Statistics and Application 统计学与应用, 2016, 5(2), 172-178
Published Online June 2016 in Hans. http://www.hanspub.org/journal/sa
http://dx.doi.org/10.12677/sa.2016.52017

Hans汉斯

Using Order Censored Data Regression Method to Infer the Normal Overall Unknown Parameter

Caiyun Sun, Mengying Chang, Qin Yue, Kunming Xie, Haiqiang Zeng

Department of Basic Subject, North China University of Science & Technology, Beijing
Email: yuncai@ncist.edu.cn

Received: Jun. 7th, 2016; accepted: Jun. 27th, 2016; published: Jun. 30th, 2016

Abstract

The characteristics such as sample data amount, good integrity, and meeting the random drawing are often required in using the method of sampling analysis for parameter estimation and inference in statistics. However, in practice we are often faced with a series of order censored data, and the parameter error obtained by the usual estimation methods may be very big. In this paper, the method of regression analysis was used to deduce the overall parameters by using the order truncation data and to do the simulation calculation. It is shown that the method is feasible by comparing the error. Finally we used the method to solve the problem about estimating the overall average score by using some sort of student achievement.

Keywords

Order Censored Data, Regression Analysis, Order Statistics

利用次序截尾数据线性回归方法推断正态总体的未知参数

孙彩云，常梦颖，岳　琴，谢昆明，曾海强

华北科技学院基础部，北京
Email: yuncai@ncist.edu.cn

文章引用: 孙彩云, 常梦颖, 岳琴, 谢昆明, 曾海强. 利用次序截尾数据线性回归方法推断正态总体的未知参数[J]. 统计学与应用, 2016, 5(2): 172-178. http://dx.doi.org/10.12677/sa.2016.52017

收稿日期：2016年6月7日；录用日期：2016年6月27日；发布日期：2016年6月30日

摘　要

统计学中，在采用抽样分析的方法进行参数估计与推断时，常要求样本数据数量、完整性好，且须满足随机抽取等特点。在实际应用中会遇到一系列按顺序排列的截尾样本数据，用普通的参数估计方法得到的参数估计值误差可能会较大，本文将回归分析方法推广到利用次序截尾数据推断正态总体参数中，且对所做结果进行模拟计算。通过误差比对，认为该方法是可行的，本文最后利用该方法通过部分排序的学生成绩推断了总体的平均成绩。

关键词

次序截尾数据，回归分析，次序统计量

1. 引言

统计学中，在采用抽样分析的方法进行参数估计与推断时，常要求样本数据数量、完整性好，且须满足随机抽取等特点。然而在实际应用中常会遇到一系列按顺序排列的截尾样本数据，例如在教学活动中，经常会组织学生参加各类学科竞赛，竞赛组织方通常只公布获奖选手的成绩(可看成顺序截尾样本)，而不公布所有参赛学生的成绩。这就涉及到参赛学校怎样根据获奖选手的成绩对本校所有参赛学生的整体成绩进行推断，从而评价各种教学指标的优劣。从数理统计的角度看，本问题可化为由次序截尾样本对总体参数进行统计推断的问题。

本文欲采用回归分析方法对总体参数进行推断。众所周知，传统的回归分析是一种强有力的数据处理工具，在自然科学和社会科学的各个领域都有广泛的应用，但是它只适用于来自正态分布的完全数据[1]。对于次序截尾数据是无法处理的。茆诗松等人提出了截尾数据的最佳线性无偏估计方法[2]，傅惠民等人又提出了最佳无偏整体估计方法[3]，本文结合这两种方法，将回归分析方法推广到利用次序截尾数据推断总体参数的问题当中，并且对所做结果进行模拟计算，通过误差比对，说明了该方法的可行性。

2. 预备知识

2.1. 次序统计量

设 $X_1, X_2, \cdots, X_n$ 是来自某个总体的一个样本。该样本的第 i 个次序统计量记为 $X_{(i)}$，它是如下的样本函数，每当该样本得到一组观测量值 $x_1, x_2, \cdots, x_n$ 时，将它们从小到大排列起来为 $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$，其中第 i 个值 $x_{(i)}$ 就是 $X_{(i)}$ 的观测值。称 $\left(X_{(1)}, X_{(2)}, \cdots, X_{(n)}\right)$ 为该样本的次序统计量。由文献[4]知，若总体的分布函数为 $F(x)$，密度函数为 $p(x)$，可推出次序统计量的密度及联合密度如下：

$X_{(i)}$ 的密度函数为 $g(y_i)$，其中 $1 \le i \le n$。

$$g(y_i) = \frac{n!}{(i-1)!(n-i)!}\left[F(y_i)\right]^{i-1}\left[1-F(y_i)\right]^{n-i} p(y_i)$$

$X_{(i)}$ 和 $X_{(j)}$ 的联合密度函数为 $g(y_i, y_j)$，其中 $1 \le i < j \le n$。

$$g(y_i, y_j) = \frac{n!}{(k-1)!(j-1-i)!(n-j)!}\left[F(y_i)\right]^{i-1}\left[F(y_j)-F(y_i)\right]^{j-1-i}\left[1-F(y_j)\right]^{n-j} p(y_i)\,p(y_j)$$

在这个等式中， $y_i \le y_j$ 都成立，在其他的场合 $g\left(y_i, y_j\right)=0$ 。

由密度函数可以计算次序统计量 $X_{(i)}$ 的期望和方差，记

$EX_{(i)}=\mu_i, i=1,\cdots,n$ ， $Cov\left(X_{(i)}, X_{(j)}\right)=v_{ij}, 1\le i, j\le n$ ，则

$$\begin{aligned}\mu_i &= \int_{-\infty}^{+\infty} y_i g\left(y_i\right)\mathrm{d}y_i = \int_{-\infty}^{+\infty} y_i \frac{n!}{(i-1)!(n-i)!}\left[F\left(y_i\right)\right]^{i-1}\left[1-F\left(y_i\right)\right]^{n-1} p\left(y_i\right)\mathrm{d}y_i \\ &= \int_0^1 \frac{n!}{(i-1)!(n-i)!} y_i\left(F\right) F\left(y_i\right)^{i-1}\left[1-F\left(y_i\right)\right]^{n-i}\mathrm{d}F\left(y_i\right)\end{aligned} \tag{1}$$

其中： $y_i\left(F\right)$ 为 $F\left(y_i\right)$ 的反函数

$$m_{ii}=\frac{n!}{(i-1)!(n-i)!}\int_0^1\left[y_i\left(F\right)\right]^2\left[F\left(y_i\right)\right]^{i-1}\left[1-F\left(y_i\right)\right]^{n-i}\mathrm{d}F\left(y_i\right)-\alpha_i^2 \tag{2}$$

$$\begin{aligned}m_{ij} &= \int_0^1\int_0^{F(y_i)} \frac{n!}{(i-1)!(j-i-1)!(n-j)!}\left[F\left(y_i\right)\right]^{k-1}\left[F\left(y_j\right)-F\left(y_i\right)\right]^{j-i-1} \\ &\quad\times\left(1-y_j\right)^{n-j}\left[F\left(y_i\right)\right]^{-1}\mathrm{d}F\left(y_i\right)\mathrm{d}F\left(y_j\right)-\alpha_i\alpha_j\end{aligned} \tag{3}$$

上面各式中， μ_i ， $m_{ij}\left(1\le i, j\le r\le n\right)$ 仅与 n ， i ， j 和 $F\left(x\right)$ 有关，可通过查表和专门程序计算[5]得到。

2.2. 广义 Gauss-Markov 模型

普通线性回归模型 $\left(Y, X\beta, \sigma^2 I_n\right)$ [5]中，若将 $Var\left(Y\right)=\sigma^2 I_n$ 改为 $Var\left(Y\right)=\sigma^2 G$ ， G 为已知正定阵，则形成所谓的广义 Gauss-Markov 模型，对此模型，因 $G>0$，存在 n 阶非奇异对称阵 B^2 ，使 $G=B^2$ 。令 $\tilde{Y}=B^{-1}Y, \tilde{X}=B^{-1}X$ ，则

$$E\tilde{Y}=B^{-1}EY=B^{-1}X\beta=\tilde{X}\beta$$

$$Var\left(\tilde{Y}\right)=B^{-1}Var\left(Y\right)B^{-1}=\sigma^2 I_n$$

由此， $\left(\tilde{Y}, \tilde{X}\beta, \sigma^2 I_n\right)$ 是一个 Gauss-Markov 模型，由该模型得到的最小二乘估计(LSE)为

$$\tilde{\beta}=\left(\tilde{X}'\tilde{X}\right)^{-1}\tilde{X}'\tilde{Y}=\left(X'G^{-1}X\right)^{-1}X'G^{-1}Y \tag{4}$$

称为 β 的加权最小二乘估计，由文献[6]，知它仍是 β 的最好线性无偏估计(BLUE)。

3. 次序截尾数据线性回归方法

设 $X_1, X_2, \cdots, X_n$ 是来自 $N\left(\mu, \sigma^2\right)$ 的一个样本，要估计 μ 和 σ ($\sigma>0$)。设 $X_{(1)}\le X_{(2)}\le\cdots\le X_{(r)}$ 为观测到的前 $r\left(r=1,2,\cdots,n\right)$ 个次序统计量，$1\le i, j\le r\le n$ 考虑这样一类估计，它们是次序统计量的线性函数。

令

$$X_{(i)}^0=\left(X_{(i)}-\mu\right)\Big/\sigma,\quad i=1,\cdots,r \tag{5}$$

则 $X_{(1)}^0\le\cdots\le X_{(r)}^0$ 相当于抽自 $\Phi\left(x\right)$ 的容量为 n 的前 r 个截尾样本。记

$$EX_{(i)}^0=\alpha_i,\ i=1,\cdots,r$$

$$Cov\left(X_{(i)}^0, X_{(i)}^0\right)=v_{ij},\ 1\le i, j\le r$$

由(1)、(2)、(3)式可知 α_i, v_{ij} 只依赖于 n， i, j 和 $\Phi(x)$，而与 μ, σ 无关，由于 $\Phi(x)$ 已知，所以当 r 取定后， α_i, v_{ij} 是可计算的。将(5)式化成

$$X_{(i)} = \mu + \sigma X_{(i)}^0 = \mu + \sigma\alpha_i + \varepsilon_i \tag{6}$$

其中

$$\varepsilon_i = \sigma\left(X_{(i)}^0 - \alpha_i\right) \quad (i = 1, 2, \cdots)$$

记 $X' = \left(X_{(1)}, \cdots, X_{(r)}\right)$， $\alpha' = (\alpha_1, \cdots, \alpha_r)$，用矩阵表示(5)式，有

$$EX = (1_r, \alpha)\begin{pmatrix}\mu\\ \sigma\end{pmatrix} \tag{7}$$

$$Var(X) = \sigma^2 V = \sigma^2 \left(v_{ij}\right)_{r\times r} \tag{8}$$

其中 1_r 表示全部由元素 1 组成的 r 维列向量。这是广义 Gauss-Markov 模型，由(4)式可求出 μ 和 σ 的 BLUE 为

$$\begin{pmatrix}\hat{u}\\ \hat{\sigma}\end{pmatrix} = \begin{pmatrix}1_r' V^{-1} 1_r & 1_r V^{-1}\alpha'\\ \alpha' V^{-1} 1_r & \alpha' V^{-1}\alpha\end{pmatrix}^{-1}\begin{pmatrix}1_r'\\ \alpha'\end{pmatrix} V^{-1} X \text{ 记作} \begin{pmatrix}L_1' X\\ L_2' X\end{pmatrix} \tag{9}$$

其协方差矩阵为

$$Var\begin{pmatrix}\hat{\mu}\\ \hat{\sigma}\end{pmatrix} = \sigma^2\begin{pmatrix}L_1' V L_1 & L_1' V L_2\\ L_1' V L_2 & L_2' V L_2\end{pmatrix} \tag{10}$$

该估计方法的优点在于，不论 n 个样品中被观测到的样品个数是多少($n \geq 2$)，上述方法都可使用。这样我们就可以由小样本进行线性回归，并且推断总体的未知参数，可以改进线性回归及统计推断在应用上的一些局限性。

4. 模拟计算

为了客观说明以上估计方法的可行性，本文由计算机随机产生正态分布 $N(2, 0.8)$ 的 15 个次序随机数作为一个样本，分别截取前 r 个，利用次序截尾数据线性回归方法来估计正态整体的参数 μ 和 σ 的值，并计算所得估计的相对误差。

产生的样本如下：

0.8621 0.8782 1.3564 1.3881 1.5075 1.8432 1.8461 1.8581

2.1582 2.2333 2.3906 2.5985 2.7109 3.1354 3.2702

所得的结果见表 1。

绘制对参数 μ 和 σ 估计的相对误差分析图，分别见图 1 和图 2。

从以上两个图可以看出当样本容量 n 固定的时候，随着截尾样本数 r 的增大，采用次序截尾数据线性回归方法来估计对正态分布整体的均值和标准差的估计值的相对误差整体基本呈下降趋势，而且相对误差控制在 10%之内，符合实际应用中的估计要求。

5. 案例研究及结论

华北科技学院建工学院在 2013 年 5 月份派出 22 名学生参加了该校基础部组织的大学生数学建模比赛的选拔考试，赛后基础部只返回了获奖学生选手的名单及参赛成绩，而其他选手的成绩未出现，获奖名单及分数见表 2。

Table 1. Results of simulation
表 1. 模拟计算的结果

样本	μ 真值	$\hat{\mu}$ 估计值	$\frac{\lvert\mu-\hat{\mu}\rvert}{\mu}$	σ 真值	$\hat{\sigma}$ 估计值	$\frac{\lvert\sigma-\hat{\sigma}\rvert}{\sigma}$
n = 15 r = 5	2	1.808759	0.09562	0.8	0.58897	0.263788
n = 15 r = 6	2	2.088105	0.044052	0.8	0.824554	0.030693
n = 15 r = 7	2	1.956924	0.021538	0.8	0.704688	0.11914
n = 15 r = 8	2	1.873929	0.063035	0.8	0.622486	0.221892
n = 15 r = 9	2	1.999523	0.000238	0.8	0.757206	0.053492
n = 15 r = 10	2	1.974166	0.012917	0.8	0.727117	0.091104
n = 15 r = 11	2	1.985079	0.00746	0.8	0.741101	0.073623
n = 15 r = 12	2	2.003595	0.001797	0.8	0.767312	0.04086
n = 15 r = 13	2	1.99054	0.00473	0.8	0.747343	0.065822
n = 15 r = 14	2	2.01821	0.009105	0.8	0.796099	0.004877
n = 15 r = 15	2	2.003455	0.001727	0.8	0.762231	0.047211

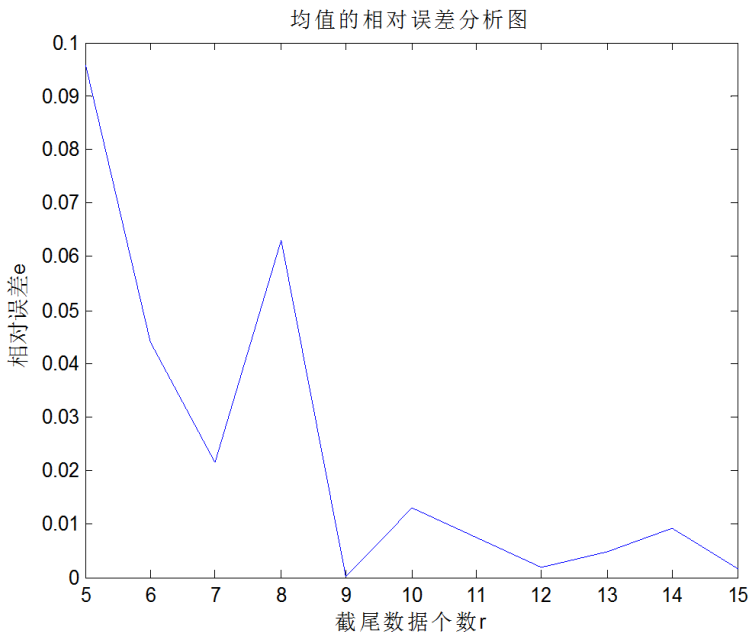

Figure 1. Relative error analysis of μ
图 1. μ 的相对误差分析图

为了解学生的学习状况，现欲利用次序截尾数据线性回归分析方法估计所有参赛选手的整体平均成绩。由经验知，学生成绩服从正态分布 $N\left(\mu,\sigma^2\right)$，现参赛学生人数为 n = 22，获奖学生个数为 r = 8，r 个学生的成绩为一组具体的次序截尾样本数据，用 $y_1 \geq y_2 \geq \cdots \geq y_r$ 表示。因为数学竞赛采用的是百分制，首先对成绩进行转换，令 $x_i = 100 - y_i$，$i = 1,2,\cdots,r$，则有 $x_1 \leq x_2 \leq \cdots \leq x_r$。由式(9)，可计算出 μ 和 σ 的 BLUE 为 $\hat{\mu} = 62.0898$，$\hat{\sigma} = 11.1909$。由于 $\hat{\mu}$ 是 X 的估计，将其进行转换，可得到整体成绩的均值 $\bar{y} = 100 - \hat{\mu} = 62.0898$。此成绩与后来与基础部落实的实际平均参赛成绩 60.8 较吻合，相对误差仅为 2%。

Table 2. Competition result
表 2. 竞赛成绩

姓名	性别	分数
杨涛	男	86
邓志明	男	81
李倩倩	女	72
李冠希	女	71
周振波	男	69
赵雅琼	女	69
牛亚超	男	68
葛志伟	男	67

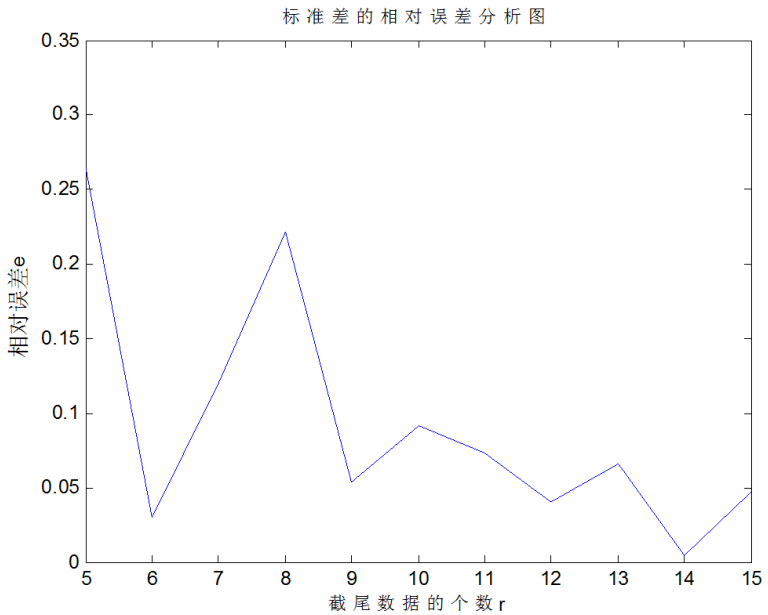

Figure 2. Relative error analysis of σ
图 2. σ 的相对误差分析图

6. 结语

本文讨论了次序统计量的期望和方差的计算公式，结合广义 Gauss-Markov 模型提出了一种次序截尾数据的线性回归分析方法，将只适用于完全数据的传统回归分析推广到了常见的次序截尾数据。通过计算机模拟计算发现，该方法对整体均值和标准差的估计值的相对误差整体基本呈下降趋势，而且相对误差控制在 10%之内，符合样本量越大估计越精准的事实，实际案例的应用也进一步验证了该方法的应用效果。

基金项目

国家级大学生创新创业训练计划项目(编号：201511104044)；华北科技学院教育科学研究课题基金资助(编号：HKJY201439)；华北科技学院应用数学重点学科资助项目(编号：HKXJZD201402)。

参考文献 (References)

[1] Jeandunn, O. and Aclark, V. (1987) Applied Statistics: Analysis of Variance and Regression. John Wiley & Sons, Inc., New York.

[2] Mao, S.S. and Wang, L.L. (1997) Accelerated Life Test. Science Press, Beijing. (In Chinese)

[3] 傅惠民, 黄伟. 最佳线性无偏整体估计方法[J]. 机械强度, 2003, 25(3): 319-324.

[4] 茆诗松, 王静龙, 濮晓松. 高等数理统计[M]. 北京: 高等教育出版社, 2006.

[5] 傅惠民, 林逢春. 大样本顺序统计量均值、方差和协方差计算与验证[J]. 机械强度, 2007, 29(1): 048-052.

[6] 王松桂, 史建红, 等. 线型模型引论[M]. 北京: 科学出版社, 2004.