文章编号:1004-4337(2009)05-0569-03　　中图分类号:R195.1;TP311.54　　文献标识码:A　　·微机应用·

Box-Cox变换及其在SPSS软件中的实现△

李运明　封宗超*　李小凯**　孙　娜　许　贵*　马　兴　倪　静

(成都军区总医院质量管理科　成都 610083)

摘　要: 目的:针对生物医学研究中Box-Cox变换问题,给予统计软件技术上的支持。方法:采用SPSS软件实现Box-Cox变换。结果:给出了Box-Cox变换的SPSS程序,并进行实例分析,估计变换参数λ。结论:给出的SPSS程序适用于Box-Cox变换。

关键词: 数据变换;　Box-Cox变换;　SPSS程序;　极大似然估计

生物医学数据的统计分析,常会遇到连续型反应变量分布不满足正态分布的情况[1]。而统计分析最常用的一般线性模型假定:$Y=X\beta+\varepsilon,\varepsilon\sim N(0,\sigma^2 I)$。其中,$Y$为反应变量的向量,$X$为设计矩阵或观测矩阵,$\beta$为未知参数向量,且要求误差项$\varepsilon$服从正态分布。但是,由于生物医学数据的特殊性,$\varepsilon$常不服从正态分布,不能直接应用一般线性模型进行数据分析。为了使ε满足正态分布,常对Y进行变量变换。变量变换的方法很多种,Box-Cox变换是其中一种常用的变换[2]。Box-Cox变换自从1964年由Box和Cox提出后,常被用于变量变换,可使得ε服从正态分布。目前,国内统计学者已采用SAS、STATA等软件对Box-Cox变换的实现问题进行了探讨[3~7],未见采用SPSS实现Box-Cox变换的报道。SPSS(Statistical Product and Service Solution)是国际上流行并具有权威性的统计分析软件之一[8],由于其易于操作而成为非统计专业人员应用最多的统计软件。因此,本研究将介绍Box-Cox变换中参数λ的估计及其在SPSS软件中的实现方法。

1　Box-Cox变换

Box-Cox变换是对反应变量y进行变换。$y(y>0)$的Box-Cox变换可用下面的公式(1)表示:

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

显然,y的Box-Cox变换是一个变换族,由可变参数λ决定具体变换的形式。当$\lambda=0$时,该变换为对数变换。更一般的,对于任意取值y的Box-Cox变换可用下面的公式(2)表示:

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

其中,参数c是为了使$y+c>0$;g的默认取值为1,g可取$\dot{Y}^{\lambda+1}$。为$\dot{Y}$的几何均数。显然参数c的取值很好确定,公式(1)、(2)中需要估计的参数为λ。为了简便表述参数λ的估计方法,假定反应变量$y>0$。

2　参数λ的估计方法

Box-Cox变换中参数λ的估计有两种方法[2]:其一为最大似然估计,其二为Bayes方法。本研究仅介绍最大似然估计方法,并给出该方法的SPSS程序。由于$Y^{(\lambda)}\sim N(X\beta,\sigma^2 I)$,固定参数λ的取值时,$\beta$、$\sigma^2$的似然函数(即$Y^{(\lambda)}$的密度函数)为:

$$L(\beta,\sigma^2)=\frac{\exp\left(-\frac{1}{2\sigma^2}(Y^{(\lambda)}-X\beta)'(Y^{(\lambda)}-X\beta)\right)}{(2\pi\sigma^2)^{n/2}}J(\lambda,y) \tag{3}$$

其中,$J(\lambda,y)=\prod_{i=1}^{n}\left|\frac{dy_i^{(\lambda)}}{dy_i}\right|=\prod_{i=1}^{n}y_i^{\lambda-1}$。

$L(\beta,\sigma^2)$分别对β、σ^2求导数,并令其等于0,可得β、σ^2的最大似然估计为:

$$\hat{\beta}(\lambda)=(X'X)^{-}XY^{(\lambda)},$$

$$\hat{\sigma}^2(\lambda)=\frac{Y^{(\lambda)\prime}(I-X(X'X)^{-}X')Y^{(\lambda)}}{n}。$$

$\hat{\sigma}^2(\lambda)$可记为MSE(λ),即为模型的误差均方。

将β、σ^2的最大似然估计值$\hat{\beta}(\lambda)$、$\hat{\sigma}^2(\lambda)$代入式(3),得到似然函数的最大值为:

$$L_{\max}(\lambda)=L(\hat{\beta}(\lambda),\hat{\sigma}^2(\lambda))=(2\pi)^{-n/2}(\mathrm{MSE}(\lambda))^{-n/2}J(\lambda,y)$$

对一系列λ的值,似然函数的最大值$L_{\max}(\lambda)$取最大时对应的λ,即为Box-Cox变换中参数λ的估计值。为计算简便,对似然函数两边取对数,略去与λ无关的常数项,得下式:

$$\ln(L_{\max}(\lambda))=-\frac{n}{2}(\mathrm{MSE}(\lambda)+\ln(J(\lambda,y))) \tag{4}$$

综上,Box-Cox变换中估计参数λ的步骤如下:

① 对于给定的λ值,计算$\hat{\beta}(\lambda)$、$\hat{\sigma}^2(\lambda)$;

② 利用式(4)计算$\ln(L_{\max}(\lambda))$;

③ 对一系列λ的值,绘制$\ln(L_{\max}(\lambda))$随λ变化的曲线图。$\ln(L_{\max}(\lambda))$取最大值时的λ值,即为Box-Cox变换中参

收稿日期:2008-12-13　　**通讯作者:**封宗超

作者简介:李运明(1982-),男,江苏徐州人,成都军区总医院质量管理科助理员,第四军医大学卫生统计学专业博士研究生。

△ 项目基金:成都军区总医院院管课题资助

* 成都军区总医院医务部医疗科　　** 成都军区总医院肿瘤中心放疗科

数 λ 的估计值。

3 Box-Cox 变换的 SPSS 程序

假定反应变量 y 均大于 0；在 SPSS 数据文件中，反应变量 y 和所有自变量对应的数据列，按照 y、$x1$、$x2$、$x3$…排列。按照上文参数 λ 的估计方法，编写 Box-Cox 变换的 SPSS 程序如下：

```
SET LENGTH=NONE.
SET MXLOOP = 1000000000.
MATRIX.
GET W/VARIABLES = all/FILE= * /missing=omit.
COMPUTE NC = NCOL(W).
COMPUTE NR = NROW(W).
COMPUTE Y=MAKE(NR,1,0).
COMPUTE XX =MAKE(NR,NC,1).
COMPUTE YLAM=MAKE(NR,1,1).
COMPUTE BOXCOX=MAKE(61,2,0).
COMPUTE YTEMP=0.
LOOP II = 1 TO NR.
COMPUTE Y(II,1)=W(II,1).
LOOP JJ = 1 TO NC-1.
COMPUTE XX(II,JJ+1)=W(II,JJ+1).
END LOOP.
COMPUTE YTEMP=YTEMP+LN(Y(II)).
END LOOP.
LOOP TEMP=1 TO 61.
COMPUTE LAMBDA=-3.1 + TEMP * 0.1.
DO IF LAMBDA=0.
COMPUTE YLAM(:)=LN(Y(:)).
ELSE.
COMPUTE YLAM(:) = (Y(:) & * * LAMBDA - 1)/
LAMBDA.
END IF.
COMPUTE BETA=INV(T(XX) * XX) * T(XX) * YLAM.
COMPUTE MSE = T(YLAM-XX * BETA) * (YLAM-
XX * BETA)/NR.
COMPUTE LOGLIKE=-1 * NR/2 * LN(MSE)+(LAMB-
DA-1) * YTEMP.
COMPUTE BOXCOX(TEMP,1)= LAMBDA.
COMPUTE BOXCOX(TEMP,2)= LOGLIKE.
END LOOP.
SAVE BOXCOX /OUTFILE= *.
END MATRIX.
RENAME VARIABLES COL1= LAMBDA COL2= LOG-
LIKE.
```

4 实例分析

为了验证 Box-Cox 变换 SPSS 程序的正确性，采用 SAS 在线帮助文件中 PROC TRANSREG 的 Box-Cox 变换实例数据进行分析[9]，并比较两种软件变换结果。数据分析目的是估计反应变量 y 随自变量 x 变化的回归直线。

4.1 按照原始数据直接进行回归分析

从不进行变量变换直接进行回归分析的结果（表 1，图 1～2），可以看出回归模型的拟合优度（$R^2=0.888$）尚可，但是剩余残差明显不符合正态分布。为了使剩余残差（误差项）的分布满足正态分布，对反应变量 y 进行变量变换。

表 1 直接进行回归分析结果

方差分析	F 值	773.892	P 值	<0.001
	R^2	0.888	校正 R^2	0.886
参数估计	参数	参数估计值	t 值	P 值
	截距 a	−23.025	−9.119	<0.001
	斜率 b	11.642	27.819	<0.001

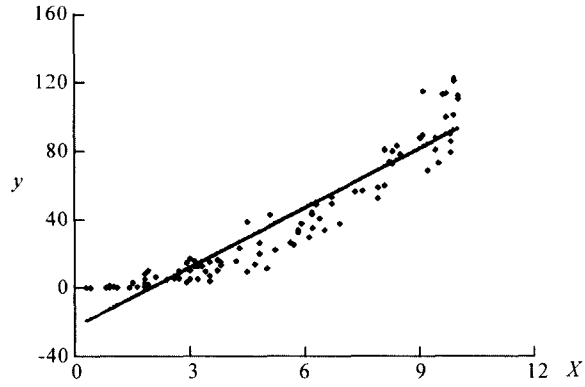

图 1 y、x 的散点图及回归直线

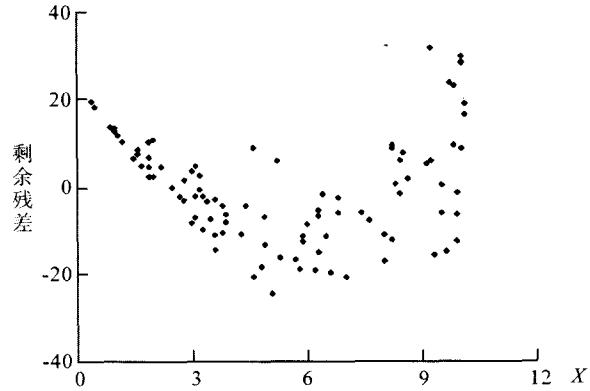

图 2 直接进行回归分析后的残差图

4.2 Box-Cox 变换参数 λ 估计结果

采用本研究给出的 Box-Cox 变换 SPSS 程序，对不同 λ 取值（$-3\leqslant\lambda\leqslant3$），计算似然函数的最大值（$\ln(L_{max}(\lambda))$），并绘制 $\ln(L_{max}(\lambda))$ 随 λ 变化的曲线图，见图 3。结果显示：$\lambda=0.5$ 时，$\ln(L_{max}(\lambda))$ 的值最大。参数 λ 估计结果与采用 SAS / PROC TRANSREG 进行 Box-Cox 变换的结果一致。

4.3 Box-Cox 变换后进行回归分析

从 Box-Cox 变换后进行回归分析的结果（表 2，图 4～5），可以看出回归模型的拟合优度（$R^2=0.954$）升高了，且剩余残

差均匀分布于直线 $y=0$ 的两侧，达到变量变换的目的。

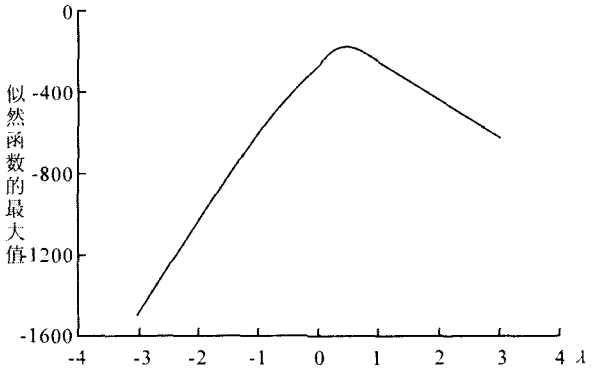

图 3　似然函数的最大值随 λ 变化的曲线图

表 2　Box-Cox 变换后进行回归分析结果

方差分析	F 值	2030.764	P 值	<0.001
	R^2	0.954	校正 R^2	0.953
参数估计	参数	参数估计值	t 值	P 值
	截距 a	−2.536	−9.002	<0.001
	斜率 b	2.104	45.064	<0.001

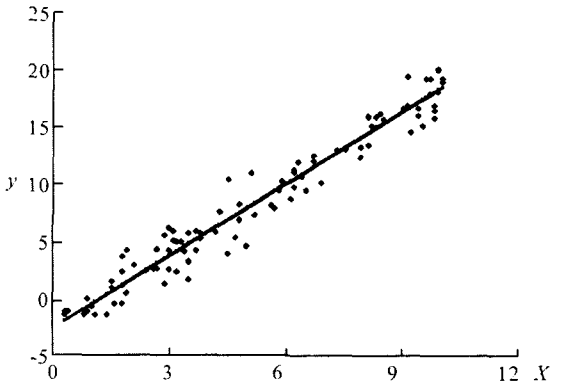

图 4　变换后 y、x 的散点图及回归直线

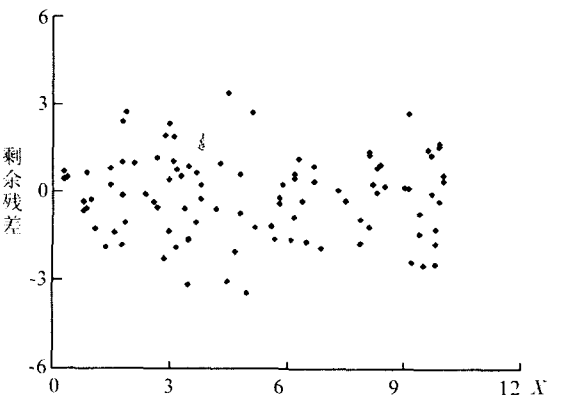

图 5　变换后进行回归分析后的残差图

5　讨论

Box-Cox 变换自从 1964 年由 Box 和 Cox 提出后，不仅在理论上得到了很多发展，而且在统计分析中得到了广泛的实际应用。在理论上，Manly(1971)、John 和 Draper(1980)、Bickel 和 Doksum(1981)、Yeo 和 Johnson(2000)对 Box-Cox 变换进行了改进，扩展了数据变换的应用范围[2]。在实际应用上，反应变量的分布不满足模型要求，常经过 Box-Cox 变换，使得数据满足模型要求[3~7]。

本研究经文献回顾，发现国内统计学者已采用 SAS、STATA 等软件对 Box-Cox 变换的实现问题进行了探讨[3~5]，并给出了相应的程序，但是未见国内采用 SPSS 实现 Box-Cox 变换的研究。而 SPSS 软件是广大生物医学科研工作者进行数据统计分析的工具，本研究给出了进行 Box-Cox 变换的 SPSS 程序，并进行了实例分析，探索出一条生物医学数据分析中利用 SPSS 软件进行 Box-Cox 变换的途径，为类似数据变换问题的解决提供参考。

参　考　文　献

1　孙振球. 医学统计学(供研究生用). 北京：人民卫生出版社，2002，24～70.

2　Pengfei Li. Box-Cox Transformation: An Overview. Department of Statistics, University of Connecticut, 2005.

3　田俊. 比数幂变换的 λ 确定方法及 SAS 程序. 数理医药学杂志，2002，15(6)：481～484.

4　陶庄. Box-Cox 变换及其在 STATA 软件中的实现. 数理医药学杂志，2007，20(3)：380～383.

5　陶庄，金水高. Box-Cox 变换及其在 SAS 软件中的实现. 中国卫生统计，2007，24(5)：541～542.

6　王小平. 医药统计中的方差齐性变换. 数理医药学杂志，2007，20(5)：615～617.

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

8　张文彤. SPSS11 统计分析教程(高级篇). 北京：北京希望电子出版社. 2002，64～90.

9　http://support. sas. com/documentation/onlinedoc/91pdf/index. html