文章编号:1671-6833(2003)04-0099-03

污染数据回归分析参数的区间估计

胡玉萍[1], 王 霞[2], 李学相[3]

(1. 郑州大学系统科学与数学系,河南 郑州 450052; 2. 郑州轻工业学院信息与计算科学系,河南 郑州 450002;3. 郑州大学工程力学系,河南 郑州 450002)

摘 要:截断数据是生存分析的重要研究内容,而关于污染数据的分析在近几年也越来越受到人们的重视.研究简单回归模型:

$$X_i^{(0)}=\gamma+\beta\mu_i+\varepsilon_i,i=1,2,\cdots,n$$

其中,$E\varepsilon_i=0,E\varepsilon_i^2=\sigma_1^2$;但 $X_1^{(0)},X_2^{(0)},\cdots,X_n^{(0)}$ 受到另一独立同分布随机变量序列 $W_1,W_2,\cdots,W_n$ 的污染,W_i 与 $X_i^{(0)}$ 独立,仅能观察到污染数据 $X_i=(1-\alpha)X_i^{(0)}+\alpha W_i,i=1,2,\cdots,n$.给出回归参数 γ,β 的区间估计.

关键词:截断数据;污染数据;区间估计

中图分类号:O 212.7 **文献标识码**:A

0 引言

近年来,截断数据(censored data)的研究获得了很大的发展.在实际问题中,除了截断数据之外,还经常会遇到一些关于所谓污染数据(contaminated data)的统计分析问题,其中一类污染数据具有形式

$$X_i=(1-\alpha)X_i^{(0)}+\alpha W_i,i=1,2,\cdots,n \quad (1)$$

即我们在观察随机变量 $X_i^{(0)}$ 时,受到随机变量 W_i 的干扰(一般我们假定 $X_i^{(0)}$ 是独立同分布的,且序列 W_i 与序列 $X_i^{(0)}$ 独立),使观察数据受到污染,我们希望由观察数据 X_i 来对 $X_i^{(0)}$ 的分布及其特征作出统计推断.

考虑简单线性模型

$$X_i^{(0)}=\gamma+\beta\mu_i+\varepsilon_i,i=1,2,\cdots,n \quad (2)$$

其中,μ_i 为固定的回归设计(常数序列);γ,β 是待估参数.设

$$\{\varepsilon_i,1\leqslant i\leqslant n\}\ i.i.d.,\varepsilon_i\sim N(0,\sigma_1^2),0<\sigma_1^2<\infty \quad (3)$$

在式(1)决定的污染状态下,我们只能观察到

$$X_i=(1-\alpha)X_i^{(0)}+\alpha W_i,0\leqslant\alpha\leqslant1,$$

设随机变量 W_i 相互独立,服从分布 $N(0,\sigma_2^2)$,且与序列 $\{\varepsilon_i\},\{X_i^{(0)}\}$ 独立,即

$$\{W_i,1\leqslant i\leqslant n\}\ i.i.d.,W_i\sim N(0,\sigma_2^2),0<\sigma_2^2<\infty \quad (4)$$

我们的目的是要用观察数据 X_i 来估计参数 γ,β,假定 σ_1^2,σ_2^2 已知,且要求

$$\sigma_1^2/\sigma_2^2>\alpha/(1-\alpha) \quad (5)$$

直观上要求因污染而引起的方差(按一定比例)小于系统部分引起的方差.注意到

$$\begin{aligned}X_i&=(1-\alpha)X_i^{(0)}+\alpha W_i\\&=(1-\alpha)(\gamma+\beta\mu_i+\varepsilon_i)+\alpha W_i\\&=(1-\alpha)(\gamma+\beta\mu_i)+[(1-\alpha)\varepsilon_i+\alpha W_i]\end{aligned} \quad (6)$$

令 $\eta_i=(1-\alpha)\varepsilon_i+\alpha W_i$,则 η_i 相互独立服从 $N(0,(1-\alpha)^2\sigma_1^2+\alpha^2\sigma_2^2)$.

令 $\gamma_1=(1-\alpha)\gamma,\beta_1=(1-\alpha)\beta$,把 γ_1,β_1 视为新的参数,文献[1]用最小二乘法可求出 γ_1,β_1 的最小二乘估计

$$\gamma_1=\frac{\sum_{i=1}^n\mu_i\sum_{i=1}^n\mu_ix_i-\sum_{i=1}^n\mu_i^2\sum_{i=1}^nx_i}{(\sum_{i=1}^n\mu_i)^2-n\sum_{i=1}^n\mu_i^2} \quad (7)$$

$$\beta_1=\frac{\sum_{i=1}^n\mu_i\sum_{i=1}^nx_i-n\sum_{i=1}^n\mu_ix_i}{(\sum_{i=1}^n\mu_i)^2-n\sum_{i=1}^n\mu_i^2} \quad (8)$$

又考虑 X_i 的方差估计

收稿日期:2002-05-05;**修订日期**:2002-07-02

基金项目:河南省自然科学基金资助项目(0311010500)

作者简介:胡玉萍(1972-),女,河南省尉氏县人,郑州大学讲师,硕士,主要从事非参数统计的研究.

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

$$S^2=\frac{1}{n-1}\sum_{i=1}^{n}[X_i-\hat{r}_1-\beta_1\mu_i]^2,$$

当$n\to\infty$时,由文献[2]知$S^2\to(1-\alpha)^2\sigma_1^2+\alpha^2\sigma_2^2\triangleq\sigma^2$(a.s)

令$S^2=(1-\alpha)^2\sigma_1^2+\alpha^2\sigma_2^2$,可解出$\alpha$的估计值

$$\hat{\alpha}=\frac{\sigma_1^2-\sqrt{(\sigma_1^2+\sigma_2^2)S^2-\sigma_1^2\sigma_2^2}}{\sigma_1^2+\sigma_2^2} \tag{9}$$

又$\gamma_1=(1-\alpha)\gamma$,$\beta_1=(1-\alpha)\beta$,由$\gamma_1$,$\beta_1$,$\alpha$的估计值$\gamma_1$,$\beta_1$,$\hat{\alpha}$可解出$\gamma$,$\beta$的估计值

$$\gamma=\gamma_1\frac{1}{1-\hat{\alpha}} \tag{10}$$

$$\beta=\beta_1\frac{1}{1-\hat{\alpha}} \tag{11}$$

本文现考虑γ,β的区间估计问题.

1 回归参数的区间估计

引理1[3] 设$X_1,X_2,\cdots,X_n$为观察到的一列相互独立数据,且$X_i=(1-\alpha)X_i^{(0)}+\alpha W_i$,$X_i^{(0)}=\gamma+\beta\mu_i+\varepsilon_i$,$i=1,2,\cdots,n$.若式(3)、(4)成立,则

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-(\gamma_1+\beta_1\mu_i))^2\sim\chi^2(n-2).$$

证明过程见文献[3].

为方便,记$SSE=\sum_{i=1}^{n}(X_i-(\gamma_1+\beta_1\mu_i))^2$,$\mu=\frac{1}{n}\sum_{i=1}^{n}\mu_i$,$l_{\mu\mu}=\sum_{i=1}^{n}(\mu_i-\mu)^2$.

引理2 在引理1的条件下,有

(1) $\beta_1\sim N(\beta_1,\sigma^2/l_{\mu\mu})$;

(2) $\gamma_1\sim N(\gamma_1,(\frac{1}{n}+\frac{\mu^2}{l_{\mu\mu}})\sigma^2)$;

(3) β_1、γ_1分别与SSE相互独立.

证明:由式(6)可知

$$X_i=(1-\alpha)(\gamma+\beta\mu_i)+[(1-\alpha)\varepsilon_i+\alpha W_i]$$

$$X_i\sim N((1-\alpha)(\gamma+\beta\mu_i),\sigma^2),$$

又γ_1,β_1分别是$\gamma_1=(1-\alpha)\gamma$,$\beta_1=(1-\alpha)\beta$的最小二乘估计,且式(3)、(4)成立,由文献[4]中第5页性质1、性质2可得

$$\beta_1\sim N(\beta_1,\sigma^2/l_{\mu\mu}),$$

$$\gamma_1\sim N(\gamma_1,(\frac{1}{n}+\frac{\mu^2}{l_{\mu\mu}})\sigma^2),$$

且β_1,γ_1分别与SSE相互独立.

定理1 $X_1,X_2,\cdots,X_n$为观察到的一列相互独立数据,且

$$X_i=(1-\alpha)X_i^{(0)}+\alpha W_i,i=1,2,\cdots,n;$$

$$X_i^{(0)}=\gamma+\beta\mu_i+\varepsilon_i,i=1,2,\cdots,n,$$

式中:$\{\mu_i\}$为已知的常数序列;序列$\{X_i^{(0)}\}$,$\{\varepsilon_i\}$,$\{W_i\}$相互独立,且ε_i与W_i相互独立.如果式(3)、(4)成立,则参数β、γ的给定置信度为$1-\alpha_1$的近似置信区间分为(b_1,b_2),(c_1,c_2),

$$b_1=(\beta_1-t_{\alpha_1/2}(n-2)\hat{\sigma}_{\beta_1})/(1-\hat{\alpha});$$

$$b_2=(\beta_1+t_{\alpha_1/2}(n-2)\hat{\sigma}_{\beta_1})/(1-\hat{\alpha});$$

$$c_1=(\gamma_1-t_{\alpha_1/2}(n-2)\hat{\sigma}_{\gamma_1})/(1-\hat{\alpha});$$

$$c_2=(\gamma_1+t_{\alpha_1/2}(n-2)\hat{\sigma}_{\gamma_1})/(1-\hat{\alpha}).$$

证明:记γ_1,β_1的标准差为σ_{γ_1},σ_{β_1},即

$$\sigma_{\gamma_1}=\sqrt{\frac{1}{n}+\frac{\mu^2}{l_{\mu\mu}}}\sigma,$$

$$\sigma_{\beta_1}=\sigma/\sqrt{l_{\mu\mu}}.$$

由文献[4]知,$S^2=\frac{SSE}{n-2}$是σ^2的无偏估计,故σ_{γ_1},σ_{β_1}的估计值为

$$\hat{\sigma}_{\gamma_1}=\sqrt{\frac{1}{n}+\frac{\mu^2}{l_{\mu\mu}}}S;$$

$$\hat{\sigma}_{\beta_1}=S/\sqrt{l_{\mu\mu}},$$

由引理2可知

$$\frac{\beta_1-\beta_1}{\sigma_{\beta_1}}\sim N(0,1),$$

β_1与SSE相互独立.根据引理1知

$$\frac{SSE}{\sigma^2}\sim\chi^2(n-2),$$

故

$$\frac{\beta_1-\beta_1}{\sigma/\sqrt{l_{\mu\mu}}}/\sqrt{\frac{SSE}{(n-2)\sigma^2}}=\frac{\beta_1-\beta_1}{\hat{\sigma}_{\beta_1}}\sim t(n-2).$$

给定置信度为$1-\alpha_1$,则

$$P\{-t_{\alpha_1/2}(n-2)<\frac{\beta_1-\beta_1}{\hat{\sigma}_{\beta_1}}<t_{\alpha_1/2}(n-2)\}=1-\alpha_1,$$

解得

$$P\{\beta_1-t_{\alpha_1/2}(n-2)\hat{\sigma}_{\beta_1}<\beta_1<\beta_1+t_{\alpha_1/2}(n-2)\hat{\sigma}_{\beta_1}\}=1-\alpha_1.$$

所以β_1置信度为$1-\alpha_1$的置信区间为

$$[\beta_1-t_{\alpha_1/2}(n-2)\hat{\sigma}_{\beta_1},\beta_1+t_{\alpha_1/2}(n-2)\hat{\sigma}_{\beta_1}].$$

由于$\beta_1=(1-\alpha)\beta$,用式(9)$\hat{\alpha}$作为α的估计值,令$\beta_1-t_{\alpha_1/2}(n-2)\hat{\sigma}_{\beta_1}=(1-\hat{\alpha})\beta$,解得

$$\beta=(\beta_1-t_{\alpha_1/2}(n-2)\hat{\sigma}_{\beta_1})/(1-\hat{\alpha});$$

令$\beta_1+t_{\alpha_1/2}(n-2)\hat{\sigma}_{\beta_1}=(1-\hat{\alpha})\beta$,解得

$$\beta=(\beta_1+t_{\alpha_1/2}(n-2)\hat{\sigma}_{\beta_1})/(1-\hat{\alpha}),$$

所以β的置信度为$1-\alpha_1$的近似置信区间为(b_1,b_2),其中,

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

$$b_1=(\beta_1-t_{\alpha_1/2}(n-2)\hat{\sigma}_{\beta_1})/(1-\hat{\alpha});$$

$$b_2=(\beta_1-t_{\alpha_1/2}(n-2)\hat{\sigma}_{\beta_1})/(1-\hat{\alpha}).$$

类似可得 γ 的置信度为 $1-\alpha_1$ 的近似置信区间为 (c_1,c_2)，其中

$$c_1=(\gamma_1-t_{\alpha_1/2}(n-2)\hat{\sigma}_{\gamma_1})/(1-\hat{\alpha});$$

$$c_2=(\gamma_1-t_{\alpha_1/2}(n-2)\hat{\sigma}_{\gamma_1})/(1-\hat{\alpha}),$$

证毕.

2　结束语

通过 γ_1,β_1 的最小二乘估计 γ_1,β_1 所服从的分布，找到 γ_1,β_1 的置信区间，从而得到回归参数 γ,β 的近似置信区间.

参考文献：

[1]　郑祖康，吴雪明，饶　刚. 污染数据处理[J]. 应用概率统计，1998，14(3)：307～312.

[2]　陈希孺，陈桂景，吴启光，等. 线性模型参数的估计理论[M]. 北京：科学出版社，1985.

[3]　胡玉萍，王　霞，李学相. 污染数据回归分析的参数估计. 郑州大学学报(工学版)，2003，24(2)：84～86.

[4]　周纪芗. 实用回归分析方法[M]. 上海：上海科学技术出版社，1990.

Interval Estimation in Regression Analysis for Contamination Data

HU Yu-ping[1], WANG Xia[2], LI Xue-xiang[3]

(1. Department of System Science & Mathematics, Zhengzhou University, Zhengzhou 450052, China; 2. Department of Information and Computing Scieces, Institute of Light Industry, Zhengzhou 450002, China; 3. Department of Engineering Mechanics, Zhengzhou University, Zhengzhou 450002, China)

Abstract: Survival analysis attaches much importance to censored data, and the analysis of contaminated data is getting more and more attention in recent years. This paper studies the simple regression model:

$$X_i^{(0)}=\gamma+\beta\mu_i+\xi_i,\ i=1,2,\cdots,n$$

where $E\xi_i=0, E\xi_i^2=\sigma_1^2$, but $X_1^{(0)}, X_2^{(0)}, \cdots, X_n^{(0)}$ are contaminated by another i.i.d. random variable sequence $W_1, W_2, \cdots, W_n$. W_i is independent of $X_i^{(0)}$. And only the contaminated data $X_i=(1-\alpha)X_i^{(0)}+\alpha W_i$ are observable, giving the interval estimation of γ and β.

Key words: censored data; contamination data; interval estimation

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net