第 20 卷第 1 期　　南 京 气 象 学 院 学 报　　Vol. 20 No. 1
1997 年 3 月　　Journal of Nanjing Institute of Meteorology　　Mar. 1997

带有变量中误差非正态模型中的 Box-Cox 变换

吕纯濂[1)]　　陈舜华[2)]

(1)南京气象学院基础科学系; 2) 南京气象学院大气物理系, 南京　210044)

摘　要　Box-Cox 变换主要用于得到近似正态分布的随机变量。线性模型中的响应(因)变量 Y (或经过变换的 Y) 通常建立在精确测量无误差的独立变量集上。本文研究独立变量中的测量误差(即变量中误差) 对Box-Cox 变换中参数估计的影响。研究结果表明, 忽略变量中的误差将产生变换参数估计的偏离, 该结论由分析论证和随机模拟得以支持。

关键词　Box-Cox 变换, 变量中误差, 非正态, 极大似然估计, 随机模拟

分类号　分类号 O212.4

线性模型响应变量 Y 中的随机干扰 e, 通常假定是服从正态分布的。若不满足该条件, 就要用响应变量的某些变换使之满足。常用的变换是 Box 和 Cox 提出的[1], 尽管许多学者对其作了新的改进[2], 但至今最常用的仍是他们最初提出的方法。

通常假定, 在 Box-Cox 方法中的线性模型预报因子变量是精确测量或无明显误差的。当预报因子变量中出现测量误差时, 就要考虑测量误差的技巧[3~7]。本文的目的主要研究测量误差对正值响应变量 Y 作 Box-Cox 变换时的影响。发现, 当忽略预报因子变量中的误差时, Box-Cox 方法中参数的极大似然估计将产生偏离, 但若考虑测量误差时, 则 Box-Cox 方法仍有效。

1　背景和方法

1.1　Box-Cox 变换

Box-Cox 变换广泛地应用于正值响应变量 Y 的正态化, Box-Cox 变换为:

$$W = \begin{cases} \dfrac{Y^{\lambda} - 1}{\lambda}, & \lambda \neq 0 \\ \ln Y, & \lambda = 0 \end{cases} \tag{1}$$

其中, λ 即为 Box-Cox 变换参数。

Box 和 Cox 讨论了在线性模型

$$\boldsymbol{W} = \boldsymbol{X}\beta + \boldsymbol{e} \tag{2}$$

式中, λ, β 和 σ_e^2 的极大似然估计[1], 其中 $\boldsymbol{W} = (W_1, W_2, \cdots, W_n)'$, $\boldsymbol{X}$ 为 $n \times p$ 回归变量矩阵, β 为 $p \times 1$ 未知回归系数向量, $\boldsymbol{e} \sim N(0, \sigma_e^2 \boldsymbol{I})$。

1.2　变量中的误差

假定在模型(2) 的预报因子变量的测量中出现随机误差, 即

收稿日期: 1996- 01- 19; 改回日期: 1996- 06- 14

$$\boldsymbol{X} = \boldsymbol{x} + \delta_{\boldsymbol{x}}$$

其中$\boldsymbol{x}$为不可观测的真正的非随机的预报因子变量矩阵, $\delta_{\boldsymbol{x}}$为随机误差矩阵, 假定$\boldsymbol{w}$与$\boldsymbol{x}$间存在真正的线性关系, 方程中无误差, 即

$$\left.\begin{aligned} &\boldsymbol{w}_i = \boldsymbol{x}_i\boldsymbol{\beta} \\ &(\boldsymbol{W}_i, \boldsymbol{X}_i) = (\boldsymbol{w}_i, \boldsymbol{x}_i) + (\boldsymbol{e}_i, \delta_{\boldsymbol{x}_i}), \epsilon_i = (\boldsymbol{e}_i, \delta_{\boldsymbol{x}_i}) \end{aligned}\right\} \quad (i = 1, \cdots, n) \quad (3)$$

该模型属函数变式[5, 7], 对这种变式中的参数估计, 通常假定测量误差协方差阵已知, 该假定亦适用于带有变量中误差的 Box-Cox 变换, 而模型误差$\boldsymbol{e}$的方差未知。

1.3 估计结果

现考虑带有变量中误差又经过 Box-Cox 变换的线性模型中的参数估计。

假定模型(3)中

$$\epsilon \sim N(0, \Sigma_{\epsilon\epsilon}), \Sigma_{\epsilon\epsilon} = \begin{pmatrix} \sigma^2 & 0 \\ 0 & \Gamma_{xx} \end{pmatrix}$$

其中, Γ_{xx}为已知的预报因子变量中误差协方差阵; σ^2为未知的模型方差。令矩估计量

$$\hat{\Gamma}_{xx} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i'\boldsymbol{X}_i, \quad \hat{\Gamma}_{xw} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i'\boldsymbol{W}_i, \quad \hat{\sigma}_{ww} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{W}_i^2$$

在模型(3)中, 假定所有矩估计量皆以概率 1 收敛, 则对$\boldsymbol{\beta}$、σ^2和$\boldsymbol{x}$有以下矩估计量[5, 7]

$$\left.\begin{aligned} &\hat{\boldsymbol{\beta}}(\lambda) = (\hat{\Gamma}_{xx} - \Gamma_{xx})^{-1}\hat{\Gamma}_{xw} \\ &\hat{\sigma}^2(\lambda) = \hat{\sigma}_{ww} - \hat{\boldsymbol{\beta}}'\hat{\Gamma}_{xw} = \hat{\sigma}_{ww} - \hat{\Gamma}_{xw}'(\hat{\Gamma}_{xx} - \Gamma_{xx})^{-1}\hat{\Gamma}_{xw} \\ &\hat{\boldsymbol{x}}_i(\lambda) = (\boldsymbol{W}_i\hat{\boldsymbol{\beta}}' + \hat{\sigma}^2\boldsymbol{X}_i\Gamma_{xx}^{-1})(\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}' + \sigma^2\Gamma_{xx}^{-1})^{-1}, \quad i = 1, \cdots, n \end{aligned}\right\} \quad (4)$$

(4)中后两式中的$\boldsymbol{\beta} = \boldsymbol{\beta}(\lambda)$, $\sigma^2 = \sigma^2(\lambda)$。由(4)可知, 带有变量中误差且因变量又经 Box-Cox 变换的线性模型中的参数估计是变换参数λ的函数。

与一般论述方法不同, 先作随机模拟, 再对模拟结果作理论分析。

2 随机模拟

本文设计了一个随机模拟试验, 用来评价和比较在 Box-Cox 变换中考虑或不考虑变量中误差的差异。

由于对所设计的回归函数、密度、样本容量、模型误差和测量误差大小, 都有许多可能的选择, 为能说明问题, 必须作一定的限制。本文仅讨论带有变量中误差的两个独立连续预报因子变量的线性模型, 两个预报因子变量的测量误差方差已知, 对连续正值因变量Y作 Box-Cox 变换(实际上, 是对已产生的适当正态随机数W, 相应给定的不同变换参数λ, 作 Box-Cox 逆变换而得到Y)。

用对每个预报因子变量的可靠性比r确定测量误差的大小, 可靠性比r是真正预报因子变量与可观测预报因子变量的方差比。用百分比误差(100(1-r)%)表示测量误差大小。我们在随机模拟中, 可靠性比r分别选了 0.95, 0.90, 0.85, 0.80, 0.75, 0.70, 0.65 和0.60, 故相应的百分比误差为 5%, 10%, 15%, 20%, 25%, 30%, 35% 和 40%。

对以下模型作了 500 次随机模拟试验, 它们的样本容量分别为 50, 100, 500。

$$\left.\begin{aligned} W_i &= \left\{\begin{array}{ll} \dfrac{Y_i^{\lambda}-1}{\lambda}, & \lambda\neq 0 \\ \ln Y_i, & \lambda=0 \end{array}\right\} = \beta_1 x_{1i} + \beta_2 x_{2i} + e_i \\ \boldsymbol{X}_i &= (x_{1i}, x_{2i}) + (\delta_{x_1 i}, \delta_{x_2 i}) = \boldsymbol{x}_i + \delta_{xi},\ i=1,\cdots,n \end{aligned}\right\} \tag{5}$$

试验时, 分别取 $\lambda=-1, 0$(对数变换), 1, 2; $x_1 \sim N(50,\ r/(1-r))$, $x_2 \sim N(50, r/(1-r))$ 相互独立; $e_i \sim N(0, \sigma^2)$, $\delta_x \sim N(0, \boldsymbol{I})$, $i=1,\cdots,n$, $r=$ 可靠性比, $\beta_1=\beta_2=0.7$, $\sigma^2=1$。即 x_1, x_2 分别由均值为 50、方差为 $r/(1-r)$ 的正态分布产生。重点是比较模型中考虑和不考虑变量中误差所得的不同结果。为简单起见, 取 x_1, x_2 同分布和 $\beta_1=\beta_2$, 这样, 模型既简单又不影响重点问题的说明和论证, 而误差 e 由均值为 0、方差为 1 的正态分布产生。x_1, x_2 均值取 50 是为保证 Y_i 取正值, 若产生了负的 Y_i(或小于 0.01%), 则舍弃而代之以新产生的值。在 486 微机上, 在 GAUSS 矩阵语言系统支持下进行随机模拟试验, 正态随机数由过程 RNDN 产生。

在无变量中误差的模型(2) 中, λ 的极大似然估计 $\hat{\lambda}$ 由解下式而得

$$\max_{\lambda}\{-\frac{n}{2}\log\hat{\sigma}_e^2(\lambda) + (\lambda-1)\sum_{i=1}^{n}\log Y_i\} \tag{6}$$

其中常数项已省略, 而

$$\hat{\sigma}_e^2 = \frac{\boldsymbol{W}'\{\boldsymbol{I}-\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\}\boldsymbol{W}}{n}$$

在有变量中误差的模型(3) 中, 由第二节可知, λ 的极大似然估计由解下式而得

$$\max_{\lambda}\{-\frac{n}{2}\log\hat{\sigma}^2(\lambda) - \frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{X}_i-\hat{\boldsymbol{x}}_i(\lambda))\hat{\Gamma}_{xx}^{-1}(\boldsymbol{X}_i-\hat{\boldsymbol{x}}_i(\lambda))' + (\lambda-1)\sum_{i=1}^{n}\log Y_i\} \tag{7}$$

其中常数项也已省略, 而

$$\hat{\sigma}^2 = \hat{\sigma}_{ww} - \hat{\beta}\hat{\Gamma}_{xw}$$

(7) 式中的中间项指出了估计真正不可观测预报因子变量 $\boldsymbol{x}_i$, $i=1,\cdots,n$ 对极大似然函数的贡献, 它渐近地趋于常数, 故当大样本时可忽略之。

图 1 仅给出了样本容量分别取 $n=50$, 500 和 $\lambda=-1, 0$ (对数变换), 1 和 2 时, Box-Cox 变换参数 λ 极大似然估计的 500 次模拟数据平均值关于不同测量误差百分比的关系图。虚线是用第二节所述的估计方法对带有变量中误差的模型估计 λ 的情况, 实线是不考虑变量中误差模型的相应情况。图中在平均值上的纵短线为 95% 的置信区间。由图可知, 在 $\lambda=-1$, 1 和 2 时, 随着测量误差百分比从 0 到 40% 的增加, 不考虑测量误差模型 λ 的估计 $\hat{\lambda}$ 相对于 λ 真值的偏离 $\delta_\lambda(r)$ 从 5% 增加到 70%, 而考虑测量误差模型 λ 的估计基本上在 λ 的真值附近, 几乎没有偏离, 在 $\lambda=0$ (对数变换), 两者基本一致。因此, 若用 Box-Cox 变换的对数变换($\lambda=0$) 时, 考虑测量误差与否对参数估计无多大影响, 但若用 Box-Cox 变换的非对数变换时, 则应考虑用带有变量中误差的模型。

3　结果分析

对以上模拟结果, 可以分别用预报因子变量真值求 λ 的极大似然估计和用带有变量中误差的观测值求 λ 的极大似然估计。对这两种情况下所得到的极大似然估计用分析方法作比较, 就可明白为什么在不考虑变量中误差的线性模型中, Box-Cox 变换参数 λ 的极大似然估计相对于其真值, 对应于变量中误差线性模型的不同误差存在不同偏离 $\delta_\lambda(r)$。假设预报因子真值模型为

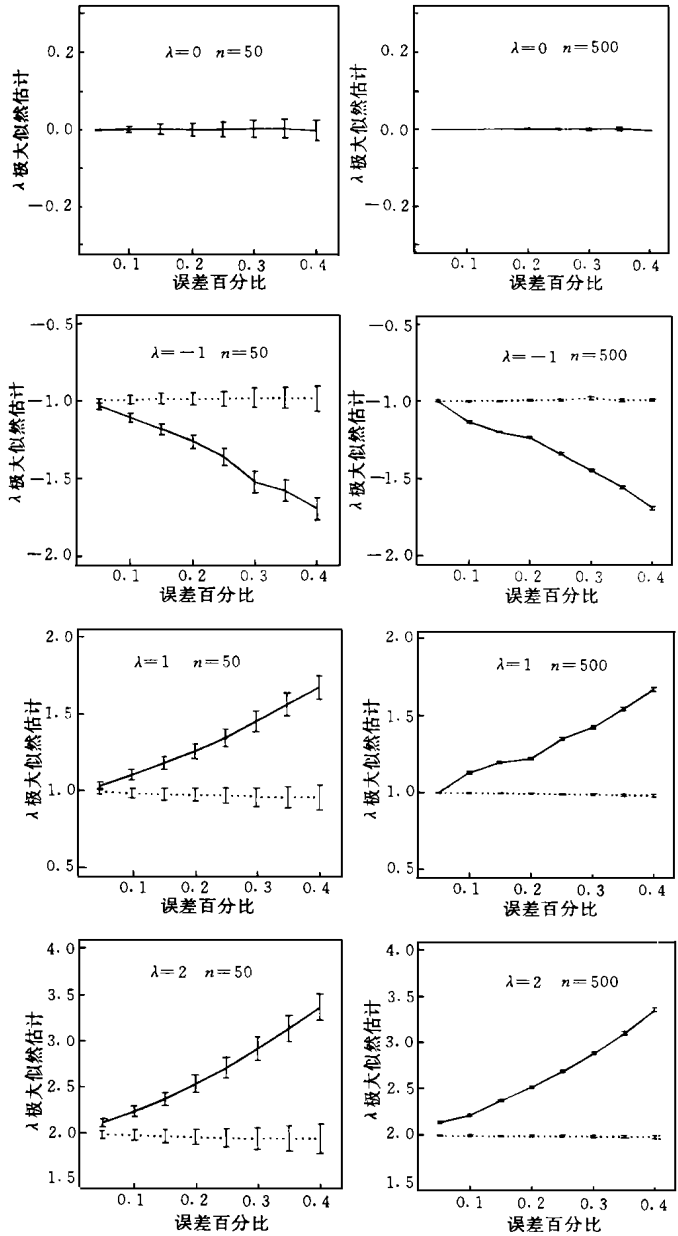

图 1 Box-Cox 变换参数 λ 极大似然估计平均值和误差百分比
(虚线带有变量中误差, 实线不带有变量中误差)

Fig. 1 Mean values estimated by maximum likelihood of Box-Cox parameter λ and the percentage errors, with errors (no errors) in the variable included denoted by dotted (solid) line

$$\boldsymbol{W}=\left\{\begin{array}{ll}\dfrac{Y^{\lambda}-1}{\lambda}, & \lambda\neq 0\\ \ln Y, & \lambda=0\end{array}\right\}=\boldsymbol{x}\boldsymbol{\beta}+\boldsymbol{e} \tag{8}$$

而带有变量中误差的模型为

$$\boldsymbol{W}=\left\{\begin{array}{ll}\dfrac{Y^{\lambda}-1}{\lambda}, & \lambda\neq 0\\ \ln Y, & \lambda=0\end{array}\right\}=\boldsymbol{X}\beta_2+\boldsymbol{e}_2 \tag{9}$$

$\boldsymbol{x}$ 为预报因子真值变量矩阵, 而 $\boldsymbol{X}=\boldsymbol{x}+\delta_x$ 为带有测量误差的预报因子观测矩阵, β_1, β_2 皆为 $p\times 1$ 未知回归系数向量, $\boldsymbol{e}_1\sim N(0,\sigma_1^2\boldsymbol{I})$, $\boldsymbol{e}_2\sim N(0,\sigma_2^2\boldsymbol{I})$, 对应 (8) 和 (9) 关于 λ 的极大似然方程分别为

$$\max_{\lambda}\{-\frac{n}{2}\log\hat{\sigma}_1^2(\lambda)+(\lambda-1)\sum_{i=1}^{n}\log Y_i\} \tag{10}$$

$$\max_{\lambda}\{-\frac{n}{2}\log\hat{\sigma}_2^2(\lambda)+(\lambda-1)\sum_{i=1}^{n}\log Y_i\} \tag{11}$$

由上两式可知, 对(10)和(11) 的比较可简化为

$$\hat{\sigma}_1^2(\lambda)=\frac{\boldsymbol{W}'(\boldsymbol{I}-\boldsymbol{x}(\boldsymbol{x}'\boldsymbol{x})^{-1}\boldsymbol{x}')\boldsymbol{W}}{n}$$

和

$$\hat{\sigma}_2^2(\lambda)=\frac{\boldsymbol{W}'(\boldsymbol{I}-\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')\boldsymbol{W}}{n}$$

的比较。为此, 利用渐近性概念, 由 (4) 和大数定律得

$$P\lim_{n\to\infty}\hat{\sigma}_1^2(\lambda)=\sigma_{ww}(\lambda)-\Gamma_{xw}'(\Gamma_{xx}-\Gamma_{xx})^{-1}\Gamma_{xw} \tag{12}$$

$$P\lim_{n\to\infty}\hat{\sigma}_2^2(\lambda)=\sigma_{ww}(\lambda)-\Gamma_{xw}'\Gamma_{xx}^{-1}\Gamma_{xw} \tag{13}$$

其中, Γ_{xx}、Γ_{xw}和 σ_{ww}分别为 $\hat{\Gamma}_{xx}$、$\hat{\Gamma}_{xw}$和 $\hat{\sigma}_{ww}$的相应渐近极限, 在随机模拟时, 令预报因子变量间是相互独立的(在线性模型 (2) 中, 若假定预报因子变量间相互独立, 即使矩估计量 $\hat{\Gamma}_{xx}$不一定为对角阵, 但其渐近极限 Γ_{xx}在变量独立的假定下应为对角阵, 因而 Γ_{xx}也是对角阵) 。因此 Γ_{xx}和 Γ_{xx}为对角阵, 则 Γ_{xx}的元素比($\Gamma_{xx}-\Gamma_{xx}$) 的相应元素小, 故在模型(9) 中, $\hat{\sigma}_2^2(\lambda)$ 的渐近极限是参数 λ 的函数, 它 "偏移 "至 $\hat{\sigma}_1^2(\lambda)$ 的渐近极限相应参数的右边。而这种 "偏移 "导致了在不考虑变量中误差线性模型中 Box-Cox 变换参数 λ 的极大似然估计对应于考虑变量中误差的线性模型中不同误差的不同偏离 $\delta_\lambda(r)$。

由随机模拟可知, 带有变量中误差线性模型中 Box-Cox 变换参数 λ 的极大似然估计 $\hat{\lambda}$, 不论百分比误差多大, 对应其真值几乎无偏离 $\delta_\lambda(r)\approx 0$(见图 1 的虚线) ; 而不带有变量中误差线性模型中 Box-Cox 变换参数 λ 的极大似然估计 $\hat{\lambda}$, 百分比误差 ($100(1-r)\%$) 越大, 则相应的偏离 $\delta_\lambda(r)$ 也越大 (见图 1 的实线) 。

参 考 文 献

1 Box G E P, Cox D R. An analysis of transformations. J Roy Statist. 1964, Soc Ser B. 26, 211～252

2 Hamilton W A, Taylor J M G. A comparison of the Box-Cox transformation method and nonparametric methods for estimating quantiles in clinical data with repeated measures. J Statist Comput Simulation, 1993, 45: 185～201

3 吕纯濂, 陈舜华, 张荣庆. 带有变量中误差的线性模型. 数学的实践与认识, 1994, (1): 6～15

4 吕纯濂, 陈舜华. 带有变量中误差的 Probit 回归模型. 系统科学与数学, 1993, 13(4): 338～344

5 吕纯濂, 陈舜华. 带有变量中误差的 Logit 模型及其计算. 数值计算与计算机应用, 1994, 15(2): 130～137

6 L Chunlian, Chen Shunhua. The Logit Regression Model with Errors in Variables and its Calculation. Chinese Journal of Numerical Mathematics and Applications, 1994, 16(4): 29～38

7 Schneeweiss H, Mittag H J. Lineare Modelle mit Fehlerbehafteten Daten. Heidelberg Wien: Physica-Verlag, 1986, 1～504

BOX-COX TRANSFORMATION USED IN NON-NORMAL MODEL WITH ERRORS-IN-VARIABLES

L Chunlian

(Department of Basic Courses, NIM, Nanjing 210044)

Chen Shunhua

(Department of Atmospheric Physics, NIM, Nanjing 210044)

Abstract The Box-Cox transformation is widely used for obtaining an approximate normally-distributed random variable. In a linear model setting, a responsive dependent variable Y (or transformed Y) is built on a set of independent variables, usually assumed to be measured with no error. This paper investigates the effects of measurement error in independent variables (known as errors-in-variables) on parameter estimate for the transformation. Ignoring such error in independent variables will cause biased transformation parameter estimate in general, a conclusion that is supported by analytical arguments and stochastical simulation.

Keywords Box-Cox transformation, errors-in-variables, non-normal, maximum likelihood estimation, random simulation