文章编号:1672 - 058X(2011)04 - 0379 - 04

随机删失下非线性回归模型均值的置信域

胡时财[1], 方连娣[2]
(1. 安徽新闻出版职业技术学院,安徽 合肥 230601;2. 铜陵学院 文学与艺术传媒系,安徽 铜陵 244000)

摘 要:考虑随机右删失数据下非线性回归模型中响应变量均值的估计问题,应用经验似然方法构造响应变量均值的调整的经验似然比统计量,证明了在一定的条件下,统计量渐近服从 χ^2 分布,所得结果给出了均值的渐近置信域。

关键词:随机删失;经验似然;χ^2 分布;非线性模型
中图分类号:O212.7 **文献标志码**:A

在医学、生物工程、临床试验等方面的研究中,通常得到的一类不完全数据是随机删失数据。例如,研究一种新药对患某种疾病的人的生命的影响,病人可以在研究时间段内任何时刻进入研究,一旦进入研究之后,病人由于中途离开或研究期间失去生命等原因,造成了数据随机删失。对退出试验或失去跟踪的病人生存时间至少是从进入到失去联系这段时间;对于仍然活着的病人,时间至少是从进入到结束研究这段时间,在这两种情况下得到的数据都属于随机删失数据。对删失数据回归模型的研究已有不少文献。Li-Gang 等[1]已经利用经验似然的方法解决了回归系数的估计问题,陈放等[2]考虑了随机右删失数据下非线性回归模型中未知参数的置信域。郑明,李四化[3]应用经验似然方法讨论了截断情况下线性回归模型中响应变量均值的估计问题。考虑下面的非线性回归模型:

$$Y = g(X,\beta) + \varepsilon \tag{1}$$

其中 Y 为响应变量,X 是 q 维协变量,β 是 p 维未知参数,$g(.,.)$ 是 X 的已知可测函数,$E(\varepsilon|X)=0$。设 C 表示删失变量,记 $Z=\min(Y,C)$,$\delta=I(Y\leqslant C)$,这里 $I(\cdot)$ 表示示性函数,当 $\delta=1$ 时表示 Y 被观察,当 $\delta=0$ 时表示 Y 删失。对于模型(1),观察到的数据 (X_i,Z_i,δ_i),$i=1,\cdots,n$,是来自 (X,Z,δ) 的独立同分布随机样本。

假定删失变量 C 的分布函数为 G,由于 $\{Y_i\}$ 被随机删失,通常的估计参数的方法不能直接运用,采用 $K-S-V$ 估计对响应变量进行一定的调整。当 G 已知时,定义:

$$Y_{iG} = \frac{\delta_i Z_i}{1 - G(Z_i)}, i = 1,\cdots,n。 \tag{2}$$

易证 $(Y_{iG}|X_i)=E(Y_i|X_i)=g(X_i,\beta)$,从直观上讲,这一方法把非删失的数据抬高,而把删失数据一律降为零。此时用 Y_{iG} 代替 Y_i,由非线性最小二乘法可得 β 的估计

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{2}\sum_{i=1}^{n}(Y_{iG} - g(X_i,\beta))^2 \tag{3}$$

对响应变量进一步调整,没有被删失的使用原来的数值,而对于被删失的用回归模型的预测值 $\hat{Y}$ 代替,即

$$Z_{in} = \delta_i Y_i + (1-\delta_i)\hat{Y}_i = \delta_i Y_i + (1-\delta_i)g(X_i,\hat{\beta}) \quad i = 1,\cdots,n. \tag{4}$$

此时,$E[Z_{in}]=E[Y_i]$,$i=1,\cdots,n$. 假设 $E[Y_i]=\mu$,当 μ 为响应变量均值真实值时,由上面的结论,$E[Z_{in}]=E[Y_i]=\mu$,$i=1,\cdots,n$. 利用响应变量的调整值 Z_{in} 可以很容易地得到 μ 的一个估计,

收稿日期:2011 - 01 - 18;**修回日期**:2010 - 04 - 20.
作者简介:胡时财(1972 -),安徽人,男,讲师,从事数学教育的研究.

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} Z_{in} = \frac{1}{n}\sum_{i=1}^{n}[\delta_i Y_i + (1-\delta_i)g(X_i,\hat{\beta}))] \tag{5}$$

此时有:

定理1 假设 $E\|X\| < \infty$, $E[\varepsilon^2] < \infty$, μ 为响应变量均值真实值,则有 $\sqrt{n}(\hat{\mu}-\mu) \to N(0,V(\mu))$,其中 $V(\mu) = S_1 + S_2 - 2S_3^{\mathrm{T}}\mu + \mu + S_4 + 2S_5$,而 $S_1 = E[\delta(Y-g(X,\beta))^2]$, $S_2 = E[g(X,\beta)]^2$, $S_3 = E[g(X,\beta)]$, $S_4 = E(1-\delta)[g(X,\hat{\beta}) - g(X,\beta)]^2$, $S_5 = E[Y-g(X,\beta)][g(X,\hat{\beta}) - g(X,\beta)]$。

当 $V(\mu)$ 未知的时候可以将 $V(\mu)$ 中期望形式换成 n 个样本的和记作 $\hat{V}_n(\mu)$ 来估计。

下面应用经验似然方法构造 μ 的置信区间。

由 $E(Z_{in}-\mu) = 0$ 可以定义 μ 的经验对数似然比统计量为

$$l(\mu) = -2\max\left\{\sum_{i=1}^{n}\log(np_i) \mid p_i \geqslant 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i Z_{in} - \mu = 0\right\}$$

由拉格朗日乘子法可得:

$$p_i = \frac{1}{n}\{1 + \lambda(Z_{in}-\mu)\}^{-1}, i = 1,2,\cdots,n$$

其中 λ 是下面方程的根

$$\frac{1}{n}\sum_{i=1}^{n}\frac{Z_{in}-\mu}{1+\lambda(Z_{in}-\mu)} = 0$$

故 $l(\mu) = 2\sum_{i=1}^{n}\log[1+\lambda_n(Z_{in}-\mu)]$,这里 Z_{in} 不再是独立的,所以似然比统计量不再是服从标准 χ^2 分布。下面给出一个调整的经验对数似然比 $\hat{l}(\mu) = r_n(\mu)l(\mu)$,这里调整因子 $r_n(\mu) = V_n(\mu)/\hat{V}_n(\mu)$,其中 $V_n(\mu) = \frac{1}{n}\sum_{i=1}^{n}(Z_{in}-\mu)^2$, $\hat{V}_n(\mu)$ 的定义如上式。

定理2 在定理1的条件下,如果 μ 为响应变量均值真实值,则 $l(\mu) \to \chi_1^2$。

基于定理2可以构造 μ 置信度为 $1-\alpha$ 的渐近置信域。

定理1的证明:

首先, $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(Z_{in}-\mu) = T_{n1} + T_{n2} + T_{n3}$,其中: $T_{n1} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\delta_i[Y_i - g(X_i,\beta)]$, $T_{n2} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(g(X_i,\beta)-\mu)$, $T_{n3} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(1-\delta_i)(g(X_i,\hat{\beta}) - g(X_i,\beta))$。

由中心极限定理,可得 $T_{n1} \to N(0,S_1)$, $T_{n2} \to N(0,S_2 - 2S_3^{\mathrm{T}}\mu + \mu^2)$, $T_{n3} \to N(0,S_4)$, $\mathrm{cov}(T_{n1},T_{n2}) = 0$, $\mathrm{cov}(T_{n2},T_{n3}) = 0$, $\mathrm{cov}(T_{n1},T_{n3}) = S_5$。

所以定理1得证。在证明定理2之前先给出几个引理。

引理1 在定理1的条件下,有:

$$\frac{1}{n}\sum_{i=1}^{n}(Z_{in}-\mu)^2 = S_1 + S_2 - 2S_3^{\mathrm{T}}\mu + \mu^2 + o_p(1)$$

证明 由非线性最小二乘法知 $\hat{\beta} \to \beta, a.s.$,故:

$$\frac{1}{n}\sum_{i=1}^{n}(Z_{in}-\mu)^2 = \frac{1}{n}\sum_{i=1}^{n}[\delta_i(Y_i - g(X_i,\beta)) + (1-\delta_i)(g(X_i,\hat{\beta}) - g(X_i,\beta)) + (g(X_i,\beta)-\mu)]^2 = R_{n1} + R_{n2} + R_{n3} + o_p(1)$$

其中, $R_{n1} = \frac{1}{n}\sum_{i=1}^{n}\delta_i[Y_i - g(X_i,\beta)]^2$, $R_{n2} = \frac{1}{n}\sum_{i=1}^{n}[g(X_i,\beta)-\mu]^2$, $R_{n3} = \frac{1}{n}\sum_{i=1}^{n}\delta_i[Y_i - g(X_i,\beta)][g(X_i,\beta)-\mu]$。

由大数定律可得:

$R_{n1} \to S_1$, $R_{n2} \to S_2 - 2S_3^{\mathrm{T}}\mu + \mu^2$, $R_{n3} \to E[\delta(Y-g(X,\beta))(g(X,\beta)-\mu)] = 0$

故引理1得证。

引理2 记 $Z_{(n)} = \max_{1\leqslant i\leqslant n}|Z_{in}|$,在定理1的条件下 $Z_{(n)} = o_p(1)$。

证明　注意到 $Z_{(n)} \leqslant \max\limits_{1\leqslant i\leqslant n} |Y_i| + \max\limits_{1\leqslant i\leqslant n} |g(X_i,\hat{\beta})|$,由文献[4]的引理 3 可得:

$$\max_{1\leqslant i\leqslant n} | Y_i | = o_p(n^{\frac{1}{2}}), \max_{1\leqslant i\leqslant n} | g(X_i,\hat{\beta}) | = o_p(n^{\frac{1}{2}})$$

所以引理 2 结论成立。

引理 3　在定理 1 的条件下,有 $\lambda = O_p(n^{-\frac{1}{2}})$。

证明　由定理 1 可得, $\frac{1}{n}\sum\limits_{i=1}^{n}(Z_{in}-\mu) = O_p(n^{-\frac{1}{2}})$,再由引理 1,2 及文献[4,5]类似地可证。

定理 2 的证明:对 $l(\mu_0)$进行 Taylor 展开,可得:

$$l(\mu) = 2\sum_{i=1}^{n}\{\lambda(Z_{in}-\mu) - \frac{1}{2}[\lambda(Z_{in}-\mu)]^2\} + \eta_n$$

其中 $|\eta_n| \leqslant C\sum\limits_{i=1}^{n}|\lambda(Z_{in}-\mu)|^3$,由引理 1 - 3 可得:

$$|\eta_n| \leqslant C|\lambda|^3 \max_{1\leqslant i\leqslant n}|Z_{in}-\mu| \sum_{i=1}^{n}(Z_{in}-\mu)^2 = o_p(1)$$

又

$$0 = \frac{1}{n}\sum_{i=1}^{n}\frac{Z_{in}-\mu}{1+\lambda(Z_{in}-\mu)} =$$

$$\frac{1}{n}\sum_{i=1}^{n}(Z_{in}-\mu) - [\frac{1}{n}\sum_{i=1}^{n}(Z_{in}-\mu)^2]\lambda + \frac{1}{n}\sum_{i=1}^{n}\frac{\lambda^2(Z_{in}-\mu)^3}{1+\lambda(Z_{in}-\mu)} =$$

$$\frac{1}{n}\sum_{i=1}^{n}(Z_{in}-\mu) - [\frac{1}{n}\sum_{i=1}^{n}(Z_{in}-\mu)^2]\lambda + o_p(1)$$

由引理 1-3 可得:

$$\lambda = [\sum_{i=1}^{n}(Z_{in}-\mu)^2]^{-1}\sum_{i=1}^{n}(Z_{in}-\mu) + o_p(1)$$

所以有

$$0 = \frac{1}{n}\sum_{i=1}^{n}\frac{\lambda(Z_{in}-\mu)}{1+\lambda(Z_{in}-\mu)} = \sum_{i=1}^{n}\lambda(Z_{in}-\mu) - [\sum_{i=1}^{n}\lambda(Z_{in}-\mu)]^2 + \sum_{i=1}^{n}\frac{[\lambda(Z_{in}-\mu)]^3}{1+\lambda(Z_{in}-\mu)}$$

又由引理可得:

$$\sum_{i=1}^{n}\frac{[\lambda(Z_{in}-\mu)]^3}{1+\lambda(Z_{in}-\mu)} = o_p(1)$$

由上面两式得:

$$\sum_{i=1}^{n}\lambda(Z_{in}-\mu) = [\sum_{i=1}^{n}\lambda(Z_{in}-\mu)]^2 + o_p(1)$$

所以, $l(\mu) = [\frac{1}{n}\sum\limits_{i=1}^{n}(Z_{in}-\mu)^2]^{-1}[\frac{1}{\sqrt{n}}\sum\limits_{i=1}^{n}(Z_{in}-\mu)]^2 + o_p(1)$, 因此,$\hat{l}(\mu) = r_n(\mu)l(\mu) = \{\frac{1}{\sqrt{n}}\sum\limits_{i=1}^{n}\frac{Z_{in}-\mu}{\sqrt{\hat{V}_n(\mu)}}\}^2 + o_p(1)$,易证 $\hat{V}_n(\mu)\to V(\mu)$,所以 $l(\mu)\to\chi_1^2$。

故定理得证。

参考文献:

[1] LI G, WANG Q H. Empirical likelihood confidence regression analysis with right censored data[J]. Statistica Sinica,2003,13(1):51-68

[2] 陈放,李高荣,冯三营,等. 右删失数据下非线性回归模型的经验似然推断[J]. 应用数学学报,2010,33(1):130-141

[3] 郑明,李四化. 截断情况下线性回归模型响应变量均值的经验似然[J]. 应用数学,2004,17(4):524-529

[4] OWEN A. Empirical likelihood ratio confidence intervals[J]. Biometrika,1988,75:237-249

[5] 方连娣. 非线性模型中参数的置信域[J]. 重庆工商大学学报:自然科学版,2009,26(1):4-7

Confidence Region of Nonlinear Regression Model Mean with Random Censoring

HU Shi-cai[1] , **FANG Lian-di**[2]

(1. Anhui Publishing Technical College, Anhui Hefei 230601, China;
2. Department of Literature, Art and Media, Tongling University, Anhui Tongling 244000, China)

Abstract: This paper considers the estimation for response variable mean in nonlinear regression model with random right censoring data, uses empirical likelihood method to construct adjusted empirical likelihood ratio statistic of response variable mean, and proves that the statistic follows asymptotic chi-square distribution under certain condition. The obtained results present asymptotic confidence region of the mean.

Key words: random censoring; empirical likelihood; chi-square distribution; nonlinear model

责任编辑:田 静

(上接第371页)

以 r_1 为例, $r_{1=}\sum_{k=1}^{4} w_k * \varepsilon_1(k) = 0.37*0.64 + 0.21*0.75 + 0.26*0.98 + 0.16*0.89 = 0.79$,由此可得,$r_2, r_3, r_4, r_5, r_6, r_7, r_8, r_9, r_{10}$的值,如表11所示。即可知:$r_1 > r_4 > r_{10} > r_7 > r_6 > r_2 = r_8 > r_9 > r_3 > r_5$,故,可得到10名准教师的评比顺序为:$B_1 > B_4 > B_{10} > B_7 > B_6 > B_2 = B_8 > B_9 > Br_3 > B_5$。

参考文献:

[1] 宋广飞,袁永博,张明媛. 基于AHP和多层模糊综合评判的购物中心选址[J]. 建筑管理现代化,2008(6):13-16
[2] 梅冬,武饮彩,施红星. 基于灰色关联分析的红旗车驾驶员评比[J]. 研究军事交通学报,2009,11(4):88-91
[3] 时丕生. 基于灰色关联分析的地下水环境质量评价[J]. 山东农业大学学报,2009,40(4):563-566
[4] 胡淑礼. 模糊数学及其应用[M]. 成都:四川大学出版社,1994
[5] 江礼政. 基于层次分析法的重庆市不同发电技术竞争力的比较[J]. 重庆工商大学学报:自然科学版,2008,25(1):33-37

Application of AHP and Gray Correlation Analysis Method to Teacher Recruitment

CHEN Zheng-min, WANG Juan

(School of Mathematics, Chongqing Normal University, Chongqing 401331, China)

Abstract: This paper firstly uses analytic hierarchy process (AHP) to obtain weight for influencing each index of the achievements of quasi-teachers and then implements gray correlation analysis method to make association sorting for quasi-teachers by calculating gray correlation degree of each evaluation index and ideal optimal effect vector so as to effectively select excellent teachers, which has strong practical significance.

Key words: teacher; AHP; gray correlation analysis; appraisal

责任编辑:田 静