

关于污染数据的区间估计

何 春

(广东工业大学 应用数学系, 广东 广州, 510643)

摘要: 讨论了具有如下形式

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

的污染数据的参数估计, 其中 $X_i^{(0)} = \gamma + \beta\mu_i + \varepsilon_i$, $\varepsilon_i \sim N(0, \sigma_1^2)$, $W_i \sim N(0, \sigma_2^2)$, 给出 α 的区间估计, 并给出 γ, β 的点估计.

关键词: 污染数据; 区间估计; 点估计

中图分类号: O212

文献标识码: A

文章编号: 1007-7162(2001)03-0102-04

0 引言

设污染数据具有形式 $X_i = (1 - \alpha)X_i^{(0)} + \alpha W_i$, 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 $X_i^{(0)} = \gamma + \beta\mu_i + \varepsilon_i$, $i = 1, 2, \dots, n$, 其中 ε_i 相互独立, 服从分布 $N(0, \sigma_1^2)$, μ_i 为固定的回归系数(常数序列), γ, β 为未知参数. 我们只能观察到 $X_i = (1 - \alpha)X_i^{(0)} + \alpha W_i$, ($0 \leq \alpha \leq 1$), 其中随机变量 W_i 相互独立, 服从分布 $N(0, \sigma_2^2)$ 且与序列 $\{X_i^{(0)}\}$, 序列 $\{\varepsilon_i\}$ 相互独立. 我们的目的是要用观察到的数据 X_i 来估计未知参数 γ, β, α . 假定 σ_1^2, σ_2^2 已知且要求 $\sigma_1^2/\sigma_2^2 > \alpha/(1 - \alpha)$, 直观上要求因污染而引起的误差(按一定比例)小于系统部分引起的误差. 注意到

$$X_i = (1 - \alpha)X_i^{(0)} + \alpha W_i = (1 - \alpha)(\gamma + \beta\mu_i + \varepsilon_i) + \alpha W_i = (1 - \alpha)(\gamma + \beta\mu_i) + [(1 - \alpha)\varepsilon_i + \alpha W_i].$$

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

把 $(1 - \alpha)\gamma, (1 - \alpha)\beta$ 视为新的参数, 且令 $(1 - \alpha)\gamma = \gamma_1, (1 - \alpha)\beta = \beta_1$, 则 $X_i = \gamma_1 + \beta_1\mu_i + \eta_i$, $i = 1, 2, \dots, n$

利用最小二乘法可求得 γ_1, β_1 的最小二乘估计:

$$\hat{\gamma}_1 = \frac{\sum_{i=1}^n \mu_i \sum_{i=1}^n \mu_i x_i - \sum_{i=1}^n \mu_i^2 \sum_{i=1}^n x_i}{\left(\sum_{i=1}^n \mu_i\right)^2 - n \sum_{i=1}^n \mu_i^2}, \quad (1)$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n \mu_i \sum_{i=1}^n x_i - n \sum_{i=1}^n \mu_i x_i}{\left(\sum_{i=1}^n \mu_i\right)^2 - n \sum_{i=1}^n \mu_i^2}. \quad (2)$$

收稿日期: 2000-10-27

作者简介: 何春(1967-), 女, 讲师, 主要研究方向为数理统计.

由于我们要估计的是 γ, β 而不是 γ_1, β_1 , 而

$$\gamma_1 = (1 - \alpha) \gamma, \quad \beta_1 = (1 - \alpha) \beta \quad (3)$$

α 为未知参数, 所以必须知道 α 的估计值才可以求出 γ, β 的估计值. 文献[1]中利用 X_i 的方差的估计

$$R_n = \frac{1}{n-2} \sum_{i=1}^n [X_i - \hat{\gamma}_1 - \hat{\beta}_1 \mu_i]^2$$

来求出 α 的估计值. 因为 $R_n \xrightarrow{P} (1 - \alpha)^2 \sigma_1^2 + \alpha^2 \sigma_2^2$ (a. s.), 所以令 $R_n = (1 - \alpha)^2 \sigma_1^2 + \alpha^2 \sigma_2^2$, 可以解出 α 的估计值

$$\hat{\alpha} = \frac{\sigma_1^2 - \sqrt{(\sigma_1^2 + \sigma_2^2) R_n - \sigma_1^2 \sigma_2^2}}{\sigma_1^2 + \sigma_2^2}.$$

把 $\hat{\alpha}$ 代入式(3), 可以求得 γ, β 的估计值. 文献[2]证明了在一定的条件下

$$\hat{\beta} \xrightarrow{P} \beta(a. s.), \quad \hat{\gamma} \xrightarrow{P} \gamma(a. s.), \quad \hat{\alpha} \xrightarrow{P} \alpha(a. s.).$$

本文用 α 的区间估计来求出 α 的估计值.

1 引理

引理 1: 设 $X_1, \dots, X_n$ 相互独立, 且 $X_i \sim N((1 - \alpha)(1 + \beta \mu_i), \sigma^2)$, 其中 $\sigma^2 = (1 - \alpha)^2 \sigma_1^2 + \alpha^2 \sigma_2^2$, α, γ, β 未知, σ_1^2, σ_2^2 已知, 则

$$\frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \bar{X})^2$$

近似地服从 $\chi^2(n-1)$ 分布. 其中 $\chi^2(n-1)$ 表示自由度为 $n-1$ 的 χ^2 分布.

证明: 因为 $X_i \sim N((1 - \alpha)(\gamma + \beta \mu_i), \sigma^2)$, 所以 $X_i (i = 1, \dots, n)$ 的方差相同, 数学期望不同, 这时取 $(1 - \alpha)(\gamma + \beta \bar{\mu})$ 作为 X_i 的数学期望的近似值, ($i = 1, 2, \dots, n$), $\bar{\mu} = \frac{1}{n} \sum_{i=1}^n \mu_i$, 那么 X_i 近似地服从 $N((1 - \alpha)(\gamma + \beta \bar{\mu}), \sigma^2)$, ($i = 1, 2, \dots, n$), 从而 $X_1, X_2, \dots, X_n$ 就可以看作是独立同分布的随机变量, 因此

$$\frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \bar{X})^2 \text{ 近似地服从 } \chi^2(n-1) \text{ 分布. 证毕.}$$

2 污染数据的参数估计

定理 1: 设 $X_1, \dots, X_n$ 为观察到的一系列数据, 且

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

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

其中 $\{\mu_i\}$ 为已知的常数序列, 序列 $\{X_i^{(0)}\}, \{\varepsilon_i\}, \{W_i\}$ 均相互独立, 且 ε_i 与 W_i 相互独立, $\varepsilon_i \sim N(0, \sigma_1^2)$, $W_i \sim N(0, \sigma_2^2)$, σ_1^2, σ_2^2 已知, 则参数 α 的给定置信度为 $1 - \alpha_1$ 的置信区间为 (b, c) ,

$$b = \frac{\sigma_1^2 - \sqrt{(\sigma_1^2 + \sigma_2^2)(n-1)S^2/x_{1-\alpha_1/2}^2(n-1) - \sigma_1^2 \sigma_2^2}}{\sigma_1^2 + \sigma_2^2},$$

$$c = \frac{\sigma_1^2 + \sqrt{(\sigma_1^2 + \sigma_2^2)(n-1)S^2/x_{\alpha_1/2}^2(n-1) - \sigma_1^2 \sigma_2^2}}{\sigma_1^2 + \sigma_2^2},$$

其中 $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$, $x_{\alpha/2}^2(n-1)$, $x_{1-\alpha/2}^2(n-1)$ 查 $x^2(n-1)$ 分布表可知.

证明: 因为 $E(X_i) = (1-\alpha)(\gamma + \beta\mu_i)$, $D(X_i) = (1-\alpha)^2\sigma_1^2 + \alpha^2\sigma_2^2 \triangleq \sigma^2$, 所以 $X_i \sim N((1-\alpha)(\gamma + \beta\mu_i), \sigma^2)$.

又因为 X_i 相互独立, 根据引理 1 知

$$\frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \bar{X})^2$$

近似地服从 $x^2(n-1)$ 分布. 给定置信度为 $1-\alpha_1$, 则由 $x^2(n-1)$ 分布的性质知

$$P\{x_{1-\alpha/2}^2(n-1) < \frac{(n-1)s^2}{\sigma^2} < x_{\alpha/2}^2(n-1)\} = 1-\alpha_1,$$

解之得

$$P\left\{\frac{(n-1)s^2}{x_{\alpha/2}^2(n-1)} < \sigma^2 < \frac{(n-1)s^2}{x_{1-\alpha/2}^2(n-1)}\right\} = 1-\alpha_1,$$

所以 σ^2 的置信度为 $1-\alpha_1$ 的置信区间为

$$\left[\frac{(n-1)s^2}{x_{\alpha/2}^2(n-1)}, \frac{(n-1)s^2}{x_{1-\alpha/2}^2(n-1)} \right].$$

由于 $\sigma^2 = (1-\alpha)^2\sigma_1^2 + \alpha^2\sigma_2^2$, 由 $(1-\alpha)^2\sigma_1^2 + \alpha^2\sigma_2^2 = \frac{(n-1)s^2}{x_{1-\alpha/2}^2(n-1)}$, 解得

$$\hat{\alpha} = \frac{\sigma_1^2 - \sqrt{(\sigma_1^2 + \sigma_2^2)(n-1)s^2/x_{1-\alpha/2}^2(n-1) - \sigma_1^2\sigma_2^2}}{\sigma_1^2 + \sigma_2^2}.$$

由 $(1-\alpha)^2\sigma_1^2 + \alpha^2\sigma_2^2 = \frac{(n-1)s^2}{x_{\alpha/2}^2(n-1)}$, 解得

$$\hat{\alpha} = \frac{\sigma_1^2 - \sqrt{(\sigma_1^2 + \sigma_2^2)(n-1)s^2/x_{\alpha/2}^2(n-1) - \sigma_1^2\sigma_2^2}}{\sigma_1^2 + \sigma_2^2},$$

所以 α 的置信度为 $1-\alpha_1$ 的置信区间为 (b, c) , 其中

$$b = \frac{\sigma_1^2 - \sqrt{(\sigma_1^2 + \sigma_2^2)(n-1)s^2/x_{1-\alpha/2}^2(n-1) - \sigma_1^2\sigma_2^2}}{\sigma_1^2 + \sigma_2^2},$$

$$c = \frac{\sigma_1^2 - \sqrt{(\sigma_1^2 + \sigma_2^2)(n-1)s^2/x_{\alpha/2}^2(n-1) - \sigma_1^2\sigma_2^2}}{\sigma_1^2 + \sigma_2^2}.$$

证毕.

定理 2: 在定理 1 的条件下, α 在区间 (b, c) 内任取一值 $\hat{\alpha}$, 则可得 γ, β 的估计值为:

$$\hat{\gamma} = \frac{\sum_{i=1}^n \mu_i \sum_{i=1}^n \mu_i x_i - \sum_{i=1}^n \mu_i^2 \sum_{i=1}^n x_i}{(\sum_{i=1}^n \mu_i)^2 - n \sum_{i=1}^n \mu_i^2} \frac{1}{1-\hat{\alpha}}, \quad (4)$$

$$\hat{\beta} = \frac{\sum_{i=1}^n \mu_i \sum_{i=1}^n x_i - n \sum_{i=1}^n \mu_i x_i}{(\sum_{i=1}^n \mu_i)^2 - n \sum_{i=1}^n \mu_i^2} \frac{1}{1-\hat{\alpha}}. \quad (5)$$

证: 由式(1)、(2)、(3) 及 α 的估计值即可得式(4)、(5).
证毕.

参考文献:

- [1] 郑祖康, 吴雪明, 饶刚. 污染数据的处理[J]. 应用概率统计, 1998, 14(3): 307-312.
- [2] 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.
- [3] 汪荣鑫. 数理统计[M]. 西安交通大学出版社, 1989. 50-69.

The Interval Estimation of Contamination Data

HE Chun

(Dept. of Applied Mathematics, GDUT, Guangzhou, 510643, China)

Abstract: This paper considers the interval estimation of a kind of contamination data, gives the interval estimation of a parameter and the point estimation of the other two parameters. All the three parameters construct the model of the contamination data.

Key words: contamination data; interval estimation; point estimation