1^2<\alpha\sigma_2^2.\end{cases}$$

其次, 考虑σ_1和σ_2有一个未知的情况. 不妨设σ_2已知而σ_1未知. 由η_i的密度函数式可得

$$\mathsf{E}(\eta_i^2)=2[(1-\alpha)^2\sigma_1^2+\alpha^2\sigma_2^2],\qquad \mathsf{E}(\eta_i^4)=24[(1-\alpha)^4\sigma_1^4+(1-\alpha)^2\sigma_1^2\alpha^2\sigma_2^2+\alpha^4\sigma_2^4].$$

故可建立下列方程组

$$\begin{cases}R=2[(1-\widehat{\alpha})^2\sigma_1^2+\widehat{\alpha}^2\sigma_2^2],\\ G=24[(1-\widehat{\alpha})^4\sigma_1^4+(1-\widehat{\alpha})^2\sigma_1^2\widehat{\alpha}^2\sigma_2^2+\widehat{\alpha}^4\sigma_2^4].\end{cases}$$

解上述方程组可得

$$\begin{cases}\widehat{\alpha}=\sqrt{\dfrac{3R-\sqrt{6G-27R^2}}{12\sigma_2^2}},\\ \widehat{\sigma}_1=\sqrt{\dfrac{3R+\sqrt{6G-27R^2}}{12(1-\widehat{\alpha})^2}}.\end{cases} \tag{8}$$

最后, 考虑$\sigma_2/\sigma_1=l$已知, 但σ_1和σ_2均未知的情况. 使用类似的方法可得

$$\begin{cases}\widehat{\alpha}=\dfrac{M}{M+1},\\ \widehat{\sigma}_1=\sqrt{\dfrac{3R+\sqrt{6G-27R^2}}{12(1-\widehat{\alpha})^2}},\\ \widehat{\sigma}_2=\sqrt{\dfrac{3R-\sqrt{6G-27R^2}}{12\widehat{\alpha}^2}},\end{cases} \tag{9}$$

其中

$$M=\frac{1}{l}\sqrt{\frac{3R-\sqrt{6G-27R^2}}{3R+\sqrt{6G-27R^2}}}.$$

注: 下记$A_n=\sum\limits_{i=1}^{n}x_ix_i^{\mathrm{T}}$, 并假定$\lim\limits_{n\to\infty}\max\limits_{1\leqslant i\leqslant n}x_i^{\mathrm{T}}A_n^{-1}x_i=0$.

定理 1 在本文假定下, 当$n\to\infty$, 有

$$\widehat{\beta}\xrightarrow{\mathrm{P}}\beta, \tag{10}$$

其中"$\xrightarrow{\mathrm{P}}$"表示依概率收敛.

定理 2 在本文假定下, 当$n\to\infty$时, 有

$$(1-\widehat{\alpha})\widehat{V}^{-1}A_n^{1/2}(\widehat{\beta}-\beta)\xrightarrow{L}\mathrm{N}(0,I_p), \tag{11}$$

其中"$\xrightarrow{L}$"表示依分布收敛, 且

$$\begin{cases}\widehat{V}=(1-\widehat{\alpha})\sigma_1+\alpha\sigma_2, & \text{当}\sigma_1\text{和}\sigma_2\text{均已知时;}\\ \widehat{V}=(1-\widehat{\alpha})\widehat{\sigma}_1+\widehat{\alpha}\sigma_2, & \text{当}\sigma_2\text{已知}\sigma_1\text{未知时;}\\ \widehat{V}=(1-\widehat{\alpha}+l\widehat{\alpha})\widehat{\sigma}_1, & \text{当}\sigma_1,\sigma_2\text{均未知但}\sigma_2/\sigma_1=l\text{已知时.}\end{cases} \tag{12}$$

§3. 定理的证明

本节证明定理1和定理2. 为了证明这两个定理, 先给出几个引理.

引理 3 在本文假定下, 当$n\to\infty$时, 有

$$\widehat{\beta}^*\xrightarrow{\mathrm{P}}\beta^*.$$

证明: 易知在模型(3)中, η_i是独立同分布的, 其密度函数f关于零点对称, 且$f(0)>0$, 故文献[12]中定理1的条件满足, 则由文献[12]的系3知引理3成立. □

引理 4 在本文的假定下, 当$n\to\infty$时, 有

$$\widehat{\alpha}\xrightarrow{\mathrm{P}}\alpha.$$

证明: 由模型(3)知

$$\begin{aligned}R&=\frac{1}{n}\sum_{i=1}^{n}(y_i^*-x_i^{\mathrm{T}}\widehat{\beta}^*)^2=\frac{1}{n}\sum_{i=1}^{n}(x_i^{\mathrm{T}}\beta^*-x_i^{\mathrm{T}}\widehat{\beta}^*+\eta_i)^2\\&=A_1+A_2+A_3,\end{aligned}$$

其中

$$A_1=\frac{1}{n}\sum_{i=1}^{n}\eta_i^2,\qquad A_2=\frac{1}{n}\sum_{i=1}^{n}[x_i^{\mathsf{T}}(\beta^*-\widehat{\beta}^*)]^2,\qquad A_3=\frac{2}{n}\sum_{i=1}^{n}x_i^{\mathsf{T}}(\beta^*-\widehat{\beta}^*)\eta_i.$$

由大数定律可知

$$A_1\to 2[(1-\alpha)^2\sigma_1^2+\alpha^2\sigma_2^2]\qquad \text{a.s.} \tag{13}$$

成立. 显然

$$A_2=\frac{1}{n}\sum_{i=1}^{n}(\beta^*-\widehat{\beta}^*)^{\mathsf{T}}x_ix_i^{\mathsf{T}}(\beta^*-\widehat{\beta}^*)=\frac{1}{n}\|A^{1/2}(\beta^*-\widehat{\beta}^*)\|^2,$$

其中"$\|\cdot\|$"表示L_2范数. 由引理3即得

$$A_2\xrightarrow{\mathsf{P}}0. \tag{14}$$

由(13)、(14)式及Cauchy不等式知

$$|A_3|\leqslant 2\sqrt{A_1A_2}\xrightarrow{\mathsf{P}}0. \tag{15}$$

故由(13)、(14)、(15)式知

$$R\xrightarrow{\mathsf{P}}2[(1-\alpha)^2\sigma_1^2+\alpha^2\sigma_2^2]. \tag{16}$$

由G的定义, 类似可证

$$G\xrightarrow{\mathsf{P}}24[(1-\alpha)^4\sigma_1^4+(1-\alpha)^2\sigma_1^2\alpha^2\sigma_2^2+\alpha^4\sigma_2^4]. \tag{17}$$

故由$\widehat{\alpha}$的定义和(16)、(17)知

$$\widehat{\alpha}\xrightarrow{\mathsf{P}}\alpha,$$

引理4成立. □

引理 5　在本文假定下, 当$n\to\infty$时, 有

$$V^{-1}A_n^{1/2}(\widehat{\beta}^*-\beta^*)\xrightarrow{L}\mathrm{N}(0,I_p),$$

其中$V=(1-\alpha)\sigma_1+\alpha\sigma_2$.

证明:　模型(3)显然满足文献[12]中定理1的条件, 故由此定理可知引理5成立. □

定理1的证明:　由(6)式及引理3和引理4知定理1成立. □

定理2的证明:　由(6)式知

$$\widehat{\beta}-\beta=\frac{\widehat{\beta}^*}{1-\widehat{\alpha}}-\beta=\frac{1}{1-\widehat{\alpha}}[\widehat{\beta}^*-(1-\alpha)\beta+(\widehat{\alpha}-\alpha)\beta].$$

由引理4有

$$\frac{1}{1-\widehat{\alpha}}\widehat{V}^{-1}A_n^{1/2}(\widehat{\alpha}-\alpha)\xrightarrow{\mathsf{P}}0.$$

再结合引理4和引理5知定理2成立. □

§4. 模拟研究

为了对本文方法的效果进行评价, 我们将本文的方法(LAD)和文献[11]中的最小二乘方法(LSE)进行比较.

我们考虑线性回归模型

$$y_i^* = (1-\alpha)(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i) + \alpha t_i, \qquad i = 1, 2, \ldots, n,$$

其中取参数真值$\beta = (\beta_0, \beta_1, \beta_2)^\intercal = (1, 1, 1)^\intercal$, $\epsilon_i \sim \text{Laplace}(0, \sigma_1)$, $t_i \sim \text{Laplace}(0, \sigma_2)$, 且$\epsilon_i, t_i$相互独立. 考虑到$\sigma_1, \sigma_2$有一个已知, 或者均未知但比值已知两种情况具有很强的类似性, 针对表1中A – D的4种实验设计方案, 我们分别在两种不同的情况(C_1: 表示σ_1, σ_2均已知, C_2: 表示$\sigma_2/\sigma_1 = l$已知但σ_1, σ_2均未知)下, 进行模拟实验. 在每种情况下, 我们均抽取1 000组样本量为100的样本进行计算, 并列出了参数估计的均值和均方误差. 由于文献[11]中的LSE要求ϵ_i和t_i的方差均已知, 故在C_2情形下我们无法计算LSE, 所以在C_2情形下我们只给出了LAD估计的结果.

表1 实验设计方案

序号	x_{i1}	x_{i2}
A	$x_{i1} \sim \text{N}(0, 2)$	$x_{i2} \sim \text{N}(0, 2)$
B	$x_{i1} \sim \text{U}(-2, 2)$	$x_{i2} \sim \text{N}(0, 2)$
C	$x_{i1} \sim \text{U}(-2, 2)$	$x_{i2} \sim \text{U}(-2, 2)$
D	$x_{i1} \sim \text{E}(1)$	$x_{i2} \sim \text{N}(0, 2)$

从表2、表3、表4、表5中可以看出:

1) 在C_1情形下, 最小一乘方法得到的估计量的偏差和均方误差均比最小二乘方法得到的估计量的要小.

2) 当$\sigma_2/\sigma_1 = l$已知但σ_1, σ_2均未知时, 最小一乘方法得到的估计量的偏差和均方误差都要比σ_1, σ_2均已知时得到的估计量的偏差和均方误差大得多. 这可能是因为在$\sigma_2/\sigma_1 = l$已知但σ_1, σ_2未知时, 要使用矩方法得到α的估计$\hat{\alpha}$, 需要用到η_i的四阶矩, 而ϵ_i和t_i均服从Laplace分布, 故η_i的分布也是一个重尾分布, 此时矩估计量$\hat{\alpha}$的偏差和均方误差均偏大, 从而影响了回归参数的估计效果.

§5. 极大似然估计

从上节的模拟结果及其分析可知, 在$\sigma_2/\sigma_1 = l$已知但σ_1, σ_2均未知的情况下, 由于矩估计方法的使用, 使得LAD估计的效果相对较差. 众所周知, 极大似然估计的效果比矩估计

表2　估计结果(数据由$\sigma_1=\sigma_2=1$, $\alpha=0.1$生成)

序号	条件	方法	$\widehat{\alpha}$		$\widehat{\beta}_0$		$\widehat{\beta}_1$		$\widehat{\beta}_2$	
			Mean	MSE	Mean	MSE	Mean	MSE	Mean	MSE
A	C_1	LSE	0.1464	0.0125	1.0745	0.0475	1.0728	0.0336	1.0735	0.0353
		LAD	0.1410	0.0118	1.0727	0.0426	1.0645	0.0288	1.0663	0.0316
B		LSE	0.1489	0.0126	1.0643	0.0542	1.0769	0.0414	1.0692	0.0307
		LAD	0.1428	0.0118	1.0569	0.0467	1.0671	0.0346	1.0630	0.0282
C		LSE	0.1541	0.0141	1.0874	0.0569	1.0836	0.0501	1.0806	0.0467
		LAD	0.1477	0.0130	1.0769	0.0456	1.0705	0.0400	1.0686	0.0373
D		LSE	0.1486	0.0128	1.0574	0.0745	1.0766	0.0350	1.0695	0.0349
		LAD	0.1425	0.0117	1.0559	0.0571	1.0664	0.0303	1.0631	0.0300
A	C_2	LSE	–	–	–	–	–	–	–	–
		LAD	0.3023	0.0498	1.3096	0.1502	1.3138	0.1347	1.3110	0.1344
B		LSE	–	–	–	–	–	–	–	–
		LAD	0.2992	0.0489	1.3105	0.1532	1.3036	0.1438	1.3068	0.1325
C		LSE	–	–	–	–	–	–	–	–
		LAD	0.2953	0.0471	1.3038	0.1489	1.3106	0.1435	1.2968	0.1335
D		LSE	–	–	–	–	–	–	–	–
		LAD	0.3039	0.0508	1.3070	0.1744	1.3201	0.1417	1.3174	0.1373

表3　估计结果(数据由$\sigma_1=\sigma_2=1$, $\alpha=0.3$生成)

序号	条件	方法	$\widehat{\alpha}$		$\widehat{\beta}_0$		$\widehat{\beta}_1$		$\widehat{\beta}_2$	
			Mean	MSE	Mean	MSE	Mean	MSE	Mean	MSE
A	C_1	LSE	0.2932	0.0090	1.0249	0.0438	1.0064	0.0233	1.0163	0.0455
		LAD	0.2890	0.0094	1.0113	0.0392	1.0009	0.0241	1.0087	0.0417
B		LSE	0.2958	0.0081	1.0143	0.0427	1.0155	0.0368	1.0078	0.0219
		LAD	0.2916	0.0085	1.0017	0.0403	1.0154	0.0349	1.0001	0.0216
C		LSE	0.2911	0.0082	0.9991	0.0403	1.0134	0.0341	1.0019	0.0343
		LAD	0.2859	0.0087	0.9919	0.0392	1.0049	0.0335	0.9985	0.0335
D		LSE	0.2950	0.0080	1.0007	0.0702	1.0109	0.0223	1.0087	0.0223
		LAD	0.2897	0.0084	0.9968	0.0649	1.0028	0.0220	1.0024	0.0224
A	C_2	LSE	–	–	–	–	–	–	–	–
		LAD	0.2982	0.0090	1.0111	0.0426	1.0158	0.0258	1.0136	0.0238
B		LSE	–	–	–	–	–	–	–	–
		LAD	0.2956	0.0098	1.0145	0.0450	1.0115	0.0358	1.0140	0.0253
C		LSE	–	–	–	–	–	–	–	–
		LAD	0.2977	0.0093	1.0183	0.0402	1.0151	0.0346	1.0130	0.0355
D		LSE	–	–	–	–	–	–	–	–
		LAD	0.2980	0.0092	0.9970	0.0676	1.0091	0.0251	1.0070	0.0260

表4 估计结果(数据由$\sigma_1 = 2$, $\sigma_2 = 1$, $\alpha = 0.1$生成)

序号	条件	方法	$\widehat{\alpha}$		$\widehat{\beta}_0$		$\widehat{\beta}_1$		$\widehat{\beta}_2$	
			Mean	MSE	Mean	MSE	Mean	MSE	Mean	MSE
A	C_1	LSE	0.1330	0.0080	1.0662	0.1019	1.0522	0.0383	1.0420	0.0362
		LAD	0.1283	0.0074	1.0578	0.0700	1.0463	0.0304	1.0427	0.0289
B		LSE	0.1354	0.0082	1.0559	0.1003	1.0645	0.0840	1.0531	0.0358
		LAD	0.1305	0.0077	1.0515	0.0725	1.0525	0.0603	1.0516	0.0298
C		LSE	0.1384	0.0090	1.0621	0.1037	1.0499	0.0839	1.0642	0.0844
		LAD	0.1331	0.0084	1.0488	0.0679	1.0435	0.0627	1.0522	0.0610
D		LSE	0.1320	0.0079	1.0333	0.1964	1.0531	0.0380	1.0457	0.0366
		LAD	0.1266	0.0074	1.0364	0.1334	1.0438	0.0318	1.0420	0.0282
A	C_2	LSE	-	-	-	-	-	-	-	-
		LAD	0.4501	0.1372	1.7049	0.7735	1.7058	0.6716	1.7178	0.6922
B		LSE	–	–	–	–	–	–	–	–
		LAD	0.4596	0.1427	1.7396	0.7065	1.7401	0.7126	1.7354	0.6683
C		LSE	–	–	–	–	–	–	–	–
		LAD	0.4517	0.1367	1.7136	0.8044	1.6953	0.7276	1.6970	0.7096
D		LSE	–	–	–	–	–	–	–	–
		LAD	0.4552	0.1412	1.7245	0.9969	1.7327	0.7097	1.7313	0.7013

表5 估计结果(数据由$\sigma_1 = 2$, $\sigma_2 = 1$, $\alpha = 0.3$生成)

序号	条件	方法	$\widehat{\alpha}$		$\widehat{\beta}_0$		$\widehat{\beta}_1$		$\widehat{\beta}_2$	
			Mean	MSE	Mean	MSE	Mean	MSE	Mean	MSE
A	C_1	LSE	0.3209	0.0083	1.0609	0.1054	1.0576	0.0483	1.0578	0.0469
		LAD	0.3146	0.0081	1.0369	0.0883	1.0485	0.0413	1.0411	0.0402
B		LSE	0.3217	0.0079	1.0511	0.1218	1.0342	0.0897	1.0512	0.0439
		LAD	0.3154	0.0077	1.0468	0.0936	1.0395	0.0729	1.0396	0.0384
C		LSE	0.3217	0.0091	1.0560	0.1236	1.0487	0.0952	1.0588	0.0992
		LAD	0.3153	0.0088	1.0446	0.1023	1.0450	0.0808	1.0418	0.0789
D		LSE	0.3201	0.0079	1.0326	0.2147	1.0530	0.0481	1.0468	0.0466
		LAD	0.3138	0.0078	1.0290	0.1667	1.0406	0.0432	1.0403	0.0395
A	C_2	LSE	–	–	–	–	–	–	–	–
		LAD	0.4526	0.0382	1.3570	0.3837	1.3460	0.2497	1.3296	0.2261
B		LSE	–	–	–	–	–	–	–	–
		LAD	0.4542	0.0377	1.3223	0.2971	1.3421	0.2759	1.3335	0.2155
C		LSE	–	–	–	–	–	–	–	–
		LAD	0.4563	0.0377	1.3452	0.3243	1.3305	0.2679	1.3482	0.2666
D		LSE	–	–	–	–	–	–	–	–
		LAD	0.4556	0.0388	1.3324	0.4323	1.3485	0.2358	1.3476	0.2254

的效果要好, 故本节考虑用极大似然方法代替矩方法估计参数α, 进一步使用(6)式即可得到回归参数β的最小一乘估计.

由η_i的密度函数和(3)可得y_i^*的密度函数为

$$f(y_i^*)=\frac{1}{2((1-\alpha)^2\sigma_1^2-\alpha^2\sigma_2^2)}\left[(1-\alpha)\sigma_1\mathrm{e}^{-|y_i^*-x_i^{\mathrm{T}}\beta^*|/[(1-\alpha)\sigma_1]}-\alpha\sigma_2\mathrm{e}^{-|y_i^*-x_i^{\mathrm{T}}\beta^*|/(\alpha\sigma_2)}\right].$$

则对数似然函数为

$$\begin{aligned}L=&-n\ln 2[(1-\alpha)^2\sigma_1^2-\alpha^2\sigma_2^2]\\&+\sum_{i=1}^{n}\ln\left[(1-\alpha)\sigma_1\mathrm{e}^{-|y_i^*-x_i^{\mathrm{T}}\beta^*|/[(1-\alpha)\sigma_1]}-\alpha\sigma_2\mathrm{e}^{-|y_i^*-x_i^{\mathrm{T}}\beta^*|/(\alpha\sigma_2)}\right].\end{aligned}\tag{18}$$

用$\widehat{\beta}^*$代替上式中的β^*, 并使用数值解法得到α的估计值, 代入(6)式中, 即可得到β的估计. 根据表1生成的数据, 再次对基于极大似然估计的LAD方法做模拟实验, 实验结果见表6和表7.

表6　估计结果(数据由$\sigma_1=1$, $\sigma_2=1$以及不同的α生成)

序号	条件	α	$\widehat{\alpha}$		$\widehat{\beta}_0$		$\widehat{\beta}_1$		$\widehat{\beta}_2$	
			Mean	MSE	Mean	MSE	Mean	MSE	Mean	MSE
A	C_1	0.1	0.1137	0.0039	1.0223	0.0207	1.0209	0.0098	1.0182	0.0088
B			0.1095	0.0037	1.0082	0.0217	1.0186	0.0167	1.0177	0.0099
C			0.1141	0.0039	1.0167	0.0200	1.0215	0.0183	1.0280	0.0165
D			0.1140	0.0038	1.0166	0.0373	1.0230	0.0107	1.0223	0.0094
A	C_2	0.1	0.1145	0.0168	1.0410	0.0494	1.0401	0.0357	1.0418	0.0367
B			0.1157	0.0164	1.0457	0.0474	1.0399	0.0460	1.0431	0.0360
C			0.1020	0.0153	1.0300	0.0467	1.0267	0.0384	1.0225	0.0416
D			0.1134	0.0169	1.0359	0.0645	1.0423	0.0365	1.0414	0.0347
A	C_1	0.3	0.3066	0.0069	1.0249	0.0373	1.0258	0.0222	1.0192	0.0218
B			0.3087	0.0067	1.0251	0.0395	1.0282	0.0323	1.0274	0.0210
C			0.3051	0.0063	1.0195	0.0392	1.0247	0.0307	1.0277	0.0341
D			0.3064	0.0069	1.0163	0.0615	1.0254	0.0228	1.0242	0.0213
A	C_2	0.3	0.2574	0.0275	0.9897	0.0679	0.9911	0.0499	0.9854	0.0493
B			0.2620	0.0291	0.9910	0.0703	1.0060	0.0646	0.9954	0.0531
C			0.2648	0.0281	0.9856	0.0636	0.9976	0.0631	0.9980	0.0641
D			0.2623	0.0270	1.0016	0.0912	0.9776	0.1519	0.9958	0.0502

比较表6–7和表2–5中相应的结果可以看出, 在绝大多数情况下, 基于极大似然的LAD方法比基于矩估计的LAD方法得到估计的偏差要明显偏小, 与预期相符合, 这说明使用极大似然的方法估计α确实可以有效地降低回归参数估计量的偏差.

表7 估计结果(数据由$\sigma_1 = 2$, $\sigma_2 = 1$以及不同的α生成)

序号	条件	α	$\widehat{\alpha}$		$\widehat{\beta}_0$		$\widehat{\beta}_1$		$\widehat{\beta}_2$	
			Mean	MSE	Mean	MSE	Mean	MSE	Mean	MSE
A	C_1	0.1	0.1140	0.0036	1.0192	0.0675	1.0163	0.0199	1.0132	0.0202
B			0.1057	0.0019	1.0281	0.0570	1.0018	0.0443	1.0050	0.0171
C			0.1135	0.0034	1.0163	0.0638	1.0149	0.0474	1.0153	0.0485
D			0.1126	0.0034	1.0222	0.1148	1.0207	0.2717	1.0180	0.0199
A	C_2	0.1	0.0791	0.0121	0.9951	0.0811	1.0002	0.0390	0.9882	0.0372
B			0.0830	0.0128	1.0050	0.0794	0.9947	0.0619	1.0059	0.0407
C			0.0773	0.0126	0.9911	0.0881	0.9928	0.0702	0.9942	0.0635
D			0.0824	0.0131	0.9992	0.1484	1.0032	0.0425	1.0023	0.0409
A	C_1	0.3	0.3084	0.0046	1.0251	0.0809	1.0186	0.0289	1.0191	0.0266
B			0.3101	0.0046	1.0266	0.0846	1.0111	0.0632	1.0218	0.0290
C			0.3099	0.0046	1.0440	0.0804	1.0251	0.0643	1.0412	0.0598
D			0.3098	0.0047	1.0203	0.1365	1.0098	0.2994	1.0245	0.0301
A	C_2	0.3	0.2960	0.0551	1.1283	0.2447	1.1163	0.1802	1.1259	0.1909
B			0.2856	0.0527	1.0798	0.2303	1.1083	0.2161	1.0953	0.1712
C			0.2781	0.0536	1.0761	0.2220	1.0767	0.2022	1.0808	0.1965
D			0.2827	0.0521	1.1173	0.3406	1.0470	0.5477	1.0939	0.1594

参 考 文 献

[1] Wedel M, DeSarbo W S. A latent class binomial logit methodology for the analysis of paired comparison choice data [J]. *Decision Sci.*, 1993, **24(6)**: 1157–1170.

[2] Frühwirth-Schnatter S. Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models [J]. *J. Amer. Statist. Assoc.*, 2001, **96(453)**: 194–209.

[3] Wang P M, Puterman M L, Cockburn I, et al. Mixed Poisson regression models with covariate dependent rates [J]. *Biometrics*, 1996, **52(2)**: 381–400.

[4] Green P J, Richardson S. Hidden Markov models and disease mapping [J]. *J. Amer. Statist. Assoc.*, 2002, **97(460)**: 1055–1070.

[5] Davis D J. An analysis of some failure data [J]. *J. Amer. Statist. Assoc.*, 1952, **47(258)**: 113–150.

[6] Huber P J. Robust estimation of a location parameter [J]. *Ann. Math. Statist.*, 1964, **35(1)**: 73–101.

[7] Yu K F. A note on the estimation of the mixing parameter in a mixture of two distributions [J]. *Comm. Statist. Theory Methods*, 1991, **20(2)**: 595–609.

[8] Copas J B. Binary regression models for contaminated data [J]. *J. Roy. Statist. Soc. Ser. B*, 1988, **50(2)**: 225–265.

[9] 郑祖康, 丁邦俊, 杨瑛, 等. 关于两类污染数据回归分析的参数估计 [J]. 高校应用数学学报, 1996, **11A(1)**: 31–40.

[10] 任哲, 陈明华. 污染数据回归分析中参数的最小一乘估计 [J]. 应用概率统计, 2000, **16(3)**: 262–268.

[11] 陈明华. 污染数据回归分析中估计的强相合性 [J]. 应用概率统计, 1998, **14(1)**: 73–78.

[12] 陈希孺, 白志东, 赵林城, 等. 线性模型中最小一乘估计的渐近正态性 [J]. 中国科学A辑, 1990, **20(5)**: 449–463.

LAD Estimation for Linear Regression Models with Contaminated Data

YE Peng ZHOU XiuQing

(*School of Mathematical Sciences and Institute of Finance and Statistics, Nanjing Normal University, Nanjing, 210023, China*)

Abstract: In this paper, linear regression models with contaminated data are considered. Estimation methods for the regression parameters based on least absolute deviations (LAD) are proposed, and properties of consistency and asymptotic normality of the proposed method are proved under some regular conditions. Simulations are done to assess the properties of the method when sample size is small, and simulation results show that the methods works well.

Keywords: contaminated data; least absolute deviations estimation; consistency; asymptotic normality

2010 Mathematics Subject Classification: 62N02