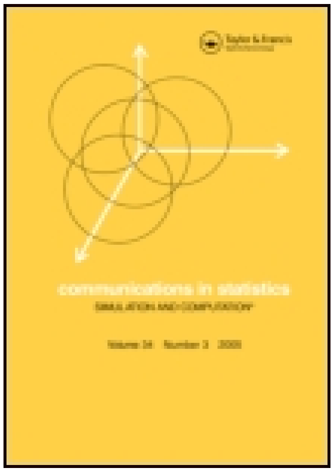

Communications in Statistics - Simulation and Computation

Publication details, including instructions for authors and subscription information: http://www.tandfonline.com/loi/lssp20

THE GENERAL BOX-COX TRANSFORMATIONS IN MULTIPLE LINEAR REGRESSION ANALYSIS

Baibing Li [a] & Bart De Moor [b]

[a] Centre for Process Analytics and Control Technology , University of Newcastle , Newcastle upon Tyne, NE1 7RU, UK

[b] ESAT/SISTA , Department of Electrical Engineering , Katholieke Universiteit Leuven , Kasteelpark Arenberg 10, Leuven-Hevenlee, B-3001, Belgium

Published online: 15 Feb 2007.

To cite this article: Baibing Li & Bart De Moor (2002) THE GENERAL BOX-COX TRANSFORMATIONS IN MULTIPLE LINEAR REGRESSION ANALYSIS, Communications in Statistics - Simulation and Computation, 31:4, 673-687

To link to this article: http://dx.doi.org/10.1081/SAC-120004319

Marcel Dekker, Inc. • 270 Madison Avenue • New York, NY 10016

COMMUN. STATIST.—SIMULA., 31(4), 673–687 (2002)

THE GENERAL BOX–COX TRANSFORMATIONS IN MULTIPLE LINEAR REGRESSION ANALYSIS

Baibing Li[1,*] and Bart De Moor[2]

[1]Centre for Process Analytics and Control Technology,
University of Newcastle,
Newcastle upon Tyne, NE1 7RU, UK
E-mail: baibing.li@ncl.ac.uk
[2]ESAT/SISTA, Department of Electrical Engineering,
Katholieke Universiteit Leuven, Kasteelpark
Arenberg 10, B-3001, Leuven-Hevenlee, Belgium

ABSTRACT

A general Box–Cox transformation method in multiple linear regressions is investigated. An algorithm is proposed to identify optimal general Box–Cox transformations based on kernel density estimation techniques. It is shown that for a multiple linear regression problem, the optimal general Box–Cox transformation can be derived through solving a matrix eigenvector problem, while the regression coefficients are estimated by least squares approach. Examples are given to illustrate the proposed method.

Key Words: Box–Cox transformation; Kernel density estimate; Least squares estimate; Multiple linear regression

[*] Corresponding author.

1. INTRODUCTION

The Box–Cox transformation is one of the most useful methods in regression analysis.[1] For an independent and identically distributed sample $(\tilde{\mathbf{x}}_1^{\mathrm{T}}, \tilde{y}_1)^{\mathrm{T}}, \ldots, (\tilde{\mathbf{x}}_n^{\mathrm{T}}, \tilde{y}_n)^{\mathrm{T}} \in R^{p+1}$ from a distribution, Box and Cox[2] considered the following model

$$\theta(\tilde{\mathbf{Y}}, \lambda) = \alpha\mathbf{1} + \tilde{\mathbf{X}}\boldsymbol{\beta} + \sigma\varepsilon \tag{1.1}$$

where $\tilde{\mathbf{Y}} = [\tilde{y}_1, \ldots, \tilde{y}_n]^{\mathrm{T}}$ is an $n \times 1$ vector of responses, $\tilde{\mathbf{X}} = [\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_n]^{\mathrm{T}}$ is an $n \times p$ design matrix ($p \leq n$), $\boldsymbol{\beta}$ is a $p \times 1$ "slope" parameter vector, α is the intercept, $\mathbf{1}$ is a vector of ones, $\sigma\varepsilon$ is an $n \times 1$ vector of errors which are independent with zero mean and constant variance σ^2, and $\theta(t, \lambda)$ is the Box–Cox transformation function:

$$\theta(t; \lambda) = \begin{cases} (t^{\lambda} - 1)/\lambda, & \text{if } \lambda \neq 0 \\ \log t & \text{if } \lambda = 0 \end{cases} \quad \text{where } t > 0 \tag{1.2}$$

To estimate the parameter λ,[2] used the maximum likelihood method assuming that the error vector, ε, has an exact normal distribution. The optimal solution of the parameter λ in the single-parameter family, (1.2), can be easily found in a plot of the log-likelihood function of the normal distribution versus λ (see, e.g., Ref. [1]).

However, the assumption of normality is often not true in practice and thus it is very important to check normality of residuals.[1] Recently, many methods have been developed which do not assume normality of residuals. Lin and Vonesh[3] constructed a non-linear regression model which is used to estimate the transformation parameter, λ, such that the normal probability plot of the data on the transformed scale is as close to linearity as possible. Halawa[4] investigated the power transformation estimation procedure using an artificial regression model. Rahman[5] proposed to estimate a Box–Cox transformation by maximising the Shapiro–Wilk W statistics. Although implemented in different ways, all of these approaches,[3–5] are based on the same idea of forcing the data to get closer to normality as much as possible.

When not all of the response data, $\tilde{y}_j$ ($j = 1, \ldots, n$), are positive, the transformation family (1.2) is not applicable. Instead a two-parameter transformation family is usually applied:

$$\theta(t; \lambda, \mu) = \begin{cases} [(t + \mu)^{\lambda} - 1]/\lambda, & \text{if } \lambda \neq 0 \\ \log(t + \mu) & \text{if } \lambda = 0 \end{cases} \tag{1.2$'$}$$

such that $\tilde{y}_j + \mu > 0$ for all $j = 1, \ldots, n$. In this case, however, when using the maximum likelihood method, it is no longer possible to spot the optimal solution through a plot for the two-parameter transformation family, $(1.2)'$, as done for the single-parameter case, incurring extra search efforts for the optimal parameters, λ and μ.

The purpose of this paper is to investigate a general Box–Cox transformation approach which can be applied to a wide area, no matter whether the transformed data are normal/positive or not. In addition, the transformation functions will be extended from the single/two parameter family, $(1.2)/(1.2)'$, to any measurable functions. The criterion of selecting an appropriate transformation function is based on minimisation of predictive errors rather than forcing the data to get close to normality as done in the approaches.[3–5] The framework of this approach is based on Ref. [6] in which a general method was developed to estimate optimal transformations for multiple regressions. We will show that, however, for the problem of seeking for a general transformation of the response variable in the linear regression Eq., (1.1), the algorithm proposed in this paper has a particular simplicity. It is non-iterative and easy to implement.

2. MAIN RESULTS

In this section, we first summarise the optimal transformation method for multiple regressions in Ref. [6], and then concentrate on the general Box–Cox transformations.

2.1. A Brief Summary of the Optimal Transformations for Regression

Suppose Y, $X_1, \dots, X_p$ are random variables with Y the response and $X_1, \dots, X_p$ the predictors. Let $\theta(Y)$, $\phi_1(X_1), \dots, \phi_p(X_p)$ be arbitrary measurable zero-mean functions of the corresponding random variables. The objective is to identify these transformations which minimise

$$e^2(\theta, \phi_1, \dots, \phi_p) = E[\theta(Y) - \sum_{j=1}^{p} \phi_j(X_j)]^2 \tag{2.1}$$

subject to $E\theta^2 = 1$, $E\theta = E\phi_1 = \cdots = E\phi_p = 0$
with $\| \cdot \| = [E(\cdot)^2]^{1/2}$. For the simple case where there exists only single predictor, $X = X_1$, Eq. (2.1) reduces to

$$e^2(\theta, \phi) = E[\theta(Y) - \phi(X)]^2 \qquad (2.1)'$$

subject to $E\theta^2 = 1$, $E\theta = E\phi = 0$

and the solution to problem (2.1)$'$ satisfies

$$\theta(Y) = E[\phi(X)|Y]/a \quad \text{and} \quad \phi(X) = E[\theta(Y)|X] \tag{2.2}$$

where $a = \|E[\phi(X)|Y]\|$ is a positive constant. The algorithm proposed by Ref. [6], termed alternating conditional expectation algorithm, is an iterative procedure in which each equation in (2.2) is substituted into another alternately until $e^2(\theta, \phi)$ in (2.1)$'$ fails to decrease. Similarly, for the case of multiple predictors, $\theta(Y)$ that minimise (2.1) satisfies

$$\theta(Y) = E\left[\sum_{i=1}^{p} \phi_i(X_i)|Y\right] \Big/ a \tag*{(2.2)$'$}$$

with $a = \|E[\sum_{i=1}^{p} \phi_i(X_i)|Y]\|$ a positive constant, and the idea of the alternating conditional expectation algorithm is the same as the case of single predictor but more complicated.

2.2. The General Box–Cox Transformations

Without loss of generality, assume that the random variables Y and X_j ($j = 1, \ldots, p$) have zero means. Instead of considering a general problem, (2.1), we restrict our interest to a special case of multiple linear regressions with a transformed response variable. Specifically, an arbitrary zero-mean measurable function of the random variable Y, $\theta(Y)$, is sought for such that θ, together with the regression coefficients, $\beta_1, \ldots, \beta_p$, minimise

$$e^2(\theta, \beta_1, \ldots, \beta_p) = E\left[\theta(Y) - \sum_{j=1}^{p} \beta_j X_j\right]^2 \tag{2.3}$$

subject to $E\theta^2 = 1$ and $E\theta = 0$.

The problem (2.3) is different from (2.1). First of all, the problem (2.3) is a constrained functional optimisation problem in the sense that all of functional ϕ_j ($j = 1, \ldots, p$) in (2.1) are restricted to be linear. Secondly, instead of being a problem of functional optimisations, seeking for the optimal regression coefficients, $\beta_1, \ldots, \beta_p$, is a problem of parameter optimisations, and thus the algorithm for this problem is expected to be much easier than the general case of functional optimisations.

It is clear that the solution of θ to the problem (2.3) is similar to the Eq. (2.2)$'$:

$$\theta(Y) = E\left[\sum_{j=1}^{p} \beta_j X_j | Y\right] \Big/ a \tag{2.4}$$

with $a = \|E[\sum_{j=1}^{p} \beta_j X_j) | Y]\|$ a positive constant, whilst the optimal regression coefficients, $\beta_1, \ldots, \beta_p$, satisfy the first-order conditions $\partial e^2(\theta, \beta_1, \ldots, \beta_p)/\partial\beta_i = 0$, i.e.,

$$\sum_{j=1}^{p} E[\beta_j X_i X_j] = E[X_i \theta(Y)] \quad i = 1, \ldots, p \tag{2.5}$$

Taking mathematical expectation for both sides of Eq. (2.4), and noting that X_j ($j = 1, \ldots, p$) have zero means, we obtain

$$E[\theta(Y)] = E\left[E\left\{\sum_{j=1}^{p} \beta_j X_j | Y\right\}\right] \Big/ a = \sum_{j=1}^{p} \beta_j E[X_j]/a = 0.$$

Therefore, $\theta(Y)$ derived from (2.4) satisfies $E\theta$=0, one of the constraints of problem (2.3). Moreover, it is clear that if $\theta(Y)$ and $\beta_1, \ldots, \beta_p$ is a solution of (2.4) and (2.5), then for any constant $c \neq 0$, $c\theta(Y)$ and $c\beta_1, \ldots, c\beta_p$ is a solution as well. Therefore, to satisfy the another constraint of the problem (2.3), $E\theta^2 = 1$, the constant c is chosen as $[E\theta^2]^{-1/2}$.

The sample version of the problem is to seek for a transformation θ and a vector of the regression coefficients, $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_p]^{\mathrm{T}}$, for the following regression problem

$$\theta(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta} + \sigma\varepsilon \tag*{(1.1)$'$}$$

such that $\boldsymbol{\beta}$ and θ minimise the following sample version of problem (2.3):

$$J(\theta, \boldsymbol{\beta}) = [\theta(\mathbf{Y}) - \mathbf{X}\boldsymbol{\beta}]^{\mathrm{T}}[\theta(\mathbf{Y}) - \mathbf{X}\boldsymbol{\beta}] \tag*{(2.3)$'$}$$

subject to $[\theta(\mathbf{Y})]^{\mathrm{T}}\theta(\mathbf{Y})/(n-1) = 1$ and $\mathbf{1}^{\mathrm{T}}\theta(\mathbf{Y}) = 0$
where $\mathbf{Y} = [y_1, \ldots, y_n]^{\mathrm{T}}$ is an $n \times 1$ vector of responses, $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^{\mathrm{T}}$ is an $n \times p$ design matrix of rank p ($p \leq n$) and $\theta(\mathbf{Y}) = [\theta(y_1), \ldots, \theta(y_n)]^{\mathrm{T}}$. Both $\mathbf{X}$ and $\mathbf{Y}$ are assumed to be mean centred, i.e., $\mathbf{X1} = \mathbf{0}$ and $\mathbf{Y1} = \mathbf{0}$. The sample version of (2.5) is then given by

$$\boldsymbol{\beta}^* = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\theta^*(\mathbf{Y}) \tag*{(2.5)$'$}$$

In order to derive a sample version of (2.4), we consider a sample conditional expectation of $E(X|Y)$ using the kernel density estimation techniques. In this paper, we adopt the well-known Nadaraya–Watson estimator which is a special case of the local polynomial kernel density estimators with zero-order.[7]

Specifically, for a chosen kernel function $k(x) \geq 0$ and a bandwidth $h > 0$, the weighting matrix of the Nadaraya–Watson estimator is constructed as $\mathbf{W} = [w_h(y_i; y_j)]$, where $w_h(t; t_j) = k[(t - t_j)/h] / \sum_{j=1}^{n} k[(t - t_j)/h]$. Note that the weighting matrix $\mathbf{W}$ of a Nadaraya–Watson estimator is a stochastic matrix with all of its elements being non-negative and satisfying $\mathbf{W1} = \mathbf{1}$. The sample conditional expectation of $E(X_j|Y)$ is then given by $\mathbf{Wx}_j$ $(j = 1, \ldots, p)$. Therefore, sample version of Eq. (2.4) is given by

$$\theta^*(\mathbf{Y}) = \mathbf{WX}\boldsymbol{\beta}^*/a \tag{2.4$'$}$$

with $a = \|\mathbf{WX}\boldsymbol{\beta}^*\|$ a positive scalar and $\|\cdot\|$ is the norm of a vector. Inserting $(2.5)'$ into $(2.4)'$ yields

$$\theta^*(\mathbf{Y}) = \mathbf{G}\theta^*(\mathbf{Y})/a \tag{2.6}$$

where $\mathbf{G} = \mathbf{WH}$, $\mathbf{H} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$ and $a = \|\mathbf{WH}\theta^*(\mathbf{Y})\| > 0$. Hence, $\theta^*(\mathbf{Y})$ is an eigenvector of the matrix $\mathbf{G}$ corresponding to a positive eigenvalue. The derived general Box–Cox transformation of the response vector is then given by a scaled $\theta^*(\mathbf{Y})$:

$$\hat{\theta}(\mathbf{Y}) = \theta^*(\mathbf{Y}) / \{[\theta^*(\mathbf{Y})]^{\mathrm{T}}\theta^*(\mathbf{Y})/(n-1)\}^{1/2}$$

and the estimate of the regression coefficients is given by $\hat{\boldsymbol{\beta}}_{\mathrm{LS}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\hat{\theta}(\mathbf{Y})$. Note that the constraint, $\mathbf{1}^{\mathrm{T}}\hat{\theta}(\mathbf{Y}) = \mathbf{0}$, is satisfied since $\mathbf{W1} = \mathbf{1}$ and $\mathbf{X1} = \mathbf{0}$.

Inserting $\hat{\boldsymbol{\beta}}_{\mathrm{LS}}$ into $(2.3)'$ and noting that $[\hat{\theta}(\mathbf{Y})]^{\mathrm{T}}\hat{\theta}(\mathbf{Y})/(n-1) = 1$, problem $(2.3)'$ becomes

$$\tilde{J}(\hat{\theta}) = (n-1) - [\hat{\theta}(\mathbf{Y})]^{\mathrm{T}}\mathbf{H}\hat{\theta}(\mathbf{Y}) \tag{2.7}$$

Hence, the global optimal solution of problem (2.3) can be found through evaluation of $\tilde{J}(\theta)$ at $\theta = \hat{\theta}(\mathbf{Y})$, the scaled eigenvectors of $\mathbf{G}$ corresponding to positive eigenvalues.

One of the problem in solving problem (2.6) is that the dimension of $\mathbf{G}$ depends on the sample size n which may become extremely large in practice. We then convert the eigenvector problem (2.6) of dimension n into another eignevector problem of dimension p which is typically much less than n in many applications. For this end, instead of inserting $(2.5)'$ into $(2.4)'$, we substitute Eq. (2.4)′ into (2.5)′, and let $\mathbf{A} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{WX}$, yielding

$$\mathbf{A}\boldsymbol{\beta}^{*} = a\boldsymbol{\beta}^{*} \tag{2.6$'$}$$

Therefore, $\boldsymbol{\beta}^{*}$ is an eigenvector of $\mathbf{A}$ associated with the positive eigenvalue, a. Note that $\mathbf{A}$ is a $p \times p$ matrix. The matrix $\mathbf{A}$ is constructed by exchanging the left part of the matrix $\mathbf{G} = \mathbf{WX}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$, $\mathbf{WX}$, with the right part, $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$. The matrices $\mathbf{G}$ and $\mathbf{A}$ are therefore have the same non-zero eigenvalues. The least squares estimate of the regression coefficients, $\hat{\boldsymbol{\beta}}_{\mathrm{LS}}$, equals to $\boldsymbol{\beta}^{*}$ after being appropriately scaled. We then have the following algorithm.

```
Given: The design matrix $\tilde{\mathbf{X}}$ and response vector $\tilde{\mathbf{Y}}$ (they do not necessarily have zero means). A kernel function $k(x) \geq 0$ and a bandwidth $h > 0$;
Step 1.  Computing the mean centred matrices $\mathbf{X}$ and $\mathbf{Y} = [y_1, \ldots, y_n]^{\mathrm{T}}$ of $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{Y}}$;
Step 2.  Computing weighting matrix $\mathbf{W} = [(w_h(y_i, y_j)]$;
Step 3.  Computing $\mathbf{H} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$ and $\mathbf{A} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{WX}$;
Step 4.  Computing the normalised eigenvector $\boldsymbol{\pi}_i$ of $\mathbf{A}$ corresponding to positive eigenvalues. Letting $\xi_i = \mathbf{WX}\boldsymbol{\pi}_i$ and standardising $\xi_i$ as $\xi_i = \xi_i/\{\xi_i^{\mathrm{T}}\xi_i/(n-1)\}^{1/2}$;
Step 5.  Letting $\hat{\theta}(\mathbf{Y}) = \arg\max_{\xi_i}\{\xi_i^{\mathrm{T}}\mathbf{H}\xi_i\}$;
Step 6.  Computing the least squares estimate $\hat{\boldsymbol{\beta}}_{\mathrm{LS}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\hat{\theta}(\mathbf{Y})$ and $\hat{\alpha}_{\mathrm{LS}} = \mathbf{1}^{\mathrm{T}}(\hat{\theta}(\mathbf{Y}) - \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_{\mathrm{LS}})/n$;
Step 7.  The transformation function is $\hat{\theta}(t) = \sum_{i=1}^{n}[w_h(t-y_i) \times \{\sum_{j=1}^{p} x_{ij}\hat{\beta}_j\}]$;
End.
```

Noted that choices of the kernel function and bandwidth may influence the resulting $\hat{\theta}$ and $\hat{\boldsymbol{\beta}}_{\mathrm{LS}}$. See Refs. [7,8] for details of selection of kernel function and bandwidth.

3. SOME PROPERTIES

In this section, we investigate properties of the solutions to (2.3) and (2.3)′.

Lemma 1. *The optimal general Box–Cox transformation* $\hat{\theta}(\mathbf{Y})$ *of the response vector* $\mathbf{Y}$, *if exists, is real-valued.*

The proof is immediate by noting that $\hat{\theta}(\mathbf{Y})$ is an eigenvector of the real-valued matrix $\mathbf{G}$ corresponding to a positive eigenvalue, μ, satisfying the linear equation systems of real-valued coefficients, $(\mu\mathbf{I}-\mathbf{G})\hat{\theta}(\mathbf{Y})=\mathbf{0}$.

Let Λ denote the set consisting of all eigenvalues of $\mathbf{G}$. Then we have

Theorem 1. *For the mean centred matrices*, $\mathbf{X}$ *and* $\mathbf{Y}$, *letting* $\mathbf{G}=\mathbf{W}\mathbf{H}$, *where* $\mathbf{H}=\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$, $\mathbf{W}=[w_h(y_i;y_j)]$, *and* $w_h(t;t_j)=k[(t-t_j)/h]/\sum_{j=1}^{n}k[(t-t_j)/h]$, $k(x)\geq 0$ *and* $h>0$, *then we have*

(i) $|\lambda|\leq 1$ *for* $\forall\ \lambda\in\Lambda$. *For the case where* $k(x)=0$ *for some real* x, *if there exists an eigenvalue* $\lambda^*=1\in\Lambda$, *then the associated eigenvector with unit variance is the optimal transformation minimising* (*2.7*).

(ii) *For a positive kernel function* $k(x)>0$ *for* $\forall x$, *we have* $|\lambda|<1$ *for* $\lambda\in\Lambda$.

Proof. (i) Noting that the maximum eigenvalue of a stochastic matrix is 1 (see, e.g., Ref. [9]) and the eigenvalues of $\mathbf{H}$ are either 0 or 1, we have $|\lambda|\leq\|\mathbf{G}\|_2\leq\|\mathbf{H}\|_2\ \|\mathbf{W}\|_2=1$ for $\lambda\in\Lambda$.

Let $\boldsymbol{\eta}_{\max}$ denote an eigenvector of $\mathbf{G}$ with unit length corresponding to the eigenvalue 1. Decompose $\boldsymbol{\eta}_{\max}$ as $\boldsymbol{\eta}_{\max}=c_1\boldsymbol{\gamma}_1+c_2\boldsymbol{\gamma}_2$, where $\boldsymbol{\gamma}_1\in N(\mathbf{H})$ and $\boldsymbol{\gamma}_2\in N^{\perp}(\mathbf{H})$, $\|\boldsymbol{\gamma}_i\|=1$, $0\leq c_i\leq 1$ $(i=1,2)$, $c_1=(1-c_2^2)^{1/2}$, $N(\mathbf{H})$ and $N^{\perp}(\mathbf{H})$ are the null space of $\mathbf{H}$ and its orthogonal complementary space respectively. Then noting that $\mathbf{H}\boldsymbol{\gamma}_1=\mathbf{0}$ and $\mathbf{H}\boldsymbol{\gamma}_2=\boldsymbol{\gamma}_2$, we have

$$\boldsymbol{\eta}_{\max}=\mathbf{G}\boldsymbol{\eta}_{\max}=\mathbf{W}\mathbf{H}(c_1\boldsymbol{\gamma}_1+c_2\boldsymbol{\gamma}_2)=c_2\mathbf{W}\boldsymbol{\gamma}_2$$

Since $1=\|\boldsymbol{\eta}_{\max}\|=c_2\|\mathbf{W}\boldsymbol{\gamma}_2\|\leq c_2\|\mathbf{W}\|_2\|\boldsymbol{\gamma}_2\|=c_2\leq 1$, we obtain $c_2=1$ and thus $c_1=0$. Hence, $\boldsymbol{\eta}_{\max}=\boldsymbol{\gamma}_2\in N^{\perp}(\mathbf{H})$. Finally, $\boldsymbol{\eta}^*=(n-1)^{1/2}\boldsymbol{\eta}_{\max}$, the standardised vector of $\boldsymbol{\eta}_{\max}$, satisfies $\tilde{J}(\boldsymbol{\eta}^*)=(n-1)-\boldsymbol{\eta}^{*\mathrm{T}}\mathbf{H}\boldsymbol{\eta}^*=0$ and thus attains the global minimum.

(ii) If $1=\lambda^*\in\Lambda$, then for any of the associated eigenvectors, $\boldsymbol{\eta}_{\max}$, we have $\boldsymbol{\eta}_{\max}\in N^{\perp}(\mathbf{H})$ and $\mathbf{H}\boldsymbol{\eta}_{\max}=\boldsymbol{\eta}_{\max}$ from (i). Since $\boldsymbol{\eta}_{\max}=\mathbf{G}\boldsymbol{\eta}_{\max}=\mathbf{W}\boldsymbol{\eta}_{\max}$, $\boldsymbol{\eta}_{\max}$ is an eigenvector of $\mathbf{W}$ associated with the eigenvalue 1. On the other hand, when $k(x)>0$ for $\forall\ x$, $\mathbf{W}$ is a positive stochastic matrix. Therefore, the algebraic multiplicity of the eigenvalue 1 is 1 (see Ref. [9]). Hence from $\mathbf{W}\mathbf{1}=\mathbf{1}$, we have $\boldsymbol{\eta}_{\max}=k\ \mathbf{1}\in N(\mathbf{H})$, where k is a scalar. This leads to a contradiction since $\boldsymbol{\eta}_{\max}\in N^{\perp}(\mathbf{H})$. This completes the proof.

In the sequel of this section, we focus on the problem of two random variables, X and Y, each with zero-mean and unit variance, such that θ and β minimise

$$e^2(\theta,\beta)=E[\theta(Y)-\beta X]^2 \tag{3.1}$$

subject to $E\theta^2=1$ and $E\theta=0$

We will investigate in what kind of circumstances an optimal general Box–Cox transformation derived in the section 2, $\theta(Y)$, reduces to the simple linear transformation, i.e., $\theta(Y) = Y$, and under what conditions it gives the same solution as the Box–Cox transformation method.

Lemma 2. *Suppose that both of the random variables Y and X have zero mean and unit variance. If* $X|Y = y \sim N(ry, v^2)$, *then the optimal solution to problem (3.1) is* $\theta(t) = \pm t$.

The proof is immediate from Eq. (2.4). Therefore, Lemma 2 gives a condition under which the optimal transformation is trivial. A related but more interesting situation is that the conditional distribution of the random variable Y given the predictor X is normal.

Theorem 2. *Suppose that both of the random variables Y and X have absolutely continuous distributions with zero mean and unit variance. Then* $Y\big|X = x \sim N(rx, v^2)$ *and the optimal solution to the problem (3.1) is* $\theta(t) = t$ *and* $\beta = r$ *(or* $\theta(t) = -t$ *and* $\beta = -r$*) if and only if* (Y, X) *have a joint normal distribution with the correlation coefficient r.*

Proof. The sufficiency is immediate from Lemma 2. We then consider proof of the necessity. Denote the marginal density functions of X and Y as $f(x)$ and $q(y)$ respectively, and the conditional density function of Y given X as $g(y|x)$. When $\theta(t) = t$ and $\beta = r$ (or $\theta(t) = -t$ and $\beta = -r$), from Eq. (2.4) we have

$$ay = \int_{-\infty}^{+\infty} rx[g(y|x)f(x)/q(y)]\,dx \tag{3.2}$$

where a is positive constant. Equation (3.2) can be rewritten as

$$-\{(1-a)y/v^2\}q(y) = d\left[\int_{-\infty}^{+\infty} q(y|x)p(x)dx\right]\Big/dy$$

or

$$dq(y)/dy = -\{(1-a)y/v^2\}q(y)$$

The above differential equation gives the solution as $q(y) = c_0\exp\{-(1-a)y^2/(2v^2)\}$ for an arbitrary constant c_0. Since $q(y)$ is a density function with unit variance, we obtain $a < 1$, $v^2 = 1 - a$, and $c_0 = (2\pi)^{-1/2}$, yielding a standard normal distribution of Y. Then for the known $q(y)$ and $g(y|x)$, we have following integral equation of the unknown $f(x)$:

$$\int_{-\infty}^{+\infty} f(x)g(y|x)\,dx = q(y)$$

which gives solution $f(x)=|r|/[(2\pi(1-v^2)^{1/2}]\ \exp\{-(rx)^2/[2(1-v^2)]\}$ through the Fourier transformation. Since X has unit variance, we obtain $v^2=1-r^2$, and thus X has a standard normal distribution. Therefore the joint distribution of X and Y, $g(y|x)f(x)$, is a bivariate normal distribution with the correlation coefficient r. This completes the proof.

Immediate from Theorem 2, we have

Corollary. *Suppose $u(t)$ is a strictly monotonic and continuously differentiable function, and both of the random variables $Z=u(Y)$ and X have absolutely continuous distributions with zero mean and unit variance. Then $u(Y)|X=x\sim N(rx,v^2)$, and the optimal solution to problem (3.1) is $\beta=r$ and $\theta(t)=u(t)$ (or $\beta=-r$ and $\theta(t)=-u(t)$) if and only if $(u(Y), X)$ have a joint normal distribution with the correlation coefficient r.*

Hence, if $u(Y)$ and X has a joint normal distribution, and $u(t)$ belongs to the Box–Cox family, (1.2) or (1.2)′, the general Box–Cox transformations will give a transformation function which is identical to that derived by the Box–Cox transformation method.

4. EXAMPLES

In this section, two examples are given to illustrate the developed general Box–Cox transformations. For both examples, the kernel function is chosen as the biweight function, i.e., $k(x)=15(1-x^2)^2/16$ for $x\in[-1,1]$, and $k(x)=0$ otherwise. Sample conditional expectation is taken as the Nadaraya–Watson estimator.

Example 1. A Box–Cox transformation for Mooney viscosity data was investigated by Ref. [1]. The predictor variables are filler level (x_1) and plasticiser level (x_2). A transformation, θ, of the response variable, Mooney viscosity MS$_4$ (V), was explored for the establishment of a linear regression model:

$$\theta(V)=\alpha+\beta_1x_1+\beta_2x_2+\varepsilon$$

Using the Box–Cox transformations, Draper and Smith[1] identified $\log(V)$ as an appropriate transformation. The regression fit of the standardised $\log(V)$, z, and the two predictors, x_1 and x_2, is given by $\hat{z} = -0.7412 + 0.0445\ x_1 - 0.0454\ x_2$.

On the other hand, we derive a general Box–Cox transformation from (2.6). The least square fitting is given by $\hat{\theta} = -0.7016 + 0.0439x_1 - 0.0468x_2$. The result is comparable to that obtained through the application of the Box–Cox transformations, $\log(V)$.

It should be noted that unlike the Box–Cox transformations, the assumption of normality for transformed data is not required for the general Box–Cox transformations. In addition, the transformation function does not necessarily belong to the Box–Cox transformation family. This is illustrated by the next example.

Example 2. In this simulation example, let the predictor X have a distribution function $F(x)$ and the response variable Y be a transformation of Z as $Y = (Z^3 + 3Z + 10)/50$, where Z is a random variable defined by $Z = 2 + X/3 + \varepsilon$. ε is independent of X and has a distribution function $H(x)$. The data vectors, $\mathbf{X}$, $\boldsymbol{\varepsilon}$, $\mathbf{Z}$ and $\mathbf{Y}$, of size 100 are generated as the outcomes of X, ε, Z and Y respectively. Consider the equation

$$\theta(\mathbf{Y}) = \alpha\mathbf{1} + \mathbf{X}\beta + \boldsymbol{\varepsilon}$$

We seek for a transformation of Y, $\theta(Y)$, such that $\mathbf{X}$ and the transformed data $\theta(\mathbf{Y})$ has a strong linear relationship.

Two simulation circumstances are considered. First, $F(x)$ is taken as $N(0,10^2)$ and $H(x)$ $N(0,1)$. The scatter plot of $\mathbf{X}$ and $\mathbf{Y}$ is shown in Figure 1. A general Box–Cox transformation $\hat{\theta}$ is then derived and the scatter plot of $\mathbf{X}$ and $\hat{\theta}(\mathbf{Y})$ is gives by Figure 2. In the second circumstance, $F(x)$ and $H(x)$ are taken as a uniform distribution in $[-5, 5]$ and $[-1, 1]$ respectively. The scatter plot of $\mathbf{X}$ and $\mathbf{Y}$ is shown is Figure 3. A general Box–Cox transformation $\hat{\theta}$ is derived and the scatter plot of $\mathbf{X}$ and $\hat{\theta}(\mathbf{Y})$ is given by Figure 4. It can be seen from these figures that for both circumstances, the general Box–Cox transformations can successfully convert the response vector $\mathbf{Y}$ into a vector $\hat{\theta}(\mathbf{Y})$ which has a strong linear relationship with the predictor vector $\mathbf{X}$.

Table 1 shows the least squares estimates of regression coefficients. The second column gives the least square estimates $\hat{\alpha}$ and $\hat{\beta}$ of α and β in both of the simulation circumstances. For comparison, suppose that the transformation function, $f(t) = (t^3 + 3t + 10)/50$, is exactly known a prior. Then the "true" transformation of the response variable should be taken as $\theta^*(t) = [f(t) - \text{mean}(\mathbf{Z})]/[\text{variance}(\mathbf{Z})]^{1/2}$ (standardisation is taken to satisfy the constraints of (3.1)). The third column of Table 1 gives the least square

Marcel Dekker, Inc. • 270 Madison Avenue • New York, NY 10016

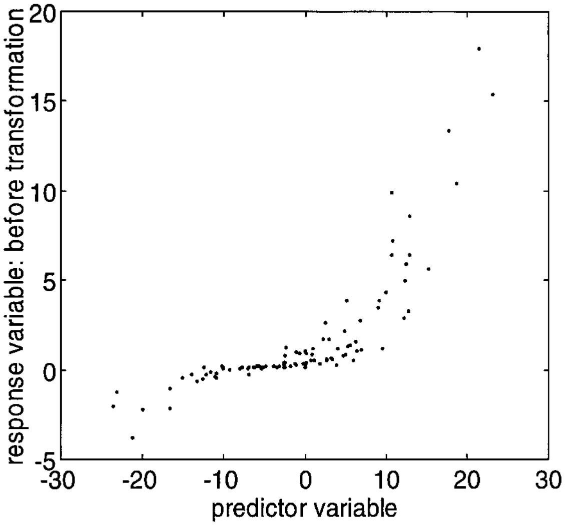

Figure 1. The scatter plot of the predictor variable X and response variable Y for the normal distribution circumstances.

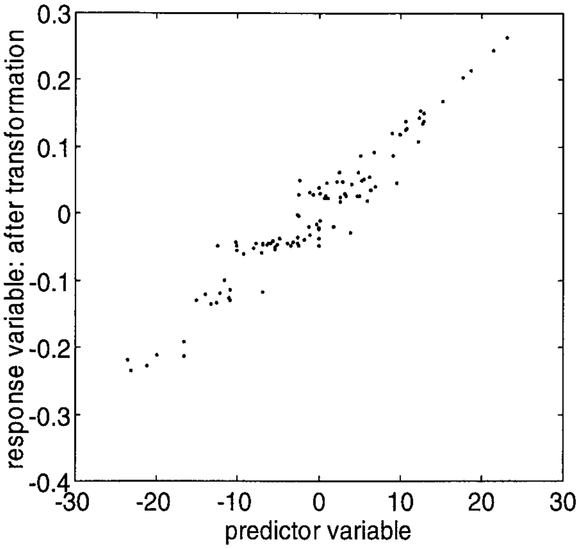

Figure 2. The scatter plot of the predictor variable X and transformed response variable $\hat{\theta}(Y)$ after applying the general Box–Cox transformation for the normal distribution circumstances.

Marcel Dekker, Inc. • 270 Madison Avenue • New York, NY 10016

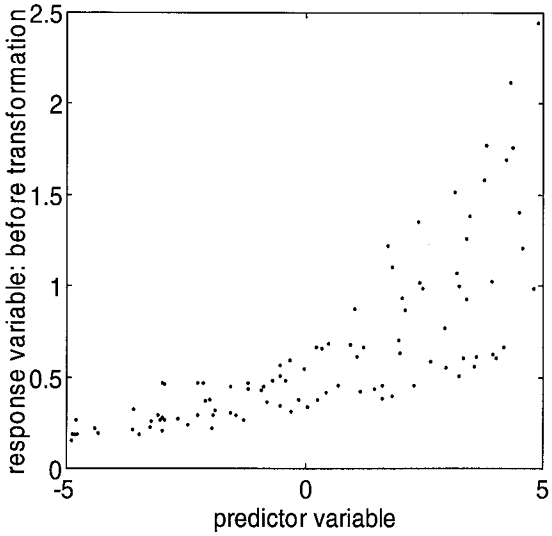

Figure 3. The scatter plot of the predictor variable X and response variable Y for the uniform distribution circumstances.

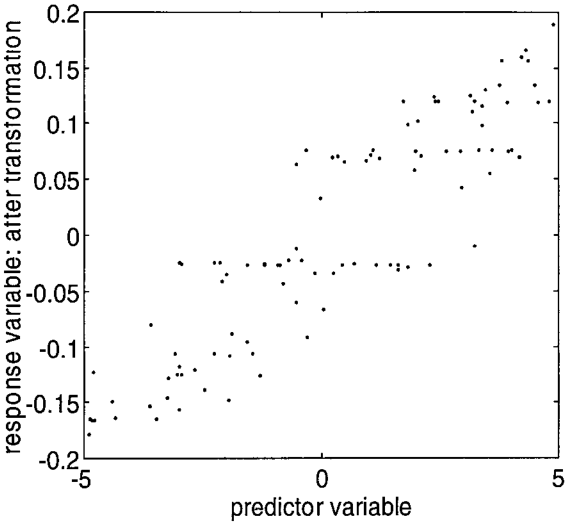

Figure 4. The scatter plot of the predictor variable X and transformed response variable $\hat{\theta}(Y)$ after applying the general Box–Cox transformation for the uniform distribution circumstances.

Table 1. The Least Squares Estimates of Regression Coefficients α and β After Applying the General Box–Cox Transformation and the "True" Transormation Respectively

Distribution	General Box–Cox Transformation $\hat{\theta}$	"True" Transformation θ^*
$F(x) \sim N(0, 10^2), H(x) \sim N(0, 1)$	$\hat{\alpha} = 0.0786$ $\hat{\beta} = 0.1006$	$\alpha^* = 0.0779$ $\beta^* = 0.0997$
$F(x) \sim U[-5, 5], H(x) \sim U[-1, 1]$	$\hat{\alpha} = -0.0908$ $\hat{\beta} = 0.3178$	$\alpha^* = -0.0880$ $\beta^* = 0.3081$

Table 2. Mean Squared Errors of the Least Squares Estimates of Regression Coefficients α and β Over 1000 Simulation Experiments After Applying the General Box–Cox Transformation and the "True" Transformation Respectively

Distribution	General Box–Cox Transformation $\hat{\theta}$	"True" Transformation θ^*
$F(x) \sim N(0, 10^2)$, $H(x) \sim N(0, 1)$	$\mathrm{MSE}(\hat{\alpha}) = 9.5426 \times 10^{-3}$ $\mathrm{MSE}(\hat{\beta}) = 4.5797 \times 10^{-5}$	$\mathrm{MSE}(\alpha^*) = 9.3892 \times 10^{-3}$ $\mathrm{MSE}(\beta^*) = 4.0679 \times 10^{-5}$
$F(x) \sim U[-5, 5]$, $H(x) \sim U[-1, 1]$	$\mathrm{MSE}(\hat{\alpha}) = 8.2997 \times 10^{-3}$ $\mathrm{MSE}(\hat{\beta}) = 1.5507 \times 10^{-4}$	$\mathrm{MSE}(\alpha^*) = 7.9367 \times 10^{-3}$ $\mathrm{MSE}(\beta^*) = 1.2505 \times 10^{-4}$

estimates α^* and β^* of α and β after applying the "true" transformation θ^*. It can be seen that for both circumstances, the results obtained by applying general Box–Cox transformations $\hat{\theta}$ and "true" transformations θ^* are very close to each other.

Finally, one thousand experiments, each with sample size $n = 100$, for both circumstances are conducted. Table 2 gives mean squared errors (MSE) of the least squares estimates of regression coefficients α and β after applying the general Box–Cox transformation and the "true" transformation under the normal and uniform distribution circumstances. It can be seen that, in comparison with the "true" transformation, the general Box–Cox transformation method performs very well, no matter the underlying distribution is normal or not.

REFERENCES

1. Draper, N.R.; Smith, H. *Applied Regression Analysis*; Wiley: New York, 1998.

2. Box, G.E.P.; Cox, D.R. An Analysis of Transformation. Journal of the Royal Statistical Society **1964**, *Series B 26*, 211–252.
3. Lin, L.I.; Vonesh, E.F. An Empirical Nonlinear Data-Fitting Approach for Transformating Data to Normality. American Statistician **1989**, *43*, 237–243.
4. Halawa, A.M. Estimating the Box–Cox Transformation via Artificial Regression Model. Commun. Statist.—Simula. **1996**, *25*, 331–350.
5. Rahman, M. Estimating the Box–Cox Transformation via Shapiro–Wilk W Statistics. Commun. Statist.—Simula. **1999**, *28*, 223–241.
6. Breiman, L.; Friedman, J.H. Estimating Optimal Transformations for Multiple Regression and Correlation (with Discussion). Journal of the American Statistical Association **1985**, *80*, 580–619.
7. Wand, M.P.; Jones, M.C. *Kernel Smoothing*; Chapman & Hall: London, 1995.
8. Bowman, A.W.; Azzalini, A. *Applied Smoothing Techniques for Data Analysis*; Clarendon Press: Oxford, 1997.
9. Horn, R.A.; Johnson, C.R. *Matrix Analysis*; Cambridge University Press: New York, 1985.

MARCEL DEKKER, INC. • 270 MADISON AVENUE • NEW YORK, NY 10016