要求回归因子显著时 Box-Cox 变换的用法*

刘迎洲[1], 仁俊杰[2], 李福生[3], 袁志发[1]

(1 西北农林科技大学 生命科学学院, 陕西 杨凌 712100;
2 西安财经学院 管理学院, 陕西 西安 710061;
3 陕西工运学院 教学部, 陕西 西安 710003)

[摘　要]　对在需要使用Box-Cox 变换, 且要求引入的回归因子均显著的情况下, 如何使用Box-Cox 变换的问题给出了 3 种方法, 并对 3 种方法及其计算结果的一些性质进行了理论上的分析与比较。结果表明, 3 种方法在计算量和计算精度上互有差异, 所估计的参数又各有其性质, 可根据实际情况选择使用。最后通过对泡桐片崩解试验数据的处理结果验证了 3 种方法的有效性。

[关键词]　Box-Cox 变换; 线性变换; 回归因子

[中图分类号]　O 212 1　　[文献标识码]　A　　[文章编号]　1671-9387(2004)08-0111-03

在实际应用中使用线性回归模型时, 经常会遇到回归模型不满足通常要求的 4 项假定: 回归函数线性, 各次试验误差线性, 各次试验误差相互独立以及各次试验误差服从正态分布。在这种情况下, 使用Box-Cox 变换可使回归模型满足上述 4 项假定[1]。另一方面, 在线性回归分析中, 经常要求回归模型中包含的回归因子都是显著的[2]。在需要使用Box-Cox 变换, 又要求引入的回归因子均显著的情况下, 如何使用Box-Cox 变换来达到这一要求的问题至今尚未见文献报道。本文给出了 3 种在一切情况下都有效的方法, 并对这些方法的性质及其使用效果进行了讨论, 最后给出了一个应用实例。

1　问题的提出

1. 1　使用Box-Cox 变换下的回归模型

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 $(x_{i1}, x_{i2}, \cdots, x_{im}, y_i)$, $i = 1, 2, \cdots, n$。设 y 关于 $x_1, x_2, \cdots, x_m$ $(m \geq 1)$ 的线性回归模型为

$$\hat{y} = b_{01} + b_{02}x_1 + \cdots + b_{0m}x_m, \tag{1}$$

当对 y 施行Box-Cox 变换

$$y^{(\lambda)} = \begin{cases} \dfrac{y^{\lambda} - 1}{\lambda}, & \lambda \neq 0 \\ \ln(y), & \lambda = 0 \end{cases} \tag{2}$$

后建立关于 $x_1, x_2, \cdots, x_m$ $(m \geq 1)$ 的线性回归模型为

$$\hat{y}^{(\lambda)} = b_{11} + b_{12}x_1 + \cdots + b_{1m}x_m, \tag{3}$$

其中的 λ 可由文献[1]或[3]所列方法估计。

1. 2　一个可能出现的问题

由模型(1)到模型(3), 关于回归因子显著性的变化情况未见有文献报道。本研究在使用Box-Cox 变换中发现, 有时在模型(1)中显著的因子, 在模型(3)中表现为不显著, 而在模型(1)中表现为不显著的因子, 在模型(3)中表现为显著。可见, 一般情况下, 先通过使用逐步回归选出回归因子, 再使用Box-Cox 变换达不到使进入最终模型的因子都显著的目的。

当要求在使用Box-Cox 变换时, 使进入回归模型的因子都显著的方法也未见文献报道。本研究给出 3 种可行的方法, 能够确保最后得到的回归模型达到这种要求, 并且由模型(1)到模型(3), 回归因子显著性如何变化都是适用的。

2　3 种可行的方法及其性质

2. 1　3 种可行的方法

2. 1. 1　方法 1　考虑把Box-Cox 变换与多元线性回归中的逐步回归法结合起来获取所要的模型。设 $x_1, x_2, \cdots, x_m$ $(m \geq 1)$ 为全部供选择的回归因子。则方法 1 的具体计算步骤如下:

首先, 给出一系列的 λ 值, 对固定的 λ 及观测数据 $(x_{i1}, x_{i2}, \cdots, x_{im}, y_i)$, $i = 1, 2, \cdots, n$, 如果 $\lambda \neq 0$, 令 $y^{(\lambda)} = y^{\lambda}$, 否则, 令 $y^{(\lambda)} = \ln x$。

其次, 对 $y^{(\lambda)}$ 及考虑的全部因子 $x_1, x_2, \cdots, x_m$ 使

* [收稿日期]　2003-06-24
[作者简介]　刘迎洲(1964-), 男, 陕西泾阳人, 副教授, 主要从事优化理论与决策研究。

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

用逐步回归法求出对应的回归方程。假设所求出的回归方程为

$$\hat{y}^{(\lambda)} = c_0^{(\lambda)} + c_1^{(\lambda)} x_{j_1}^{(\lambda)} + \cdots + c_s^{(\lambda)} x_{j_s}^{(\lambda)}, \quad (4)$$

其中 $x_{j_1}^{(\lambda)}, x_{j_2}^{(\lambda)}, \cdots, x_{j_s}^{(\lambda)}$ $(s \le m)$ 是 $x_1, x_2, \cdots, x_m$ 中的 s 个变量。同时还需求出

$$zsse(\lambda, y) = \sum (y_i^{(\lambda)} - y_i)^2, \quad (5)$$

其中

$$y_i^{(\lambda)} = \begin{cases} (\hat{y}_i^{(\lambda)})^{\frac{1}{\lambda}}, \lambda \neq 0 \\ \exp(\hat{y}_i^{(\lambda)}); \lambda = 0 \end{cases}$$

$$\hat{y}_i^{(\lambda)} = c_0^{(\lambda)} + c_1^{(\lambda)} x_{ij_1} + \cdots + c_s^{(\lambda)} x_{ij_s}, i = 1, 2, \cdots, n;$$

$(x_{ij_1}, x_{ij_2}, \cdots, x_{ij_s}, y_i)$ 为 $(x_{j_1}^{(\lambda)}, x_{j_2}^{(\lambda)}, \cdots, x_{j_s}^{(\lambda)}, y)$ 的第 i 组观测数据, $i = 1, 2, \cdots, n$。

最后, 将对不同的 λ 求得的 $zsse(\lambda, y)$ 列表或作图, 从中找出使 $zsse(\lambda, y)$ 达到最小的 λ, 则在上一步中对应于这个 λ 值使用逐步回归法求得的回归方程, 就是使用本方法得到的最后结果。

2.1.2 方法 2 考虑把 Box-Cox 变换与多元线性回归中求一切可能回归方程的方法结合起来获得所需的模型。设 $x_1, x_2, \cdots, x_m$ $(m \ge 1)$ 为全部回归因子, y 为因变量。以 A 表示由 $x_1, x_2, \cdots, x_m$ 为元素的集合, V 表示 A 的一切非空子集构成的集合, $T = \{x_{j_1}, x_{j_2}, \cdots, x_{j_s}\}$ 为 V 的任意一个元素。则方法 2 的计算步骤如下:

首先, 以 y 为因变量, T 中的元素 $x_{j_1}, x_{j_2}, \cdots, x_{j_s}$ 为自变量, 用文献[1]或[3]中的方法确定(3)式中的 λ, 如果 $\lambda \neq 0$, 则令 $y^{(\lambda)} = y^{\lambda}$, 否则, 令 $y^{(\lambda)} = \ln x$。

其次, 记上一步得到的 $y^{(\lambda)}$ 为 $y^{(T,\lambda)}$, 建立 $y^{(T,\lambda)}$ 关于 $x_{j_1}, x_{j_2}, \cdots, x_{j_s}$ 的回归方程,

$$\hat{y}^{(T,\lambda)} = c_0^{(T,\lambda)} + c_1^{(T,\lambda)} x_{j_1} + \cdots + c_s^{(T,\lambda)} x_{j_s}。 \quad (6)$$

同时求出式(6)中各个因子的回归系数的 F 检验值和残差平方和

$$zsse(T, \lambda, y) = \sum (y_i^{(T,\lambda)} - y_i)^2。 \quad (7)$$

其中,

$$y_i^{(T,\lambda)} = \begin{cases} (\hat{y}_i^{(T,\lambda)})^{\frac{1}{\lambda}}, \lambda \neq 0 \\ \exp(\hat{y}_i^{(T,\lambda)}); \lambda = 0 \end{cases}$$

$$\hat{y}_i^{(T,\lambda)} = c_0^{(T,\lambda)} + c_1^{(T,\lambda)} x_{ij_1} + \cdots + c_s^{(T,\lambda)} x_{ij_s}, i = 1, 2, \cdots, n,$$

$(x_{ij_1}, x_{ij_2}, \cdots, x_{ij_s}, y_i)$ 为 $(x_{j_1}^{(\lambda)}, x_{j_2}^{(\lambda)}, \cdots, x_{j_s}^{(\lambda)}, y)$ 的第 i 组观测数据, $i = 1, 2, \cdots, n$。

再次, 从上一步得到的一切回方程中选择回归因子全部显著的回归方程。

最后, 从上一步得到的结果中选择残差平方和 $zsse(T, \lambda, y)$ 最小的回归方程, 就是使用本方法得到的结果。

2.1.3 方法 3 考虑方法 1 的简化情形。首先, 以 y 为因变量, 以 $x_1, x_2, \cdots, x_m$ 为自变量, 使用文献[1]或[3]所列方法确定(3)式中的 λ, 如果所确定的 $\lambda \neq 0$, 则令 $y^{(\lambda)} = y^{\lambda}$, 否则, 令 $y^{(\lambda)} = \ln x$。

其次, 以 $y^{(\lambda)}$ 为因变量, 以 $x_1, x_2, \cdots, x_m$ 为自变量, 使用逐步回归法即可得到一个回归系数全是显著的回归方程。

2.2 3 种可行方法的比较及其性质

2.2.1 残差的比较 假设由方法 3 得到的回归方程如(4)式, 由(5)式计算的结果记为 $zsse(y^{(\lambda)}, \lambda)$, 则由文献[3, 4]的分析及各方法的计算过程可知

$$zsse(T, \lambda, y) \le zsse(\lambda, y) \le zsse(y^{(\lambda)}, \lambda)。$$

2.2.2 估计结果的一个性质 方法 1 以(5)式最小为目标, 所以方法 1 得到的回归方程中的系数是一种最小二乘估计; 方法 2 总是取特定自变量下 λ 的最大似然估计量, 所以方法 2 得到的回归方程中的系数是一种最大似然估计量[3]; 方法 3 中得到的 λ 对全部自变量来说是一种最大似然估计量, 但对最终经过逐步回归后得到的回归方程来说已不是最大似然估计量了。由于方法 3 是方法 1 的简化, 最终结果是方法 1 中供最后选择的结果中的一种, 所以最终结果中的 λ 也不是最小二乘估计量。但很明显, 方法 3 的计算量比其余两种方法都小。在实际应用中, 方法 3 有时能够满足要求。

3 在多项式回归上的应用

关于下面数据的拟合问题是一个多项式回归问题, 不但要解决回归系数的相关性问题, 而且还要解决方差的齐一性问题。所以, 可以使用本文第 2 部分所述的方法。

由表 1 数据[4]的散点图可知, y 关于 x 对数据的拟合为一多项式回归。当回归阶次为 6 时已出现过拟合, 于是以 x, x^2, x^3, x^4, x^5 为回归因子集进行逐步回归, 但发现结果的残差不能满足齐一性的要求。所以为获得较好的回归模型, 以 x, x^2, x^3, x^4, x^5 为回归因子集, 采用本文第 2 部分所述方法, 使用 MATLAB 可得 3 种方法的计算结果。

使用方法 1 的计算结果为(取 $\lambda = -1, -0.5, 0, 0.2, 0.4, 0.6, 0.7, 0.8, 0.9, 1, 2$)

$$y^{0.8} = 0.3910 + 0.0096x^2 - 0.0002x^3,$$

$$(130.672) \quad (41.287)$$

$zsse(\lambda, y) = 1.0434, F = 451.9802,$

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

$R^2 = 0.972$。

使用方法 2 的计算结果为

$$y^{0.2038} = 0.79436 + 0.00307x^2 - 1.0687e^{-5}x^4 + 2.8606e^{-7}x^5,$$

(142.79)　(31.361)　(19.915)

$zsse(T, \lambda, y) = 0.8147$, $F = 394.36$, $R^2 = 0.9793$。

使用方法 3 的计算结果为

$$y^{0.29836} = 0.68531 + 0.033835x,$$

(689.18)

$zsse(y^{(\lambda)}, \lambda) = 1.741$, $F = 689.18$, $R^2 = 0.9623$。

上述 3 个回归模型系数下括号中数值为该回归系数的 F 检验值。结果表明, $zsse(T, \lambda, y) \leq zsse(\lambda, y) \leq zsse(y^{(\lambda)}, \lambda)$, 3 个回归方程等号右端出现因子的回归系数均显著, 且回归模型本身也显著。

表 1　泡桐片的重量及其崩解所用时间的试验数据[4]

Table 1　Test data of weight and disintegration time of *Paulownia pill*

序号 Order	重量(x)/g Weight	崩解时间(y)/min Time of disintegration	序号 Order	重量(x)/g Weight	崩解时间(y)/min Time of disintegration	序号 Order	重量(x)/g Weight	崩解时间(y)/min Time of disintegration
1	0.10	0.29	11	6.00	2.39	21	16.00	0.71
2	0.30	0.32	12	7.00	2.40	22	17.00	1.15
3	0.50	0.34	13	8.00	2.68	23	18.00	1.19
4	0.70	0.36	14	9.00	3.31	24	19.00	1.78
5	0.80	0.53	15	10.00	3.40	25	20.00	2.09
6	1.00	0.67	16	11.00	0.31	26	21.00	2.56
7	2.00	0.65	17	12.00	0.34	27	22.00	2.35
8	3.00	0.99	18	13.00	0.42	28	23.00	2.57
9	4.00	1.31	19	14.00	0.32	29	24.00	3.46
10	5.00	2.37	20	15.00	0.35			

[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] 袁志发, 顾天骥. 概率基础与数理统计[M]. 北京: 农业出版社, 1987.

[3] 刘迎洲, 吴养会, 袁志发, 等. Box-Cox 变换的黄金分割法实现[J]. 西北农林科技大学学报(自然科学版), 2001, 29(4): 125-127.

[4] 罗延红. 除虫脲泡桐片剂的研制[D]. 陕西杨陵: 西北农林科技大学, 2000.

Methods of using Box-Cox transformation to meet the need of the factors being significant in the regression equation

LIU Ying-zhou[1], **REN Jun-jie**[2], **LI Fu-sheng**[3], **YUAN Zhi-fa**[1]

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

2 *Management College, Xi'an Finance and Economics College, Xi'an, Shaanxi* 710061, *China*;

3 *Shaanxi Trade Union Cadres' Training College, Xi'an, Shaanxi* 710003, *China*)

Abstract: In order to deal with some data, we need to use Box-Cox transformation and to meet the need of the factors being significant in the regression equation. In this situation, the question is how to use Box-Cox transformation. This paper presents three kinds of the methods that they can be used to solve the question, and studies some characters of the methods. Because of some differences in efficiency of computation, error of regression and characters estimate parameters, one of the methods should be selectted according to the factual situation and the need. The validity of the methods is demonstrated by the result to apply them in dealing with the data of Paulownia pill disintegration test.

Key words: Box-Cox transformation; linear regression; regression factor

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net