

稳健经验似然估计方法

苏依官, 樊亚莉, 徐孝琳

(上海理工大学 理学院, 上海 200093)

摘要: 在经验似然方法的基础上提出了一种稳健的经验似然估计方法, 在约束条件的估计方程中使用权重函数以及有界得分来限制异常点的影响。通过数值模拟研究该方法的稳健性。模拟结果表明, 相比普通的经验似然估计, 提出的稳健经验似然方法在数据中存在异常值的情况下所得估计的均方误差更小。同时对于非正态的厚尾分布数据, 提出的稳健经验似然估计也在均方误差意义下更优。

关键词: 经验似然; 稳健性; 估计方程

中图分类号: O 212.1 **文献标志码:** A

Robust Empirical Likelihood Estimation

SU Yiguan, FAN Yali, XU Xiaolin

(College of Science, University of Shanghai for Science and Technology, Shanghai 200093, China)

Abstract: Based on introducing the empirical likelihood method, a robust empirical likelihood estimator was presented. The weight function and the bounded score were used to limit the influence of outlier on the estimation equation with the constraint condition. The performance of the proposed estimator was studied through simulation. The simulation results show that the proposed robust empirical likelihood estimator has smaller mean square error, compared to ordinary empirical likelihood estimators. At the same time, the proposed estimator also performs better for heavy tailed data sets in the term of mean square error.

Keywords: empirical likelihood; robustness; estimation equation

普通最小二乘(OLS)估计和最大似然(ML)估计是线性回归模型^[1]中常用的参数估计方法。这些方法在分布假设合理的情况下表现良好, 但是, 当实际的数据不一定满足这些假设的情况下, 这些方法所得到的估计会有较大偏差, 这时需要考虑其他非参数估计方法。经验似然(EL)方法就是

一种非参数推断方法, 它是由 Owen 于 1988 年在完全样本下提出的一种统计推断方法^[2], 它有类似于 boot-strap 方法的抽样特性。这一方法与经典的统计方法相比优点突出, 当用经验似然方法构造置信区间时, 它具有域保持性、变换不变性、Bartlett 纠偏性等诸多优点。同时, 经验似然置信

收稿日期: 2018-04-19

基金项目: 国家自然科学基金资助项目 (11401383)

第一作者: 苏依官(1994-), 男, 硕士研究生. 研究方向: 概率论与数理统计. E-mail: 1135883358@qq.com

通信作者: 樊亚莉(1978-), 女, 讲师. 研究方向: 概率论与数理统计. E-mail: yalifan@usst.edu.cn

域的形状只由样本决定,无需构造枢轴统计量^[2-3]。正因为如此,经验似然统计方法引起了许多学者的重视,他们将这一方法应用到各种统计模型及统计推断问题中,如总体均值的估计、线性回归模型的拟合、分位数的估计^[4]、估计方程^[5]的建立,同时它也可以在不完全的删失数据^[3]中得到很好的应用。实际上,EL方法的本质是带约束的最优化问题,在经验似然方法中,对每个观测值赋予一个权重,EL估计就是最大化权重乘积,而人们感兴趣的参数一般体现在该最大化问题的约束条件中。

众所周知,OLS,ML估计对方对异常值较为敏感,当数据中存在异常值时,用普通的OLS和ML估计方法所得估计会有较大的偏差和方差,而稳健的估计方法恰好可以弥补这一点,有许多学者讨论了稳健估计的重要性以及如何构造稳健估计量^[6-10]。它可以在异常值存在的情况下,选择适当的抑制异常值影响的技术,使参数估值尽可能不受其影响,得出正常模式下的最优或接近最优的参数估值。由于EL方法目标函数的约束条件继承了OLS估计方程的原理,而异常值的存在也可能对EL估计产生较大的影响,因此,如何改进EL方法约束条件中的估计方程使之能够有效抑制异常值的影响,这是一个值得研究的问题。文献^[11]考虑将稳健估计方程和经验似然方法结合起来,构造稳健的经验似然估计,但是,他们的估计方程只考虑了限制因变量中的异常值,对自变量中可能存在的异常值并没有采取相应的措施。

本文考虑稳健的经验似然估计方法,基于经验似然方法,在约束条件的估计方程中用一个依赖于自变量的权重函数使得杠杆点对估计方程的影响降低,并且对残差项使用一个有界的得分函数来限制因变量中异常点的影响,从而在数据分布来源未知的情况下有效地抑制可能的异常值对估计的影响。

1 模型和方法

1.1 普通经验似然估计方法

考虑一般线性模型

$$Y = X^T \beta + \varepsilon \quad (1)$$

式中: $X = (X_1, X_2, \dots, X_n)$ 为自变量, $X_i = (X_{i1}, X_{i2}, \dots, X_{ip})^T$; $Y = (Y_1, Y_2, \dots, Y_n)^T$ 为因变量; 随机

误差项 $\varepsilon = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)^T$, 假设 $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ 独立同分布, 且 $E(\varepsilon_i) = 0$, $\text{Var}(\varepsilon_i) = \sigma^2$, $i = 1, 2, \dots, n$, σ^2 可以用非稳健与稳健估计方式分别对应为

$$\hat{\sigma} = \sqrt{\frac{1}{n} \sum_{i=1}^n (Y_i - X_i^T \hat{\beta}^*)^2}$$

$$\hat{\sigma}_R = 1.483 \text{ median} \left\{ \left| (Y_i - X_i^T \hat{\beta}^*) - \text{median}(Y_i - X_i^T \hat{\beta}^*) \right| \right\}$$

这里 $\hat{\beta}^* = (X^T X)^{-1} X^T Y$ 即为最小二乘估计。

对线性模型(1), EL估计是下面优化问题的最优解^[4,12]。

最大化目标函数

$$L(\beta) = \prod_{i=1}^n p_i$$

其约束条件为

$$p_i \geq 0, \quad \sum_{i=1}^n p_i = 1, \quad \sum_{i=1}^n p_i X_i (Y_i - X_i^T \beta) = 0$$

记

$$g_i(\beta) = X_i (Y_i - X_i^T \beta)$$

建立普通的经验似然(EL)函数

$$L(\beta) = \sup \left\{ \prod_{i=1}^n p_i \mid \sum_{i=1}^n p_i = 1, \sum_{i=1}^n p_i g_i(\beta) = 0, p_i \geq 0 \right\} \quad (2)$$

于是, $L(\beta)$ 的最大值点 β 就是 β 的普通经验似然估计。

1.2 稳健的经验似然估计方法

考虑到实际数据受各种因素的影响,会有不同程度的污染,若仅按式(2)估计 β , 势必会有较大的偏差。于是,在经验似然估计的基础上进行稳健化处理^[13-14],即稳健的经验似然估计。

a. 定义有界得分函数

$$\psi(x) = \begin{cases} x, & |x| < c \\ c, & x \geq c \\ -c, & x \leq -c \end{cases} \quad (3)$$

式中, c 为可调节的参数,一般介于 1~2 之间。

b. 定义权重函数

$$w_i = \min \left\{ 1, \left[\frac{p_0}{(X_i - m_x)^T S_x^{-1} (X_i - m_x)} \right]^{\frac{r}{2}} \right\} \quad (4)$$

其中, p_0 取 $\chi_{0.95}^2(p)$, 即自由度为 p 的卡方分布的 0.95 分位数, $r \geq 1$, r 为控制模型稳健的力度, m_x 为 X 按行取中位数所得的 p 维向量。

$$S_x = \text{diag}(S_1, S_2, \dots, S_k, \dots, S_p)$$

$$S_k = 1.483 (\text{median}(|X_{1k} - m_x(k)|, |X_{2k} - m_x(k)|, \dots, |X_{nk} - m_x(k)|), k = 1, 2, \dots, p$$

$m_x(k)$ 为 $\mathbf{m}_x$ 的第 k 个分量。对于 $\mathbf{X}_i$ 与整体 $\mathbf{X}$ 偏离较大的情况, $(\mathbf{X}_i - \mathbf{m}_x)^T \mathbf{S}_x^{-1} (\mathbf{X}_i - \mathbf{m}_x)$ 会较大, 因此, $\left[\frac{p_0}{(\mathbf{X}_i - \mathbf{m}_x)^T \mathbf{S}_x^{-1} (\mathbf{X}_i - \mathbf{m}_x)} \right]^{\frac{r}{2}}$ 会较小, 显然, w_i 的值就小了, 即该 $\mathbf{X}_i$ 的对应的权重小了; 相反, 对于 $\mathbf{X}_i$ 与整体 $\mathbf{X}$ 偏离较小的情况, w_i 会较大, 这时 w_i 往往取 1。那么, 通过该函数可对 $\mathbf{X}$ 中可能的异常数据实现稳健化处理。通过结合以上 2 个稳健化函数得到稳健的经验似然函数

$$L_R(\boldsymbol{\beta}) = \sup \left\{ \prod_{i=1}^n p_i \left| \sum_{i=1}^n p_i = 1, \sum_{i=1}^n p_i g_{i,R}(\boldsymbol{\beta}) = 0, \right. \right. \\ \left. \left. g_{i,R}(\boldsymbol{\beta}) = X_i W_i \psi \left((Y_i - X_i^T \boldsymbol{\beta}) / \sigma \right), p_i \geq 0 \right\} \quad (5)$$

于是, $L_R(\boldsymbol{\beta})$ 的最大值点 $\hat{\boldsymbol{\beta}}_R$ 就是 $\boldsymbol{\beta}$ 的稳健经验似然估计。

2 理论性质

现针对提出的稳健经验似然估计解的存在性以及估计量、统计量的性质给出一些理论证明, 证明方法类似于文献[4]。记 $R_i = g_{i,R}(\boldsymbol{\beta})$, $R = g_R(\boldsymbol{\beta})$, 定义

$$l_E(\boldsymbol{\beta}) = \sum_{i=1}^n \log(n(1 + \lambda_1^T R_i))$$

显然, 上式的最小值点为式(5)的解^[15]。

引理 1(解的存在性) 假设 $E[\mathbf{R}\mathbf{R}^T]$ 是正定的, $\partial \mathbf{R} / \partial \boldsymbol{\beta}$ 关于 $\boldsymbol{\beta}$ 在真实值 $\boldsymbol{\beta}_0$ 邻域内是连续的, $\|\partial \mathbf{R} / \partial \boldsymbol{\beta}\|$ 和 $\|\mathbf{R}\|^3$ 在该邻域内受某些可积函数 $G(x)$ 的约束, 并且 $E(\partial \mathbf{R} / \partial \boldsymbol{\beta})$ 的秩为 p 。那么, 随着 n 趋向于无穷, $l_E(\boldsymbol{\beta})$ 在球 $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \leq n^{-1/3}$ 内部某些点 $\hat{\boldsymbol{\beta}}_R$ 以概率 1 达到最小值。

定理 1(渐近正态性) 基于引理 1 的条件, 假设 $\partial^2 \mathbf{R} / \partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T$ 关于 $\boldsymbol{\beta}$ 在真实值 $\boldsymbol{\beta}_0$ 邻域内是连续的, 那么, 如果 $\|\partial^2 \mathbf{R} / \partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T\|$ 在该邻域内受某些可积

$$D\mathbf{G}_R(\mathbf{A}) = \begin{bmatrix} 1 & 1 & \cdots & 1 & \mathbf{O}_{1 \times p} \\ X_1\{Y_1 - X_1^T \boldsymbol{\beta}\} & X_2\{Y_2 - X_2^T \boldsymbol{\beta}\} & \cdots & X_n\{Y_n - X_n^T \boldsymbol{\beta}\} & - \left[\sum_{i=1}^n p_i X_i X_i X_i^T \right]^T \end{bmatrix}$$

那么, 式(7)的 Hessian 矩阵可表示为

$$H\mathbf{L}(\mathbf{A}, \mathbf{A}) = \begin{bmatrix} \mathbf{B}(\mathbf{A}, \mathbf{A}) & D\mathbf{G}(\mathbf{A})^T \\ D\mathbf{G}(\mathbf{A}) & \mathbf{O}_{p \times p} \end{bmatrix}$$

这里

函数 $G(x)$ 的控制, 则有

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_R - \boldsymbol{\beta}_0) \rightarrow N(0, V)$$

这里

$$V = \left[E \left(\frac{\partial \mathbf{R}}{\partial \boldsymbol{\beta}} \right)^T (E(\mathbf{R}\mathbf{R}^T))^{-1} E \left(\frac{\partial \mathbf{R}}{\partial \boldsymbol{\beta}} \right) \right]^{-1}$$

3 算 法

3.1 普通经验似然估计求解

式(2)的解等价于如下目标函数的解^[16]:

$$L(p, \boldsymbol{\beta}) = - \min_{p_i \in (0,1)} \left[\sum_{i=1}^n \ln p_i \right] \quad (6)$$

满足约束条件 $\sum_{i=1}^n p_i = 1, \sum_{i=1}^n p_i X_i \{Y_i - X_i^T \boldsymbol{\beta}\} = 0$ 。

对于一般的带约束的函数求极值普遍采用拉格朗日乘数法, 得到下式:

$$L(\mathbf{p}, \boldsymbol{\beta}, \lambda_0, \lambda_1^T) = - \sum_{i=1}^n \ln p_i + \lambda_0 \left(\sum_{i=1}^n p_i - 1 \right) - \\ \lambda_1^T \sum_{i=1}^n p_i g_i(\boldsymbol{\beta}) \quad (7)$$

其中, $\mathbf{p} = [p_1, p_2, \cdots, p_n]$, $\lambda_0 \in R^1$, $\lambda_1^T \in R^p$, 均为拉格朗日乘子, 记

$$\mathbf{A} = \begin{bmatrix} \mathbf{p} \\ \boldsymbol{\beta} \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} \lambda_0 \\ \lambda_1^T \end{bmatrix}$$

式(7)的一阶梯度为

$$\nabla L(\mathbf{A}, \mathbf{A}) = \begin{bmatrix} \nabla L(\mathbf{A}) + (D\mathbf{G}(\mathbf{A}))^T \mathbf{A} \\ \mathbf{G}(\mathbf{A})^T \end{bmatrix} \quad (8)$$

其中,

$$\nabla L(\mathbf{A}) = \left[-\frac{1}{p_1}, -\frac{1}{p_2}, \cdots, -\frac{1}{p_n}, \mathbf{O}_{1 \times p} \right]^T \\ \mathbf{G}(\mathbf{A}) = \left[\sum_{i=1}^n p_i - 1, \sum_{i=1}^n p_i X_i \{Y_i - X_i^T \boldsymbol{\beta}\} \right]$$

$\mathbf{G}(\mathbf{A})$ 关于 $\mathbf{A}$ 的 Jacobi 矩阵为

$$B(A,\Lambda)=\begin{bmatrix}\frac{1}{p_1^2}&0&\cdots&0&-\lambda_1^T X_1 X_1^T\\0&\frac{1}{p_2^2}&\cdots&0&-\lambda_1^T X_2 X_2^T\\\vdots&\vdots&&\vdots&\vdots\\0&0&\cdots&\frac{1}{p_n^2}&-\lambda_1^T X_n X_n^T\\-\left[\lambda_1^T X_1 X_1^T\right]^T&-\left[\lambda_1^T X_2 X_2^T\right]^T&\cdots&-\left[\lambda_1^T X_n X_n^T\right]^T&\boldsymbol{O}_{p\times p}\end{bmatrix}$$

为一个 $(n+p)\times(n+p)$ 矩阵。

根据牛顿迭代原理，可以得到

$$\boldsymbol{H}\boldsymbol{L}(A_t,\Lambda_t)\begin{bmatrix}A_{t+1}-A_t\\ \Lambda_{t+1}-\Lambda_t\end{bmatrix}+\nabla L(A_t,\Lambda_t)=0$$

这里， A_t,Λ_t 的值表示在第 t 次迭代 $\boldsymbol{A},\boldsymbol{\Lambda}$ 的当前值， A_{t+1},Λ_{t+1} 的值是由当前 A_t,Λ_t 计算得到。

可得 $\boldsymbol{A},\boldsymbol{\Lambda}$ 的迭代式

$$\begin{bmatrix}A_{t+1}\\\Lambda_{t+1}\end{bmatrix}=\begin{bmatrix}A_t\\\Lambda_t\end{bmatrix}-\boldsymbol{H}\boldsymbol{L}(A_t,\Lambda_t)^{-1}\nabla L(A_t,\Lambda_t)\tag{9}$$

该算法的步骤可以归结如下：

- a. 设定 $p,\beta,\lambda_0,\lambda_1$ 的初值和阈值 ξ 。取 $p=(\frac{1}{n},\frac{1}{n},\cdots,\frac{1}{n})$ ， $\beta=(0,0,\cdots,0)$ ， $\lambda_0=0.001$ ， $\lambda_1=(0.001,0.001,\cdots,0.001)$ 作为初值， ξ 取 10^{-5} 。
- b. 对于 $t=1,2,3\cdots$ ，计算 $\boldsymbol{H}\boldsymbol{L}(A_t,\Lambda_t)$ 和 $\nabla L(A_t,\Lambda_t)$ 。
- c. 根据迭代式(9)计算 A_{t+1},Λ_{t+1} 。
- d. 重复上述步骤直到满足 $\|\beta_{t+1}-\beta_t\|<\xi$ 。

$$D\boldsymbol{G}_R(A)=\begin{bmatrix}1&1&\cdots&1&\boldsymbol{O}_{1\times p}\\X_1w_1\psi(\{Y_1-X_1^T\beta\})&X_2w_2\psi(\{Y_2-X_2^T\beta\})&\cdots&X_nw_n\psi(\{Y_n-X_n^T\beta\})&-\left[\sum_{i=1}^np_iw_iX_iX_i^T\right]^T\end{bmatrix}$$
$$B_R(A,\Lambda)=\begin{bmatrix}\frac{1}{p_1^2}&0&\cdots&0&-\lambda_1^Tw_1X_1X_1^T\\0&\frac{1}{p_2^2}&\cdots&0&-\lambda_1^Tw_2X_2X_2^T\\\vdots&\vdots&&\vdots&\vdots\\0&0&\cdots&\frac{1}{p_n^2}&-\lambda_1^Tw_nX_nX_n^T\\-\left[\lambda_1^Tw_1X_1X_1^T\right]^T&-\left[\lambda_1^Tw_2X_2X_2^T\right]^T&\cdots&-\left[\lambda_1^Tw_nX_nX_n^T\right]^T&\boldsymbol{O}_{p\times p}\end{bmatrix}$$

其算法迭代步骤同前。

4 模拟研究

为了研究稳健经验似然估计方法在有限样本下的表现效果，通过模拟试验将稳健方法与非稳健方法以及最小二乘方法进行比较。

4.1 数据的产生

从模型(1)产生数据，这里取 $\beta=(\beta_1,\beta_2,\beta_3)^T=$

3.2 稳健的经验似然估计求解

由于稳健方法与非稳健方法的区别只在估计方程中，因此，算法类似，只是将 $\boldsymbol{H}\boldsymbol{L}(A,\Lambda)$ 和 $L(A,\Lambda)$ 变成 $\boldsymbol{H}\boldsymbol{L}_R(A,\Lambda)$ 和 $L_R(A,\Lambda)$ 。其拉格朗日函数为

$$L_R(p,\beta,\lambda_0,\lambda_1^T)=\sum_{i=1}^n\ln p_i+\lambda_0\left(\sum_{i=1}^np_i-1\right)-\lambda_1^T\sum_{i=1}^np_i g_{i,R}(\beta)$$

新的迭代式为

$$\begin{bmatrix}A_{t+1}\\\Lambda_{t+1}\end{bmatrix}-\begin{bmatrix}A_t\\\Lambda_t\end{bmatrix}=-\boldsymbol{H}\boldsymbol{L}_R(A_t,\Lambda_t)^{-1}\nabla L_R(A_t,\Lambda_t)\tag{10}$$

在 $\boldsymbol{H}\boldsymbol{L}_R(A_t,\Lambda_t)$ 与 $L_R(A_t,\Lambda_t)$ 中，

$$G_R(A)=\left[\sum_{i=1}^np_i-1,\sum_{i=1}^np_iw_i\psi\left(\left(Y_i-X_i^T\beta\right)/\sigma\right)\right]$$

$(1,2,3)^T,n=100,p=3,x_{i1},x_{i2},\cdots,x_{ip}$ 从 $N(0,1)$ 中抽取，且它们之间独立。

对于 ε_i 作2组试验。

试验1 对于一般的参数估计方法，误差项往往取自标准正态分布，为了说明本文所述方法同样适用，在试验1中 ε_i 取自 $N(0,1)$ 。

试验2 考虑到本文的估计方法是一个非参数方法，不同于参数的方法，其对误差项分布来源没有太严格的要求，为了验证这一点，在试验

2 中 ε_i 取自 $t(1)$ 。

在模拟中, 将式(4)权重函数 w 中的 r 取 1, 结合表达式特点, 此时使得 $\mathbf{x}$ 异常值对估计的影响最小; 式(3) ψ 函数中的 c 一般在 1~2 之间取值, 本文取 1.5。

4.2 数据污染与方式

为了考察本文的稳健方法的优势, 对于试验 1 通过几种污染数据的方式产生数据中的异常值。

污染 1 随机挑选试验 1 中 4% 的 y_i 将其替换成 $y_i + 4$ 。

污染 2 随机挑选试验 1 中 2% 的 x_i 将其替换成 $x_i + 2$ 。

污染 3 污染 1, 污染 2 同时进行。

4.3 结果与分析

重复上述模拟, $N = 100$, 即可得到 β 的 100 次估计, 通过估计的偏差(Bias)、方差(Var)、均方误差(Mes)来衡量估计的优劣。

在表 1 中, NR 表示在无污染下的非稳健方法, R 表示在无污染下的稳健方法, NR-C1 表示在污染 1 下的非稳健方法, R-C1 表示在污染 1 下的稳健方法; 在表 2 中, OSL- $t(1)$ 表示在残差项服从 $t(1)$ 厚尾分布下的最小二乘法方法, 其余符号的含义类似。记录下 3 个分量 $\beta_1, \beta_2, \beta_3$ 在各指标下的结果, 每个指标下各设置 1 个综合量, 利用 3 个结果的二范数来更直观地比较各种估计方法的优劣。

表 1 试验 1 的模拟结果
Tab.1 Simulation results of experiment 1

统计量		方法与污染							
		NR	R	NR-C1	R-C1	NR-C2	R-C2	NR-C3	R-C3
Bias	分量 1	0.003	0.003	-0.003	-0.002	-0.367	-0.057	-0.259	-0.062
	分量 2	0.007	0.009	-0.016	-0.010	-0.405	-0.068	-0.278	-0.085
	分量 3	-0.004	-0.003	-0.015	-0.014	-0.390	-0.068	-0.262	-0.063
	综合	0.009	0.010	0.022	0.017	0.671	0.112	0.461	0.123
Var	分量 1	0.009	0.010	0.016	0.011	0.032	0.010	0.020	0.012
	分量 2	0.009	0.009	0.014	0.010	0.029	0.011	0.022	0.014
	分量 3	0.014	0.014	0.016	0.012	0.033	0.011	0.022	0.013
	综合	0.019	0.020	0.027	0.020	0.054	0.019	0.038	0.022
mes	分量 1	0.009	0.010	0.016	0.011	0.167	0.013	0.087	0.016
	分量 2	0.009	0.009	0.015	0.011	0.192	0.016	0.100	0.021
	分量 3	0.014	0.014	0.016	0.012	0.185	0.016	0.091	0.017
	综合	0.019	0.020	0.027	0.020	0.315	0.026	0.161	0.031

结论:

a. 稳健与非稳健方法处理无污染数据的情况。对于误差项来自 $N(0,1)$ 的情况, 大体上稳健方法的 3 个指标的综合量略微大于非稳健方法的, 因为, 稳健方法损失了数据中的一些信息, 这完全是正常的。对于误差项来自 $t(1)$ 的情况, 稳健方法的 3 个指标的综合量均大幅小于非稳健方法的, 在表 2 中, $\hat{\beta}, \hat{\beta}_R$ 的偏差分别为 0.777, 0.061, 方差分别为 43.996, 0.381, 均方误差分别为 44.402, 0.384, 说明稳健方法的优势尤为明显。

b. 从偏差角度观察, 有污染或是误差项来自

$t(1)$ 的情况, 稳健方法的该指标各个分量基本都小于非稳健方法, 从而确保估计更加接近无偏。

c. 从方差角度观察, 污染 1(仅对 $\mathbf{X}$ 污染的情况), 以及误差项来自 $t(1)$ 的情况, 稳健的方法的该指标各个分量基本都小于非稳健方法的。而对于污染 2、污染 3(即包含对 $\mathbf{x}$ 的污染), 大体上稳健方法的该指标各个分量基本只是略微大于非稳健方法的, 说明该稳健方法在 $\mathbf{x}$ 受污染时对方差的控制也比较稳定。

d. 从均方误差角度观察, 除了无污染且误差项来自 $N(0,1)$ 的情况, 稳健方法的该指标各个分量

表 2 试验 2 的模拟结果

Tab.2 Simulation results of experiment 2

统计量		方法与污染		
		OLS-t(1)	NR-t(1)	R-t(1)
Bias	分量 1	-10	0.345	0.055
	分量 2	-28	-0.623	0.017
	分量 3	-13	-0.309	-0.021
	综合	33	0.777	0.061
Var	分量 1	9 944	19.715	0.220
	分量 2	69 405	36.130	0.236
	分量 3	16 254	15.542	0.204
	综合	71 973	43.996	0.381
Mse	分量 1	10 046	19.834	0.223
	分量 2	70 206	36.519	0.237
	分量 3	16 433	15.638	0.204
	综合	72 800	44.402	0.384

基本都小于非稳健方法的，这主要是偏差小决定的。

e. 在误差项来自 $N(0,1)$ 的情况，最小二乘方法与经验似然方法的比较。在无污染、污染 1、污染 2、污染 3 情况下最小二乘估计的均方误差分别为 0.018, 0.028, 0.315, 0.163。通过与表 1 数据进行比较，对于无污染的情况，最小二乘方法均方误差小于稳健与非稳健经验似然方法；对于有污染的情况，其均方误差会略大于 2 种经验似然方法。

5 实证分析

将本文提出的稳健经验似然估计方法应用到全国各地区房价预测中，这批数据共收集了 35 个地区在 2014 年的住宅商品房平均销售价格及在岗职工平均工资、生产总值、住宅房屋竣工面积 3 个可能的影响因素数据。其数据来源于国家统计局及各城市的统计年鉴 <http://www.stats.gov.cn/>。

将每个地区看作一次观测，每次观测有 4 个数据，将观测到的住宅商品房平均销售价格作为因变量，将剩下的 3 个观测指标作为基本自变量，在此基础上添加 3 个自变量平方项、交叉项作为生成自变量。对于交叉项，经实际数据的回归分析只取在岗职工平均工资与生产总值乘积项。

用 Y_i 表示第 i 个地区的因变量观测值， X_{ij} 表示第 i 个地区的第 j 个指标的观测值， $\beta_1, \cdots, \beta_7$ 分别代

表在岗职工平均工资、生产总值、住宅房屋竣工面积、在岗职工平均工资平方项、生产总值平方项、在岗职工平均工资与生产总值乘积项、住宅房屋竣工面积平方项的系数，截距项为 β_0 ，对该批数据拟合的线性回归模型为

$$Y_i = \beta_0 + \sum_{j=1}^7 X_{ij}\beta_j + \varepsilon_i, \quad i = 1, 2, \cdots, 35 \quad (11)$$

式中， ε_i 为独立同分布的随机误差项，假定 $E(\varepsilon_i) = 0$ 。

考虑到实际数据各项指标单位不同，数值差异较大，因此，先将各项指标标准化，使其均值为 0，方差为 1。现将本文的稳健经验似然估计方法与非稳健经验似然方法应用到这批数据中。为了考察本文方法的稳健性，进一步用交叉验证方法比较稳健与非稳健方法的两种表现，每次剔除 1 个地区数据，用剩下 34 个地区数据估计式(11)中的回归系数，并用 CV 来衡量两种方法在交叉验证过程的稳健性，定义

$$CV = \frac{1}{35} \sum_{i=1}^{35} |Y_i - X_i \hat{\beta}_{(-i)}|^2$$
$$X_i = (X_{i1}, X_{i2}, \cdots, X_{i7})$$

式中， $\hat{\beta}_{(-i)}$ 表示去掉第 i 个地区之后回归估计的系数。

在表 3 中，NR 为普通经验似然估计，R 为稳健经验似然估计，VNR 为普通经验似然估计的方差，VR 为稳健经验似然估计的方差。从表 3 数据可以看出，稳健经验似然估计在稳定性方面优于普通经验似然估计，这一点可以通过对比 VNR 与 VR 在各系数下的值来得到。

表 3 稳健与非稳健方法对房价数据的结果分析

Tab.3 Analysis results on the price data by the robust and unsteady methods

方法	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_0	CV
NR	-1.71	3.71	-0.75	3.84	1.56	-6.03	0.05	0.00	19.97
R	-0.89	2.74	-0.60	2.39	1.36	-4.29	-0.04	-0.02	18.66
VNR	1.75	1.96	0.34	2.96	1.26	4.14	0.28	0.09	-
VR	1.74	1.93	0.29	2.84	1.28	4.03	0.25	0.08	-

观察前 3 个比较直观的变量，分别为在岗职工平均工资、生产总值、住宅房屋竣工面积。从经济学角度来说，供给和需求是决定商品价格的 2 个方面。住宅房屋竣工面积是通过供给方面来影响房价的。从经济意义上讲，住宅房屋竣工面积

越多,说明对商品房的供给越多,房价的增长速度应该放缓,直至下降。因此, β_3 的理论值应是负的,本文的估计结果与之相符。在岗职工平均工资、生产总值是通过需求方面来影响房价的。人均可支配收入决定个人消费水平的高低,随着收入水平的提高,消费能力也会相应提高,增加的居民可支配收入会提高房屋的销售价格。收入的上涨可以收缩买房时间,这直接影响房地产的销售量和销售额。人均可支配收入较高的城市,其居民有相对充足的资金购房,其房价自然较高;反之,若房价相同,人均可支配收入较低城市的居民没有足够的资金购房,房价无疑会下降。因此, β_1 的理论值应是正的,由于受到部分地区的异常值影响,导致经验似然估计结果较差,为-1.71,而本文的稳健方法可以较大程度地减小这种影响,其估计结果为-0.89。房价是商品价格的体现,当生产总值增长时,房价也会随之增长。日本在1991年之前长时间的经济增长与房地产泡沫可以体现这点。美国的经济飞速增长,房价也随之增长。反之,如我国东三省因为依赖的重工业发展缓慢,导致生产总值增长缓慢,人口外流,从而间接导致房价下降。因此, β_2 的理论值应是正的。

CV 值越小,表示该方法稳健效果越好。计算非稳健方法的 CV 值为19.97,稳健方法的 CV 值为18.66,小于前者,因此,稳健方法得到的结果比较稳定,同时也说明了数据中存在某些异常值。实际上,杭州、北京、深圳等几个城市的房价确实存在过高现象,这与本文的分析结果相符。

6 总结与展望

基于稳健估计方程和经验似然方法,提出稳健经验似然估计,研究了相应的算法,并作了大量模拟研究。研究表明,当数据中不含异常值时,稳健经验似然估计的效果和普通的经验似然估计的效果是相近的,但是,当数据中含有异常值时,稳健经验似然估计的效率明显高于普通的经验似然估计的效率。随着污染力度和污染比例的增加,稳健经验似然估计方法的优势更为明显。同时,模拟表明,作为一种非参数估计方法,本文的稳健经验似然估计适用于分析非正态的数据。因此,对于实际生活中的数据分析,建议使用稳健经验似然估计。

在本文研究的基础上,进一步还可以考虑将稳健经验似然估计方法运用到纵向数据分析中,同时还可以考虑在稳健估计的基础上作变量选择等,这些问题都值得进一步研究。

参考文献:

- [1] 陈希孺,王松桂.近代线性回归——原理方法及应用[M].合肥:安徽教育出版社,1987.
- [2] OWEN A B. Empirical likelihood ratio confidence intervals for a single functional[J]. *Biometrika*, 1988, 75(2): 237-249.
- [3] ADIMARI G. An empirical likelihood statistic for quantiles[J]. *Journal of Statistical Computation and Simulation*, 1998, 60(1): 85-95.
- [4] QIN J, LAWLESS J. Empirical likelihood and general estimating equations[J]. *The Annals of Statistics*, 1994, 22(1): 300-325.
- [5] 王启华.经验似然统计推断方法发展综述[J].*数学进展*, 2004, 33(2): 141-151.
- [6] 王海娜.线性回归模型的若干稳健估计方法及应用实例[D].济南:山东大学,2013.
- [7] 吴喜之.复杂数据统计方法:基于R的应用[M].2版.北京:中国人民大学出版社,2013.
- [8] 郑忠国.关于稳健估计的若干问题[J].*数学进展*, 1986, 15(1): 1-8.
- [9] 樊亚莉,徐群芳.稳健的变量选择方法及其应用[J].*上海理工大学学报*, 2013, 35(3): 256-260.
- [10] HE X M, FUNG W K, ZHU Z Y. Robust estimation in generalized partial linear models for clustered data[J]. *Journal of the American Statistical Association*, 2005, 100(472): 1176-1184.
- [11] ÖZDEMİR Ş, ARSLAN O. Combining empirical likelihood and robust estimation methods for linear regression models[EB/OL]. [2018-04-02]. <http://arxiv.org/abs/1801.08812v1>.
- [12] OWEN J. Empirical likelihood for linear models[J]. *Empirical Likelihood for Linear Models*, 1991, 19(4): 1725-1747.
- [13] 周江文,黄幼才,杨元喜,等.抗差最小二乘法[M].武汉:华中理工大学出版社,1997.
- [14] HUBER P J, RONCHETTI E M. Robust statistics[M]. 2nd ed. New York: Wiley, 2009.
- [15] 俞志敏.基于 L_q 惩罚经验似然的变量选择[D].长沙:湖南师范大学,2013.
- [16] ÖZDEMİR Ş, ARSLAN O. An alternative algorithm of the empirical likelihood estimation for the parameter of a linear regression model[J]. *Communications in Statistics-Simulation and Computation*, 2018, 502: 1-9.