The Korean Communications
in Statistics Vol. 11 No. 1, 2004
pp. 189-196

Outliers in Multivariate Box-Cox Transformed Data

Myung Geun Kim[1)]

Abstract

The sensitivity of the multivariate Box-Cox transformation model to simultaneous perturbations of all the data based on the likelihood displacement is studied in order to detect outliers. An example is given for illustration.

Keywords : Box-Cox transformation, likelihood displacement, outliers.

1. Introduction

Box and Cox (1964) discussed the family of transformations for improving the normality in linear model. Andrews et al. (1971) extended the Box-Cox transformations to multivariate data. It is well recognized that the maximum likelihood estimate of the transformation parameter is very sensitive to outliers. However, few diagnostic methods for multivariate transformations have been developed. Velilla (1995) developed deletion diagnostics and presented a robust estimator of transformation parameter. Riani and Atkinson (2000) suggested a forward searching method which starts with an initial subset of the data containing no masked outliers and monitors the effect of adding observations to the subset, but they did not clearly show that focusing on a few plausible combination of the transformation parameters with one-at-a-time searches will be sufficient generally, as indicated by the discussant.

The local influence method introduced by Cook (1986) is a general method of assessing the influence of minor perturbations of a model and it is used for identifying observations that influence the assumptions underlying the model. It enables us to handle all the data by allowing assessment of simultaneous perturbations affecting all the data, unlike case-deletions. The local influence method relies on the surface of the likelihood displacement for investigating the influence of observations. It is based on the maximum curvature and its corresponding direction vector of a certain curve on the surface formed by the perturbation vector and the likelihood displacement.

In this work we will study the sensitivity of the multivariate Box-Cox transformation model to simultaneous perturbations of all the data based on the likelihood displacement in order to

1) Professor, Department of Applied Statistics, Seowon University, Cheongju, 361-742.
E-mail: mgkim@seowon.ac.kr.

identify outliers. An illustrative example is given.

2. Box-Cox Transformation

The Box-Cox transformation of a p-variate random vector $\boldsymbol{x}_r = (x_{r1}, \ldots, x_{rp})^T$ is defined by

$$x_{ri}(\lambda_i) = \begin{cases} (x_{ri}^{\lambda_i} - 1)/\lambda_i & \text{if } \lambda_i \neq 0 \\ \log(x_{ri}) & \text{if } \lambda_i = 0 \end{cases}$$

for all $r = 1,\ldots,n$ and $i = 1,\ldots,p$. It is assumed that all of p components in each $\boldsymbol{x}_r$ take positive values. The Box-Cox family of transformations is indexed by the vector of transformation parameters, $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)^T$. We assume that the transformed random vectors $\boldsymbol{x}_r(\boldsymbol{\lambda}) = (x_{r1}(\lambda_1), \ldots, x_{rp}(\lambda_p))^T$ are independent and identically distributed as a p-variate normal distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

In what follows, for an m x n matrix $\boldsymbol{A}$, $v(\boldsymbol{A})$ indicates the mn x 1 vector formed by stacking the columns of $\boldsymbol{A}$ from the first column to the last one. For a symmetric n x n matrix $\boldsymbol{A}$, $vh(\boldsymbol{A})$ implies the $n(n+1)/2$ x 1 vector obtained from $v(\boldsymbol{A})$ by deleting all of the elements that are above the diagonal of $\boldsymbol{A}$. We denote by $\boldsymbol{\theta}$ the column vector of parameters formed by stacking the elements of $\boldsymbol{\lambda}$, $\boldsymbol{\mu}$ and $vh(\boldsymbol{\Sigma}^{-1})$ in this order. More details about matrix calculus can be found in Magnus and Neudecker (1988) and Schott (1997).

3. Local Influence Based On Likelihood Displacement

Let $\boldsymbol{w} = (w_1, \ldots, w_n)^T$ be an n x 1 vector of perturbations. We denote the log-likelihoods for the unperturbed and perturbed models by $L(\boldsymbol{\theta})$ and $L(\boldsymbol{\theta}|\boldsymbol{w})$, respectively. The likelihood displacement $LD(\boldsymbol{w})$ is defined by $2[L(\hat{\boldsymbol{\theta}}) - L(\hat{\boldsymbol{\theta}}_w)]$, where $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}_w$ are the maximum likelihood estimators of $\boldsymbol{\theta}$ under the unperturbed and perturbed models, respectively. The surface of interest is formed by the $n+1$ x 1 vector of the values $\boldsymbol{w}$ and $LD(\boldsymbol{w})$ as $\boldsymbol{w}$ varies over a certain space.

We consider case-weight perturbations for which the transformed random vectors $\boldsymbol{x}_r(\boldsymbol{\lambda})$ $(r = 1,\ldots,n)$ are independent and distributed as a p-variate normal distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma}/w_r)$. We write as $\mathbf{1}_m$ the m x 1 vector with all elements equal to 1. When $\boldsymbol{w} = \mathbf{1}_n$, the

perturbed model reduces to the unperturbed model so that $L(\theta) = L(\theta | 1_n)$.

Define the $p(p+5)/2$ x n matrix

$$\Delta = \frac{\partial^2 L(\theta | w)}{\partial \theta \partial w^T}$$

evaluated at $\theta = \hat{\theta}$ and $w = 1_n$, and the $p(p+5)/2$ x $p(p+5)/2$ matrix

$$\dot{L} = \frac{\partial^2 L(\theta)}{\partial \theta \partial \theta^T}$$

evaluated at $\theta = \hat{\theta}$ Let

$$\dot{F} = \Delta^T \dot{L}^{-1} \Delta.$$

Let l_{max} be the eigenvector corresponding to the largest absolute eigenvalue of $-2\dot{F}$ and let $1_{(i)}$ be the i x 1 vector with its i-th element equal to 1 and the others being zero. Then the largest absolute eigenvalue is the maximum curvature of the curve which is the portion of the surface cut out by the plane spanned by the vectors $1_{(n+1)}$ and $(l_{max}^T, 0)^T$ (Cook, 1986, pp.138-9). Observations that correspond to large elements of the first direction vector l_{max} are locally influential and potential outliers.

In the following subsections we will derive $\dot{L}$ and Δ in order to get $\dot{F}$.

3.1 *Derivation of* $\dot{L}$. The log-likelihood function, ignoring unimportant constants, becomes

$$L(\theta) = \sum_{i=1}^{p} (\lambda_i - 1) \sum_{r=1}^{n} \log(x_{ri}) + \frac{n}{2} \log|\Sigma^{-1}| - \frac{1}{2} tr\, E(\lambda, \mu)^T E(\lambda, \mu) \Sigma^{-1},$$

where $E(\lambda, \mu) = X(\lambda) - 1_n \mu^T$ and $X(\lambda)^T = (x_1(\lambda), \ldots, x_n(\lambda))$. We have alternative expressions for the last term of $L(\theta)$ as follows

$$\begin{aligned} tr\{E(\lambda, \mu)^T E(\lambda, \mu) \Sigma^{-1}\} &= v^T\{E(\lambda, \mu)\}(\Sigma^{-1} \otimes I_n) v\{E(\lambda, \mu)\} \\ &= v^T\{E(\lambda, \mu)^T E(\lambda, \mu)\} D_p vh(\Sigma^{-1}) \end{aligned} \quad (1)$$

where D_p is the p^2 x $p(p+1)/2$ duplication matrix such that $v(\Sigma^{-1}) = D_p vh(\Sigma^{-1})$ and the notation $\otimes$ denotes the Kronecker product.

Let

$$\dot{x}_{ri}(\lambda_i) = \frac{dx_{ri}(\lambda_i)}{d\lambda_i} = \begin{cases} -\lambda_i^{-2}(x_{ri}^{\lambda_i} - 1) + \lambda_i^{-1} x_{ri}^{\lambda_i} \log(x_{ri}) & \text{if } \lambda_i \neq 0 \\ (1/2) x_{ri}^{\lambda_i} \{\log(x_{ri})\}^2 & \text{if } \lambda_i = 0 \end{cases}$$

$$\ddot{x}_{ri}(\lambda_i)=\frac{d^2x_{ri}(\lambda_i)}{d\lambda_i^2}=\begin{cases}-2\lambda_i^{-1}(\dot{x}_{ri}(\lambda_i)+\lambda_i^{-1}x_{ri}^{\lambda_i}\{\log(x_{ri})\}^2 & \text{if } \lambda_i\neq 0\\ (1/3)x_{ri}^{\lambda_i}\{\log(x_{ri})\}^3 & \text{if } \lambda_i=0\end{cases}$$

and

$$\boldsymbol{R}(\boldsymbol{\lambda})=[\boldsymbol{d}_{p,1}\otimes\dot{\boldsymbol{x}}_{(1)}(\lambda_1),\ldots,\boldsymbol{d}_{p,p}\otimes\dot{\boldsymbol{x}}_{(p)}(\lambda_p)],$$

where $\boldsymbol{d}_{p,i}$ is the i-th column of the identity matrix $\boldsymbol{I}_p$ of order p and $\dot{\boldsymbol{x}}_{(i)}(\lambda_i)=(\dot{x}_{1i}(\lambda_i),\ldots,\dot{x}_{ni}(\lambda_i))$. Using the first equality of (1), differentiation of $L(\boldsymbol{\theta})$ with respect to $\boldsymbol{\lambda}$ yields

$$\frac{\partial L(\boldsymbol{\theta})}{\partial\boldsymbol{\lambda}}=-\boldsymbol{R}(\boldsymbol{\lambda})(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{I}_n)v\{\boldsymbol{E}(\boldsymbol{\lambda},\boldsymbol{\mu})\} \tag{2}$$

and then the second order differentiation is given by

$$\frac{\partial^2 L(\boldsymbol{\theta})}{\partial\boldsymbol{\lambda}\,\partial\boldsymbol{\lambda}^T}=-diag[\{\boldsymbol{d}_{p,i}^T\otimes\ddot{\boldsymbol{x}}_{(i)}(\lambda_i)\}v\{\boldsymbol{E}(\boldsymbol{\lambda},\boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}\}]-\boldsymbol{R}(\boldsymbol{\lambda})(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{I}_n)\boldsymbol{R}(\boldsymbol{\lambda})^T \tag{3}$$

where the first term in the right-hand side of (3) is the p x p diagonal matrix whose i-th diagonal element is given by that surrounded by brackets and $\ddot{\boldsymbol{x}}_{(i)}(\lambda_i)=(\ddot{x}_{1i}(\lambda_i),\ldots,\ddot{x}_{ni}(\lambda_i))$. Since $\partial v(\mathbf{1}_n\boldsymbol{\mu}^T)/\partial\boldsymbol{\mu}^T=\boldsymbol{I}_p\otimes\mathbf{1}_n$, we get

$$\frac{\partial^2 L(\boldsymbol{\theta})}{\partial\boldsymbol{\lambda}\,\partial\boldsymbol{\mu}^T}=\boldsymbol{R}(\boldsymbol{\lambda})(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{I}_n). \tag{4}$$

Since the right-hand side of (2) can be written as $-\boldsymbol{R}(\boldsymbol{\lambda})\{\boldsymbol{I}_p\otimes\boldsymbol{E}(\boldsymbol{\lambda},\boldsymbol{\mu})\}\boldsymbol{D}_p vh(\boldsymbol{\Sigma}^{-1})$, we have

$$\frac{\partial^2 L(\boldsymbol{\theta})}{\partial\boldsymbol{\lambda}\,\partial vh^T(\boldsymbol{\Sigma}^{-1})}=-\boldsymbol{R}(\boldsymbol{\lambda})\{\boldsymbol{I}_p\otimes\boldsymbol{E}(\boldsymbol{\lambda},\boldsymbol{\mu})\}\boldsymbol{D}_p. \tag{5}$$

Since $\partial L(\boldsymbol{\theta})/\partial\boldsymbol{\mu}=(\boldsymbol{\Sigma}^{-1}\otimes\mathbf{1}_n^T)v\{\boldsymbol{E}(\boldsymbol{\lambda},\boldsymbol{\mu})\}$, we easily obtain

$$\frac{\partial^2 L(\boldsymbol{\theta})}{\partial\boldsymbol{\mu}\,\partial\boldsymbol{\mu}^T}=-n\boldsymbol{\Sigma}^{-1}. \tag{6}$$

Since we have an alternative expression $\partial L(\boldsymbol{\theta})/\partial\boldsymbol{\mu}=[\boldsymbol{I}_p\otimes\{\mathbf{1}_n^T\boldsymbol{E}(\boldsymbol{\lambda},\boldsymbol{\mu})\}]\boldsymbol{D}_p vh(\boldsymbol{\Sigma}^{-1})$, it is easily shown that

$$\frac{\partial^2 L(\boldsymbol{\theta})}{\partial\boldsymbol{\mu}\,\partial vh^T(\boldsymbol{\Sigma}^{-1})}=[\boldsymbol{I}_p\otimes\{\mathbf{1}_n^T\boldsymbol{E}(\boldsymbol{\lambda},\boldsymbol{\mu})\}]\boldsymbol{D}_p. \tag{7}$$

Using Theorem 8.4 of Schott (1997, p.336) and the second equality of (1) provides

$$\frac{\partial^2 L(\theta)}{\partial vh(\Sigma^{-1})\,\partial vh^T(\Sigma^{-1})} = -\frac{n}{2} D_p^T(\Sigma\otimes\Sigma) D_p. \tag{8}$$

Replacing the unknown parameters in (3) to (8) by their maximum likelihood estimates, we obtain $\ddot{L}$ and in this case the derivative in (7) vanishes. Using the formula for finding the inverse of a partitioned matrix, a little computation yields

$$\ddot{L}^{-1} = \begin{bmatrix} I_p \\ Q_2^T \\ Q_3^T \end{bmatrix} Q_1^{-1}[\, I_p \;\; Q_2 \;\; Q_3] + diag(\,0, -\frac{1}{n}\Sigma, -\frac{2}{n}[\, D_p^T(\Sigma\otimes\Sigma) D_p]^{-1})$$

evaluated at $\theta = \hat{\theta}$ where

$$Q_0 = -\Sigma^{-1}\otimes I_n + \frac{1}{n}(\Sigma^{-1}\otimes 1_n)\Sigma(\Sigma^{-1}\otimes 1_n^T) + \frac{2}{n}\{I_p\otimes E(\lambda,\mu)\} D_p[\, D_p^T(\Sigma\otimes\Sigma) D_p]^{-1} D_p^T\{I_p\otimes E(\lambda,\mu)^T\}$$

$$Q_1 = -diag[\{d_{p,i}^T\otimes \ddot{x}_{(i)}(\lambda_i)\} v\{E(\lambda,\mu)\Sigma^{-1}\}] + R(\lambda) Q_0 R(\lambda)^T$$

$$Q_2 = \frac{1}{n} R(\lambda)(\Sigma^{-1}\otimes 1_n)\Sigma$$

$$Q_3 = -\frac{2}{n} R(\lambda)\{I_p\otimes E(\lambda,\mu)\} D_p[\, D_p^T(\Sigma\otimes\Sigma) D_p]^{-1}.$$

3.2 *Derivation of* Δ The perturbed log-likelihood, ignoring unimportant terms for computing Δ is given by

$$\begin{aligned} L(\theta|\, w) &= -\frac{1}{2} tr\{E(\lambda,\mu)^T W E(\lambda,\mu)\Sigma^{-1}\} \\ &= -\frac{1}{2} v(W)^T\{E(\lambda,\mu)\otimes E(\lambda,\mu)\} D_p vh(\Sigma^{-1}) \\ &= -\frac{1}{2} v\{E(\lambda,\mu)\}^T(\Sigma^{-1}\otimes W) v\{E(\lambda,\mu)\} \end{aligned} \tag{9}$$

where W is the n x n diagonal matrix whose r-th diagonal element is w_r. Let $e_r(\lambda,\mu)^T$ be the r-th row of $E(\lambda,\mu)$. Similarly to the derivation of (2), from the third equality of (9) we can derive

$$\frac{\partial L(\theta|\, w)}{\partial \lambda} = -R(\lambda)[\{\Sigma^{-1} E(\lambda,\mu)^T\}\otimes I_n] v(W)$$

and then a little algebra shows that

$$\frac{\partial^2 L(\theta|\, w)}{\partial \lambda\, \partial w^T} = -R(\lambda)[\{\Sigma^{-1} E(\lambda,\mu)^T\}\otimes I_n]\, U \tag{10}$$

where $U=[\, d_{n,1}\otimes d_{n,1}, \ldots, d_{n,n}\otimes d_{n,n}]$.

Since we have $\partial L(\theta|\, w)/\partial \mu=[\{\Sigma^{-1} E(\lambda, \mu)^T\}\otimes 1_n^T]v(W)$ from the third equality of (9), it is easily shown that

$$\frac{\partial^2 L(\theta|\, w)}{\partial \mu\, \partial w^T} = \Sigma^{-1} E(\lambda, \mu)^T. \tag{11}$$

We get $\partial L(\theta|\, w)\partial vh(\Sigma^{-1})=-(1/2) D_p^T\{E(\lambda, \mu)^T\otimes E(\lambda, \mu)^T]v(W)$, from the second equality of (9) and therefore we have

$$\frac{\partial^2 L(\theta|\, w)}{\partial vh(\Sigma^{-1})\, \partial w^T} = -\frac{1}{2} D_p^T G(\lambda, \mu) \tag{12}$$

where $G(\lambda, \mu)$ is the p^2 x n matrix whose r-th column is $e_r(\lambda, \mu)\otimes e_r(\lambda, \mu)$. Evaluating the derivatives in (10) to (12) at $\theta=\hat{\theta}$ and $w=1_n$ yields Δ

4. A Numerical Example

For illustration we consider bivariate radiation measurements recorded through closed (x_1) and open (x_2) doors of 42 microwave ovens (Johnson & Wichern, 1998, p.192 and p.212). The maximum likelihood estimates of the transformation parameters are $\hat{\lambda}_1$ = 0.161 and $\hat{\lambda}_2$ = 0.151.

Figure 1 shows the index plot of the elements of l_{max} described in Section 3 for the multivariate Box-Cox transformation model. It indicates that observations 13, 14, 19 and 35 are outliers for this model.

The likelihood ratio statistic for the hypothesis $\lambda=0$ has a value 2.34 and the associated p-value is 0.311. Thus it is reasonable to take log-transformation of both variables. Figure 2 shows the scatterplot for log-transformed data. We can see that observations 13, 14, 19 and 35 identified by the local influence method are located at the outer side of the data cloud. Observations 13 and 19 can influence the variances of both log-transformed variables. Observations 14 and 35 can influence the probabilistic relation between both log-transformed variables.

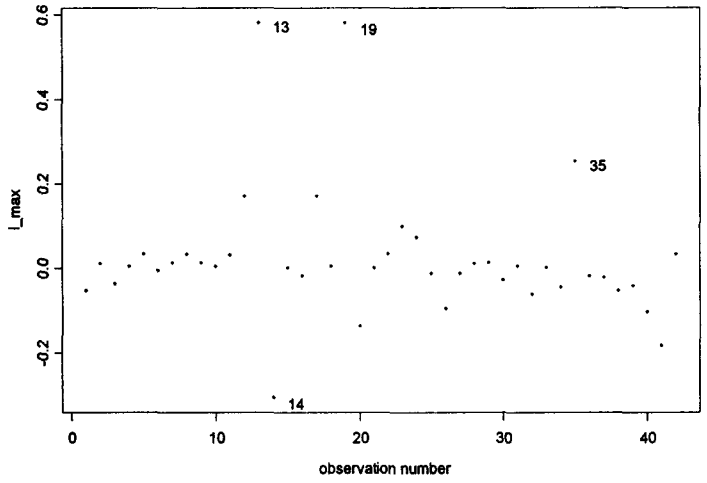

Figure 1. Index plot of the elements of l_{max} for the Box-Cox transformation model

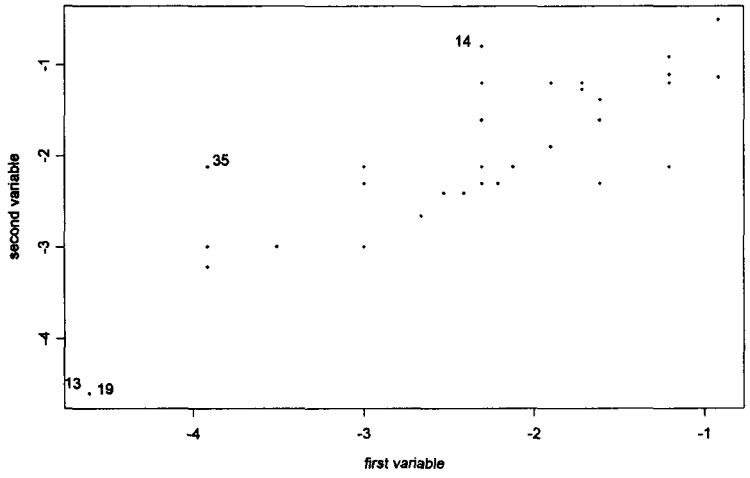

Figure 2. Scatter plot for log-transformed data

References

[1] Andrews, D.F., Gnanadesikan, R. and Warner, J.L. (1971) Transformations of multivariate data, *Biometrics*, **27**, 825-840.

[2] Box, G.E.P. and Cox, D.R. (1964) An analysis of transformations (with discussions),

Journal of the Royal Statistical Society (B), **26**, 211-252.

[3] Cook, R. D. (1986) Assessment of local influence (with discussions), *Journal of the Royal Statistical Society (B)*, **48**, 133-169

[4] Johnson, R.A. and Wichern, D.W. (1998) *Applied Multivariate Statistical Analysis*, 4th ed., Prentice-Hall.

[5] Magnus, J.R. and Neudecker, H. (1988) *Matrix Differential Calculus with Applications in Statistics and Econometrics*, John Wiley, New York.

[6] Riani, M. and Atkinson, A.C. (2000) Robust diagnostic data analysis: transformations in regression (with discussion), *Technometrics*, **42**, 384-397.

[7] Schott, J.R. (1997) *Matrix Analysis for Statistics*, John Wiley, New York.

[8] Velilla, S. (1995) Diagnostics and robust estimation in multivariate data transformations, *Journal of the American Statistical Association*, **90**, 945-951.

[Received January 2004, Accepted April 2004]