Statistics and Application 统计学与应用, 2021, 10(2), 278-283
Published Online April 2021 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2021.102027

Hans汉斯

基于水平镜像算法的改进Box-Cox变换

陈　鸿

云南师范大学数学学院，云南　昆明
Email: 1435227867@qq.com

收稿日期：2021年3月31日；录用日期：2021年4月15日；发布日期：2021年4月26日

摘　要

基于服从负偏态分布数据的一种水平镜像算法，本文提出一种改进Box-Cox变换——镜像Box-Cox变换，并进行数值实验，实验结果显示，与传统的Box-Cox变换相比较，镜像Box-Cox变换在处理正偏态分布数据的效果上与传统Box-Cox变换处理效果相同的基础上，其处理负偏态分布数据的效果要优于传统Box-Cox变换。再进行模拟回归模型实验，实验结果表明，经过镜像Box-Cox变换的数据建立的回归模型的拟合和预测效果有所提高，且效果优于使用传统Box-Cox变换后的数据。

关键词

Box-Cox变换，水平镜像算法，极大似然估计法

Improved Box Cox Transform Based on Horizontal Mirror Algorithm

Hong Chen

School of Mathematics, Yunnan Normal University, Kunming Yunnan
Email: 1435227867@qq.com

Received: Mar. 31st, 2021; accepted: Apr. 15th, 2021; published: Apr. 26th, 2021

Abstract

Based on a horizontal mirror algorithm for data with negative skew distribution, this paper proposes an improved Box-Cox transform: mirror Box-Cox transform, and carries out numerical experiments. The experimental results show that, compared with the traditional Box-Cox transform, mirror Box-Cox transform can process negative skewness on the basis of the same effect as the traditional Box-Cox transform. The effect of distributed data is better than that of traditional

文章引用: 陈鸿. 基于水平镜像算法的改进 Box-Cox 变换[J]. 统计学与应用, 2021, 10(2): 278-283.
DOI: 10.12677/sa.2021.102027

Box-Cox transform. Then the simulated regression model experiment is carried out. The experimental results show that the fitting and prediction effect of the regression model established by the mirror Box-Cox transformation data is improved, and the effect is better than the data after using the traditional Box-Cox transformation.

Keywords

Box-Cox Transform, Horizontal Mirror Algorithm, Maximum Likelihood Estimation Method

1. 引言

1.1. 研究背景

现实中，我们会遇到的数据纷繁复杂，不同的数据根据我们所做的假设的不同，需要进行不同的变换，以便我们能够在已有理论上对其进行分析。例如：股票收益率等数据的特殊性，不可观测的误差可能是和预测变量相关的，但其不服从正态分布，于是给线性回归的最小二乘估计系数的结果带来误差，为了满足线性回归的四个假设条件而又不丢失信息，有时需要对数据进行处理变换；又例如方差分析需要试验误差具有独立性、无偏性、方差齐性和正态性的条件，若不满足这些条件就需要对数据进行处理[1]。

Box-Cox 变换是 George Box 和 David Cox 在 1964 年提出的一种参数化广义幂变换方法[2]，其主要特点是引入一个参数 λ，通过数据本身估计该参数 λ，从而确定应采取数据变换形式[3]。常用于稳定方差、减少数据在统计建模中的非正态性和增强关联性度量的有效性。

基于正偏和负偏的数据互为镜像关系，本文提出的镜像 Box-Cox 变换，提高偏态分布数据正态化效果，且便于运算，并进行数值实验以验证该结论。

1.2. 正态性检验及回归模型评价指标说明

1) Shapiro-Wilk 检验[4] (W 检验)

W 检验是用来检验数据是否符合正态分布的。可计算得到一个相关系数，它越接近 1 就越表明数据和正态分布拟合得越好。且 W 检验还会给出一个 P 值，若 P 值大于 0.05，就无法拒绝其符合正态分布。若统计量 W 值接近 1，但 P 值小于 0.05，我们仍然拒绝其符合正态分布。W 检验计算公式为：

$$W = \frac{\left(\sum_{i=1}^{n} a_i y_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}$$

其中，y_i 为样本的次序统计量，a_i 为一个待估常量。

2) MAPE [5] (Mean Absolute Percentage Error，平均绝对百分比误差)

MAPE 常用于描述准确度，它是一个百分比值，因此比其他统计量更容易理解。MAPE 的值越小，说明预测模型拥有更好的精确度。其数学表达式为：

$$\text{MAPE}=\frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i-\hat{y}_i}{y_i}\right|\times 100\%$$

2. Box-Cox 变换

假设样本里一共有 n 个数据点，分别是：$y=(y_1,y_2,\cdots,y_n)'$，我们把变换后新的数据点记为：$y^{(\lambda)}=\left(y_1^{(\lambda)},y_2^{(\lambda)},\cdots,y_n^{(\lambda)}\right)'$

当 $y\geq 0$ 时，Box-Cox 变换是对原始数据做如下变换：

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

当存在 $y_i<0, i=1,2,\cdots,n$ 时，Box-Cox 变换是对原始数据进行如下变换：

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

λ 是一个待定变换参数，对不同的 λ，所做的变换自然就不同，所以这是一个变换族。我们将(1)式称为 Box-Cox 变换的基本公式；将(2)式称为 Box-Cox 变换的扩展公式。

3. 镜像 Box-Cox 变换

众所周知，在面对偏态分布数据时，我们需要使其变换为正态分布，要在保持数据大小次序不变的同时，减小数据之间的距离。对于处理正偏态分布的数据就需要一类增长率单调递减的函数，例如对数函数、平方根函数等等。而根据 Box-Cox 变换公式：

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

要使变换函数增长率单调递减，就需要变换的二阶导数小于 0，即：

$$\left(y^{(\lambda)}\right)''<0 \quad\Rightarrow\quad \begin{cases}\lambda(\lambda-1)y^{\lambda}<0, & \lambda\neq 0\\ -\dfrac{1}{y^2}<0, & \lambda=0\end{cases} \quad\Rightarrow\quad \lambda\in[0,1)$$

故 Box-Cox 变换在处理正偏态分布数据时，待定变换参数 $\lambda\in[0,1)$，同理可得，其在处理负偏态分布数据时，待定变换参数 $\lambda\in[1,+\infty)$。

所以，Box-Cox 变换在处理正偏态分布数据时，相较于其处理负偏态分布数据时，除了幂函数，还多了对数变换，且变换参数的取值范围较小，易于求解。

理论介绍

$\forall y_i\in\mathbb{R}$，镜像 Box-Cox 变换是对原始数据做如下变换：

$$y^{(\lambda)}=\begin{cases}\alpha\cdot\dfrac{(\alpha\cdot y+\beta)^{\lambda}-1}{\lambda}, & \lambda\neq 0\\ \alpha\cdot\log(\alpha\cdot y+\beta), & \lambda=0\end{cases} \tag{3}$$

若原始数据服从正偏态分布，且 $\forall y_i \geq 0$，则：$\alpha = 1$，$\beta = 0$；若原始数据服从正偏态分布，且 $\exists y_i < 0$，则：$\alpha = 1$，$\beta = \max(|y|)+1$；若原始数据服从负偏态分布，则：$\alpha = -1$，$\beta = \max(|y|)+1$ 其中，$\max(|y|)$ 表示原始数据取绝对值后，数据中的最大数。

无论选择传统的 Box-Cox 变换还是镜像 Box-Cox 变换，最关键的问题在于怎样选定一个最优的 λ，使得变换后的样本(及总体)正态性最好[6]。求解最优参数 λ，我们可以采用极大似然估计法和 Bayes 方法[6] [7]。

4. 数值实验

我们将分三种情况进行数值实验：1) 数据全为正数，2) 数据全为负数，3) 数据有一部分正数和一部分负数。

本节使用 Python 软件随机生成上述三种情况的负偏态分布[8]数据，之后将此类数据分别使用传统的 Box-Cox 变换和本文提出的镜像 Box-Cox 变换进行处理，对变换后的数据进行偏度、峰度[9]和 Shapiro-Wilk [10]检验，并画出数据的频率直方图和 P-P 图，据此可以比较两种变换的效果。

4.1. 符号说明

本节将实验中处理的不同数据使用不同的符号表示，便于之后的实验结果的描述，符号说明如表 1 所示。

Table 1. Symbol description
表 1. 符号说明

符号	说明
Ⅰ⁻型数据	表示全为正数的服从负偏态分布的原始模拟数据
Ⅱ⁻型数据	表示全为负数的服从负偏态分布的原始模拟数据
Ⅲ⁻型数据	表示部分负数、部分正数的服从负偏态分布的原始模拟数据
M*x* 数据	表示 *x* 型数据经过镜像 Box-Cox 变换后的数据，*x* 取 Ⅰ⁻、Ⅱ⁻、Ⅲ⁻
T*x* 数据	表示 *x* 型数据经过传统 Box-Cox 变换后的数据，*x* 取 Ⅰ⁻、Ⅱ⁻、Ⅲ⁻

4.2. 实验结果

4.2.1. 数据正态性检验图示结果

图 1 从左到右分别表示的是 Ⅰ⁻型、Ⅱ⁻型、Ⅲ⁻型、TⅠ⁻型、TⅡ⁻型、TⅢ⁻型、MⅠ⁻型、MⅡ⁻型和 MⅢ⁻型数据的直方图和 P-P 图。从图中可以看出经过镜像 Box-Cox 变换后的数据更加接近正态分布。

4.2.2. 数据正态性假设检验结果

从表 2 中可以看出无论是 Ⅰ⁻型、Ⅱ⁻型还是 Ⅲ⁻型数据，经过镜像 Box-Cox 变换后的数据在偏度、峰度和 W 值的表现均比传统 Box-Cox 变换后的数据更加接近正态分布。

5. 回归模型模拟

5.1. 实验分析

生成一组服从负偏态分布的回归因变量，数据中即包含正数、也包含负数，计算原始数据的偏度、峰度，并进行 Shapiro-Wilk 检验，之后对原始数据建立回归模型，进行预测并计算 MAPE；再使用传统 Box-Cox 变换及镜像 Box-Cox 变换对原始数据进行处理，计算经过变换后的数据的偏度、峰度，并进行

Shapiro-Wilk 检验，之后使用处理后的数据建立回归模型，进行预测并计算 MAPE。

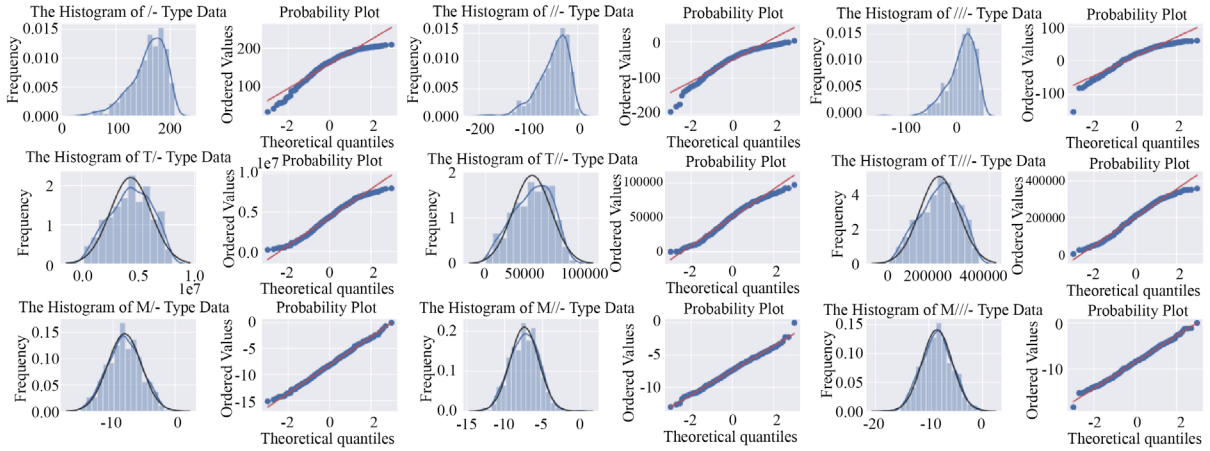

Figure 1. The histogram and P-P plot of three type data
图 1. 三类数据的频率直方图和 P-P 图

Table 2. Data normality test results and optimal parameters (λ)
表 2. 数据正态性检验结果及最优参数(λ)

统计量 / 数据类型	偏度(skewness)	峰度(kurtosis)	*W* 值	*P* 值	最优参数(λ)
Ⅰ⁻型数据	−0.9919	0.8031	0.9327	3.213e−14	
TⅠ⁻型数据	−0.2097	−0.7730	0.9805	3.0208e−06	2.5746
MⅠ⁻型数据	0.0256	−0.2940	0.9974	0.6136	0.3642
Ⅱ⁻型数据	−1.1063	1.5032	0.9277	8.4380e−15	
TⅡ⁻型数据	−0.2745	−0.7039	0.9809	3.8808e−06	2.3344
MⅡ⁻型数据	−0.0098	0.0173	0.9967	0.4014	0.3023
Ⅲ⁻型数据	−1.0621	1.7657	0.9346	5.4640e−14	
TⅢ⁻型数据	−0.2344	−0.6995	0.9827	1.1579e−05	2.5655
MⅢ⁻型数据	0.0170	−0.1521	0.9977	0.7196	0.3909

5.2. 实验结果

从表 3 可以得到经过镜像 Box-Cox 变换后的数据更接近正态分布，并且其预测值与实际值的平均绝对百分比误差(RMSE)为 2.802%，远小于传统 Box-Cox 变换后的数据的 70.77%。

Table 3. Data normality test results and model fitting effect evaluation
表 3. 数据正态性检验结果及模型拟合效果评价

统计量 / 数据类型	偏度 (skewness)	峰度(kurtosis)	*W* 值	*P* 值	RMSE
原始数据	−1.1638	2.4959	0.9379	5.342e−20	11.15%
传统 Box-Cox 变换数据	−0.0543	−0.1824	0.9983	0.4239	70.77%
镜像 Box-Cox 变换数据	0.0062	0.1481	0.9987	0.7126	2.802%

6. 结语

通过数值实验，我们发现无论何种类型(全为正数、全为负数和部分正数、部分负数)的负偏态分布数据经过镜像 Box-Cox 变换后，数据基本服从正态分布，且效果要优于使用传统 Box-Cox 变换。

进行模拟回归模型实验结果表明，使用传统 Box-Cox 变换后的数据建立的回归模型，进行预测后其 RMSE 为 70.77%；使用镜像 Box-Cox 变换后的数据建立的回归模型，进行预测后其 RMSE 为 2.802%，经过镜像 Box-Cox 变换的数据建立的回归模型的拟合和预测效果有所提高，且效果优于使用传统 Box-Cox 变换后的数据。

本文提出的镜像 Box-Cox 变换通过计算机易于实现，效果相较传统 Box-Cox 变换有所提高，可以作为处理非正态数据的一种可靠的方法。

参考文献

[1] 张彦玲. 处理非正态数据[J]. 中国质量, 2002(8): 22-24.

[2] Box, G. and Cox, D. (1964) An Analysis of Transformations (with Discussion). *Journal of the Royal Statistical Society*, *Series B*, **26**, 211-252. https://doi.org/10.1111/j.2517-6161.1964.tb00553.x

[3] 王松桂, 陈敏, 陈立萍. 线性统计模型——线性回归与方差分析[M]. 北京: 高等教育出版社, 1999: 52-55.

[4] Shapiro, S.S. and Wilk, M.B. (1965) An Analysis of Variance Test for Normality (Complete Samples). *Biometrika*, **52**, 591-611. https://doi.org/10.1093/biomet/52.3-4.591

[5] Hyndman, R.J. and Koehler, A.B. (2006) Another Look at Measures of Forecast Accuracy. *International Journal of Forecasting*, **22**, 679-688. https://doi.org/10.1016/j.ijforecast.2006.03.001

[6] 钟登华, 刘豹. Box-Cox 变换模型参数估计方法研究[J]. 系统工程学报, 1993, 8(2): 40-46.

[7] 胡宏昌, 樊献花, 等. 广义 Box-Cox 变换[J]. 周口师范学院学报, 2006, 23(5): 17-18.

[8] Azzalini, A. and Capitanio, A. (1999) Statistical Applications of the Multivariate Skew-Normal Distribution. *Journal of the Royal Statistical Society*: *Series B*, **61**, 579-602. https://doi.org/10.1111/1467-9868.00194

[9] 茆诗松, 周纪芗. 概率论与数理统计[M]. 北京: 中国统计出版社, 2013: 260-262, 420-422.

[10] Rigby, R.A. and Stasinopoulos, D.M. (2010) Smooth Centile Curves for Skew and Kurtotic Data Modelled Using the Box-Cox Power Exponential Distribution. *Stats in Medicine*, **23**, 3053-3076. https://doi.org/10.1002/sim.1861