Box-Cox 变换的黄金分割法实现*

刘瀛洲, 吴养会, 袁志发, 王乃信
(西北农林科技大学 生命科学学院, 陕西 杨陵 712100)

[摘 要] 提出可以使用黄金分割法确定Box-Cox 变换中的参数, 给出了一种计算步骤。提出了通过多搜索几个区间和回归诊断来更大限度地保证所确定出的参数是合用的。并从理论分析和实例检验两方面说明了与一维格点搜索法相比, 使用黄金分割法确定Box-Cox 变换中的参数计算量较小, 容易使计算结果达到较高的精确度, 且易用性也有所提高。

[关键词] 线性回归; Box-Cox 变换; 一维格点搜索法; 黄金分割法
[中图分类号] O 241 [文献标识码] A [文章编号]1000-2782(2001)04-125-03

通常在讨论线性回归模型时, 总要假定: 回归函数线性, 各次试验误差齐性, 各次试验误差相互独立以及各次试验误差服从正态分布[1]。但在实际问题中这些假定不一定都能得到满足。如果实际问题中这些假定的部分或全部得不到满足时, 在这些假定基础之上的各种讨论就根据不足了。在这种情况下, 为了继续使用上述假定基础之上的各种讨论结果, 可以使用Box 与Cox 提出的一类应用较广的Box-Cox 变换, 有望使最后的回归模型满足上述 4 项假定的要求[1~3]。要使用Box-Cox 变换, 首先必须对这个变换中的参数作出合适的估计。现有的确定Box-Cox 变换中参数的方法是一维格点搜索法[1~2]。Zarembka[4]、李选举[5]曾对如何快速方便地实现确定Box-Cox 变换中参数的一维格点搜索法作过研究。除此而外, 关于如何确定Box-Cox 变换中参数的方法, 未见有其他报道。黄金分割法是一种解最优化问题的直线搜索方法。本文将对如何使用黄金分割法确定Box-Cox 变换中的参数进行探讨, 旨在寻找出一种确定Box-Cox 变换中参数的新方法, 与一维格点搜索法相比, 新方法计算量较小, 容易使计算结果达到较高的精确度, 且易于使用。

1 Box-Cox 变换及参数的似然估计量

1.1 Box-Cox 变换

Box-Cox 变换为

$$y^{(\lambda)} = \begin{cases} \dfrac{y^{\lambda} - 1}{\lambda}, & \lambda \neq 0 \\ \ln(y), & \lambda = 0 \end{cases} \qquad (1)$$

其中 $y > 0$, λ 为待定参数。虽然此变换要求 $y > 0$, 但当此条件不满足时, 只要作一平移即可, 以下假定 $y > 0$。

1.2 λ 的似然估计量

考虑可观测的随机变量 y 关于一般变量 $x_1, x_2, \cdots, x_m$ $(m \geq 1)$ 的线性回归问题, 且有观测数据 $(x_{i1}, x_{i2}, \cdots, x_{im}, y_i)$, $i = 1, 2, \cdots, n$。对于任意的 $\lambda \in (-\infty, +\infty)$, 令

$$z_j^{(\lambda)} = \frac{y_j^{(\lambda)}}{\left(\prod_{i=1}^{n} y_i\right)^{\frac{\lambda-1}{n}}}, \quad j = 1, 2, \cdots, n \qquad (2)$$

其中 $y^{(\lambda)}$ 由(1)式计算, 则可建立 $z^{(\lambda)}$ 关于 $x_1, x_2, \cdots, x_m$ 的线性回归方程, 求出其残差平方和, 这里记为 $SSE(\lambda, z^{(\lambda)})$。如果 $\hat{\lambda}$ 使

$$SSE(\hat{\lambda}, z^{(\hat{\lambda})}) = \min\{SSE(\lambda, z^{(\lambda)}), \lambda \in (-\infty, +\infty)\},$$

则 $\hat{\lambda}$ 为(1)式中 λ 的似然估计量[1~3]。

2 用黄金分割法确定 λ 的算法

黄金分割法是适应面比较广的一种直线搜索方法, 适应于在搜索区间$[a, b]$上的任何单谷函数求极小值点的问题[6]。所以, 完全可以使用黄金分割法在给定的区间$[a, b]$上搜寻出一个 $\hat{\lambda}$, 并用这个 $\hat{\lambda}$ 作为 λ 的似然估计值。

* [收稿日期] 2000-09-07
[作者简介] 刘瀛洲(1964-), 男, 陕西泾阳人, 副教授, 硕士, 主要从事优化理论与决策研究。

参照实施黄金分割法的一般步骤[6]和笔者使用黄金分割法的经验, 下面列出使用黄金分割法在区间$[a, b]$上搜索参数λ的似然估计值的计算步骤。

①输入$(x_{i1}, x_{i2}, \cdots, x_{im}, y_i)$, $i= 1, 2, \cdots, n$;

②令$\beta= 0.5(\sqrt{5}- 1)$, $\gamma= 1- \beta$;

③确定λ的初始搜索区间$[a, b]$, 终止限ϵ;

④计算$\lambda_2= a+ \beta(b- a)$, $s_2= SSE(\lambda_2, z^{(\lambda_2)})$;

⑤计算$\lambda_1= a+ \gamma(b- a)$, $s_1= SSE(\lambda_1, z^{(\lambda_1)})$;

⑥如果$|\lambda_2- \lambda_1|< \epsilon$, 则输出$\hat{\lambda}= 0.5(\lambda_1+ \lambda_2)$; 否则, 即$|\lambda_2- \lambda_1|\geq \epsilon$, 转⑦;

⑦如果$s_1\leq s_2$, 则置$b= \lambda_2$, $\lambda_2= \lambda_1$, $s_2= s_1$, 然后转⑤; 否则, 即$s_1> s_2$, 则置$a= \lambda_1$, $\lambda_1= \lambda_2$, $s_1= s_2$, $\lambda_2= a+ \beta(b- a)$, $s_2= SSE(\lambda_2, z^{(\lambda_2)})$, 然后转⑥。

从理论上讲, 在上述算法步骤⑤中完全可以用$\lambda_1= a+ b- \lambda_2$代替$\lambda_1= a+ \gamma(b- a)$。因为$a+ \gamma(b- a)= a+ b- \lambda_2$。但如果这样, ④中$\lambda_2$所产生的误差将可能在⑤到⑦的循环中积累过大, 从而使⑥中的终止条件始终不能到达。

3 可能出现的问题及解决办法

至今, 还未有文献证明$SSE(\lambda, z^{(\lambda)})$为$(-\infty, +\infty)$上的单谷函数。如果$SSE(\lambda, z^{(\lambda)})$不是$(-\infty, +\infty)$上的单谷函数, 上述黄金分割法所得$\hat{\lambda}$不一定是全局最优的。其实, 一维格点搜索法也存在这一问题。笔者认为可以采取如下的办法来解决这一问题。

1) 可以多搜索几个区间(每个区间可小些), 然后取$SSE(\lambda, z^{(\lambda)})$的最小者所对应的$\lambda$作为$\hat{\lambda}$。

2) 求出$y\sim y^{(\hat{\lambda})}$与$x_1, x_2, \cdots, x_m$的线性回归方程$\hat{y}'= b_0'+ b_1'x_1+ \cdots+ b_m'x_m$和残差$e'= y'- \hat{y}'$, 然后利用回归诊断的各种方法对回归模型是否满足本文开头所列的 4 项假定进行诊断(如考察残差图$\hat{y}'$—e'等), 以确定在$\lambda= \hat{\lambda}$时 Box-Cox 变换是否合用。

4 与一维格点搜索法的比较

4.1 一维格点搜索法

确定λ的一维格点搜索法是, 先给出一系列的λ值(一般为具有相同间隔的一系列点), 对固定的λ值, 由(2)式计算出$z_j^{(\lambda)}$, $j= 1, 2, \cdots, n$, 求出相应的$SSE(\lambda, z^{(\lambda)})$。如果$\hat{\lambda}$使

$$SSE(\hat{\lambda}, z^{(\hat{\lambda})})= \min\{SSE(\hat{\lambda}, z^{(\hat{\lambda})}),$$

λ为给定的一系列值}

则认为这个$\hat{\lambda}$为要求的λ的似然估计值。

很明显, 寻找λ的一维格点搜索法的原理比较直观, 容易理解, 使用也比较方便; 但也存在着一些不能令人满意之处: 对于所取的一系列λ值, 如果相临λ值之间的间隔过大, 将可能导致寻找的$\hat{\lambda}$不够精确; 如果相临λ值之间的间隔过小, 又可能会导致大的计算量。

4.2 与一维格点搜索法的比较

4.2.1 易用性 当使用一维格点搜索法时, 如果要给的一系列λ值之间有规律可循(例如等距节点), 只要给出少量的几个数值, 计算机即可完成搜索。在这种情况下, 黄金分割法与一维格点搜索法在使用方便性上基本相同。因为在实施黄金分割法时, 只需给出搜索区间和终止限即可。当要给的一系列λ值之间无规律可循时, 使用一维格点搜索法就不如黄金分割法方便了。

4.2.2 计算量 用黄金分割法确定λ值是通过按照 0.618 的比率不断缩短搜索区间的办法来确定λ值的, 与之相比, 在相同计算精度要求下, 一维格点搜索法要计算更多次数的$SSE(\lambda, z^{(\lambda)})$。所以在相同的搜索区间和相同的搜索精度要求下, 黄金分割法一般比一维格点搜索法计算量小。

4.2.3 计算精确度 由本文前面的讨论知, 当$SSE(\lambda, z^{(\lambda)})$是区间$[a, b]$上的单谷函数时, 且使用一维格点搜索法所给的一系列λ值都处于区间$[a, b]$上时, 一定有

$$SSE(\lambda_1, z^{(\lambda_1)})\leq SSE(\lambda_2, z^{(\lambda_2)}), \quad (3)$$

其中, λ_1和λ_2分别为黄金分割法和一维格点搜索法在区间$[a, b]$上搜索到的λ值。因为

$$SSE(\lambda_1, z^{(\lambda_1)})= \min\{SSE(\lambda, z^{(\lambda)}), \lambda\in[a, b]\}。$$

当$SSE(\lambda, z^{(\lambda)})$不是区间$[a, b]$上的单谷函数时, 通过本文第 3 部分所提方法, 亦可有(3)式。可见黄金分割法比一维搜索法更容易使计算结果达到较高的精确度。

5 应用举例

表 1 列出了某地区所产原棉的纤维强力y与纤维的公制支数x_1、纤维的成熟度x_2的 28 组数据[1]。

对于表 1 中的数据, 在计算机上用 MATLAB

语言, 使用一维格点搜索法寻找 $\hat{\lambda}$, 为使精度达到小数点后 2 位, 取 $\lambda_i = -1 + 0.01i$, $i = 0, 1, \cdots, 600$, 结果得 $\hat{\lambda} = 2.09$, 费时 0.44 s; 同样, 在区间 $[-1, 5]$ 上寻找 $\hat{\lambda}$, 使用本文所提寻找 $\hat{\lambda}$ 的方法, 取 $a = -1$, $b = 5$, $\epsilon = 0.000\,1$, 亦可得 $\hat{\lambda} = 2.09$, 而显示时间却是 0。

表 1　原棉纤维强力数据

Table 1　Strength data of raw cotton fiber

序号 Order	公制支数 x_1/(m·g^{-1}) Metric count	成熟度 x_2 Maturity	纤维强力 y/g Strength	序号 Order	公制支数 x_1/(m·g^{-1}) Metric count	成熟度 x_2 Maturity	纤维强力 y/g Strength
1	5 415	1.58	4.03	15	6 208	1.70	3.81
2	5 700	1.38	4.01	16	5 798	1.59	4.00
3	5 674	1.57	4.00	17	5 551	1.61	4.19
4	5 698	1.55	4.09	18	6 089	1.57	3.81
5	6 165	1.52	3.73	19	6 060	1.53	3.96
6	5 929	1.60	4.09	20	6 059	1.55	3.93
7	7 505	1.14	2.95	21	6 370	1.45	3.72
8	5 920	1.50	3.90	22	6 102	1.49	3.84
9	7 646	1.18	2.89	23	6 245	1.50	3.88
10	6 556	1.27	3.48	24	6 644	1.45	3.38
11	6 475	1.50	3.60	25	6 191	1.58	3.76
12	5 907	1.50	3.77	26	6 352	1.50	3.79
13	5 697	1.54	3.94	27	5 999	1.59	3.79
14	6 618	1.20	3.66	28	5 815	1.70	4.09

注: 数据来源为文献[1]第 38 页。

Note: The data refer to the reference document No. 1, on page 38.

[参考文献]

[1] 周纪芗. 回归分析[M]. 上海: 华东师范大学出版社, 1993.

[2] William H G. 经济计量学[M]. 北京: 中国社会科学出版社, 1998.

[3] Box G, Cox D. An analysis of transformations[J]. Journal of the Royal Statistical Society, Series B, 1964: 211-264.

[4] Zarembka P. Functional form in the demand for money[J]. Journal of the American Statistical Association, 1968, 63: 502-511.

[5] 李选举. Box-Cox 变换及其在 MathCAD 上的实现[J]. 数量经济技术经济研究, 2000, 4: 42-44.

[6] 薛嘉庆. 最优化原理[M]. 北京: 冶金工业出版社, 1983.

[7] 张宜华. 精通 MATLAB5[M]. 北京: 清华大学出版社, 1999.

Golden section search for actualizing Box-Cox transformation

LIU Ying-zhou, WU Yang-hui, YUAN Zhi-fa, WANG Nai-xin

(*College of Life Sciences, Northwest Sci-Tech University of Agriculture and Forestry, Yangling, Shaanxi* 712100, *China*)

Abstract: This paper brings up that golden section search can be used to search for the parameter in Box-Cox transformation and offers a kind of steps in computing it. By searching for the parameter in more sections and using regression diagnosis, the established parameter is ensured to be more suitable. Theoretical analysis and experiments show: comparing with a dimension lattice search, the searching for the parameter in Box-Cox transformation by golden section search has less calculation, and it is easy to get more accurate result and is more applicable.

Key words: linear regression; Box-Cox transformation; a dimension lattice search; golden section search