CHAPTER II

ANALYSIS OF TRANSFORMATIONS IN REGRESSION MODELS STATISTICS
2.1 INTRODUCTION

One of the reasons for the popularity of the linear regression model is that even when linearity is an unreasonable assumption for the original data, it often is reasonable for data which have been appropriately transformed. Many different transformations can be used to obtain linearity and approximate normality of the residuals. However, the only family of transformations which is widely used in linear models is the Box – Cox family. This family has many desirable properties, among them the fact that it includes both the logarithmic transformation and no transformation at all as special cases.

Box and Cox (1964) suggested estimation of data transformations that yield a model satisfying the standard assumptions of linearity, homoscedasticity and normality.

Box and Cox transformation is not always appropriate. It cannot be applied to variables which can take on zero or negative values.

Box and Cox (1964) have introduced the usual techniques for the analysis of linear models as exemplified by the analysis of variance and by multiple regression analysis are usually justified by assuming

(i) Simplicity of linear structure of dependent vector

(ii) Constancy of error variance
(iii) Normality of distributions
(iv) Independence of observations

In the analysis of variance, a very important example of (i) is the assumption of additivity.

To achieve additivity in the analysis of variance, selection might be based on

(a) Minimization of the F-value for the degree of freedom for non-additivity
(b) Minimization of the F-ratio for interaction versus error; or
(c) Maximization of the F-ratio for treatments versus error

In regression problems, where both dependent and independent variables can be transformed, there are more possibilities to be considered.

Transformation of the independent variables Box and Tidwell, (1962) can be applied without affecting the constancy of variance and normality of error distributions. An important application is to convert a monotonic non-linear regression relation into a linear one. Obviously it is useless to try to linearize a relation which is not monotonic, but a transformation is sometimes useful in such cases;
2.2 BOX-COX TRANSFORMATION OF THE DEPENDENT VARIABLE

Consider a parametric family of transformations from $Y$ to $Y^{(\lambda)}$, the parameter $\lambda$, possibly a vector, defining a particular transformation.

Two important transformations considered here are

\[
Y^{(\lambda)} = \begin{cases} 
Y^\lambda - 1, & (\lambda \neq 0) \\
\log Y, & (\lambda = 0)
\end{cases} \quad \text{... (2.2.1)}
\]

and

\[
Y^{(\lambda)} = \begin{cases} 
(Y + \lambda_2)^{\lambda_1} - 1, & (\lambda_1 \neq 0) \\
\log(Y + \lambda_2), & (\lambda_1 = 0)
\end{cases} \quad \text{... (2.2.2)}
\]

The transformations (2.2.1) hold for $Y > 0$ and (2.2.2) for $Y > -\lambda_2$ since an analysis of variance is unchanged by a linear transformation (2.2.1) is equivalent to

\[
Y^{(\lambda)} = \begin{cases} 
Y^\lambda (\lambda \neq 0) \\
\log Y (\lambda = 0)
\end{cases} \quad \text{... (2.2.3)}
\]

The form (2.2.1) is slightly preferable for theoretical analysis because it is continuous at $\lambda = 0$. In general, it is assumed that for each $\lambda$, $Y^{(\lambda)}$...
is a monotonic function of $Y$ over the admissible range. Suppose that one observe an $n \times 1$ vector of observations $Y = y_1, y_2, y_3, \ldots, y_n$, and that the appropriate linear model for the problem is specified by

$$E[Y^{(d)}] = a \theta \quad \ldots \quad (2.2.4)$$

Where $Y^{(d)}$ is the column vector of transformed observations, 'a' is a known matrix and $\theta$ a vector of unknown parameters associated with the transformed observations.

One may now assume that for some unknown $\lambda$, the transformed observations $Y_i^{(d)}$ (i = 1, 2, ..., n) satisfy the full normal theory assumptions, i.e. They are independently normally distributed with constant variance $\sigma^2$, and with expectations (2.2.4). The probability density function for the untransformed observations, and hence the likelihood function in relation to these original observations, is obtained by multiplying the normal density function by the Jacobian of the transformation.

The likelihood function in relation to the original observations $Y$ is thus

$$L(\lambda) = \frac{1}{(2\pi)^{\frac{n}{2}} \sigma^n} e^{-\frac{1}{2\sigma^2} \left[ \frac{(Y^{(d)} - a\theta)^T (Y^{(d)} - a\theta)}{2\sigma^2} \right]} J(\lambda; y), \quad \ldots \quad (2.2.5)$$
Where

\[ J(\lambda; y) = \prod_{i=1}^{n} \frac{dY_i^{(\lambda)}}{dY_i} \]

One can find the maximum likelihood estimates in two steps. First, for given \( \lambda \), (2.2.5) is expect for a constant factor, the likelihood for a standard least squares problem.

Hence, the maximum likelihood estimates of the \( \theta' \)'s are the least squares estimates for the dependent variable \( Y^{(\lambda)} \) and the estimate of \( \sigma^2 \), denoted for fixed \( \lambda \) by \( \hat{\sigma}^2(\lambda) \), is

\[ \hat{\sigma}^2(\lambda) = \frac{Y^{(\lambda)'}}{n} a_r Y^{(\lambda)} = \frac{S(\lambda)}{n} \]

\[ \cdots (2.2.6) \]

Where, when 'a' is of full rank,

\[ a_r = I - a (a'a)^{-1} a' \]

\[ \cdots (2.2.7) \]

and \( S(\lambda) \) is the residual sum of squares in the analysis of variance of \( Y^{(\lambda)} \).

Thus, for fixed \( \lambda \), the maximized log likelihood function is, except for a constant,
\[ L_{\text{max}}(\lambda) = -\frac{1}{2} \log \hat{\sigma}^2(\lambda) + \log J(\lambda, y) \quad \ldots \ (2.2.8) \]

In the important special case (2.2.1) of the simple power transformation, the second term in (2.2.8) is

\[ \log J(\lambda, y) = (\lambda - 1) \sum \log y_i \quad \ldots \ (2.2.9) \]

In (2.2.2), when an unknown origin \( \lambda_0 \) is included, the term becomes

\[ \log J(\lambda, \lambda_0, y) = (\lambda - 1) \sum \log(y_i + \lambda_0) \quad \ldots \ (2.2.10) \]

It will now be informative to plot the maximized log likelihood \( L_{\text{max}}(\lambda) \) against for a trial series of values. From this plot, the maximizing value \( \hat{\lambda} \) may be read off and one can obtain an appropriate 100(1-\( \alpha \))% confidence region from

\[ L_{\text{max}}(\hat{\lambda}) - L_{\text{max}}(\lambda) < \frac{1}{2} \chi^2_{v_\lambda}(\alpha) \quad \ldots \ (2.2.11) \]

Where

\( v_\lambda \) is the number of independent components in \( \lambda \).

In the special case of the one parameter power transformation

\[ y^{(\lambda)} = \frac{y^\lambda - 1}{\lambda}, \]
\[
\frac{d}{dy} L_{\text{max}}(\lambda) = -n \frac{Y^{(1)} a_r u^{(1)}}{Y^{(1)} a_r Y^{(2)}} + \frac{n}{\lambda} + \sum \log y \quad \cdots (2.2.12)
\]

Where

\( u^{(1)} \) is the vector of components \( \{a_t y_t \log y_t \} \)

The numerator in (2.2.12) is the residual sum of products in the analysis of covariance of \( Y^{(2)} \) and \( u^{(1)} \).

The above results can be expressed very simply if we work with the normalized transformation

\[
Z^{(2)} = \frac{Y^{(2)}}{J^{n}}
\]

Where

\[
J = J(\lambda; Y).
\]

Then

\[
L_{\text{max}}(\lambda) = -\frac{1}{2} \log \hat{\sigma}^2(\lambda; Z) \quad \cdots (2.2.13)
\]

Where

\[
\hat{\sigma}^2(\lambda; Z) = \frac{Z^{(1)} a_r Z^{(2)}}{n} = \frac{S(\lambda; Z)}{n},
\]

Where

\( S(\lambda; Z) \) is the residual sum of squares of \( Z^{(a)} \).
The maximized likelihood function is thus proportional to \( \{S(\lambda; Z)\}^n \) and the maximum likelihood estimate is obtained by minimizing \( S(\lambda; Z) \) with respect to \( \lambda \).

For the simple power transformation, one can have

\[
Z^{(2)} = \frac{Y^\lambda - 1}{\lambda Y^\lambda - 1} \quad \ldots \quad (2.2.14)
\]

Where

\( Y \) is the geometric mean of the observations.

For the power transformation with shifted location, it is given by

\[
Z^{(2)} = \frac{(Y + \lambda_2)^\lambda - 1}{\lambda_1 \{gm(Y + \lambda_2)^\lambda\}^{\lambda-1}} \quad \ldots \quad (2.2.15)
\]

Where

\( gm(Y + \lambda_2) \) is the sample geometric mean of the \((Y + \lambda_2)\)'s.

### 2.3 AN EXTENSION OF BOX-COX TRANSFORMATION

Consider a set \( n \) of observations \( y_1, y_2, \ldots, y_n \) and there exists a monotonic transformation \( Z = f(y) \), such that \( Z \) is normally distributed with constant variance \( \sigma^2 \) and mean, \( \bar{Z} \).

Box and Cox considered transformations of the form

\[
f(y) = f(y, \phi),
\]
Where the parameter, $\phi$, possibly a vector, defines a transformation. They have shown that, for fixed $\phi$, the maximized log likelihood of the original observations is, expect for a constant,

$$L_{\text{max}}(\phi) = -\frac{1}{2} n \log \hat{\sigma}^2 + \log J(\phi; y) \quad \ldots (2.3.1)$$

Where

$\hat{\sigma}^2$ is the maximum likelihood estimate of $\sigma^2$, and

$$J(\phi; y) = \prod_{i=1}^{n} \frac{dZ_i}{dy_i} \quad \ldots (2.3.2)$$

The Jacobian of the transformation.

To use this, $Z_i$ must be specified as a function of a set of parameters, and Box and Cox assumed that $Z_i$ is a linear function of a vector of parameters, $\beta$.

However, in several important practical situations it is necessary to relax this assumption and permit $Z_i$ to be a non-linear function of the parameters.

In particular,

Let $Y = f'(Z)$, The transformation of the expectations of the $Z$'s to the original scale, and suppose there exists a transformation
\[ U = g(Y, \theta), \]

Such that \( U \) is a linear function of a vector of parameters, \( \beta \), and the parameter, \( \theta \), possibly a vector, defines the transformation necessary to give linearity. This model is an extension of the generalized linear model of Nelder and Wedderburn (1972) in which the data, \( y \), are replaced by \( Z = f(Y, \phi) \) and the linking function, \( f \), involved a function, \( g' \), which depends on parameters, \( \theta \).

If \( g(Y, \theta) = f(Y, \phi) \), for all \( Y \), this is the situation Considered by Box and Cox.

For fixed \( \phi, \theta \), if the log likelihood function is maximized with respect to the parameters, \( \beta \),

\[ L_{\max}(\phi, \theta) = -\frac{1}{2} n \log \hat{\sigma}^2 + \log J(\phi; Y), \quad \cdots \quad (2.3.3) \]

apart from a constant, and maximum likelihood estimates of \( \phi, \theta \) may be found by maximizing this with respect to \( \phi \) and \( \theta \).

Box and Cox suggested replacing the \( Z \)'s by \( \frac{Z}{\{ J(\phi; Y) \}^{1/2}} \) since the maximum likelihood function then becomes

\[ L_{\max}(\phi, \theta) = -\frac{1}{2} n \log \hat{\sigma}^2 \quad \cdots \quad (2.3.4) \]

This form is more convenient analytically, and will often reduce rounding errors when the method is used in practice. In the following it is assumed that the \( Z \)'s have been normalized in this way.
Let \( \hat{Z} \) denotes a fitted value of \( Z \) calculated from maximum likelihood estimates of the parameter, \( \theta \),

\[
\hat{\sigma}^2 = \frac{\sum (Z - \hat{Z})^2}{n} \quad \text{(2.3.5)}
\]

Define \( S(\phi, \theta) \) as the value of \( \sum (Z - \hat{Z})^2 \) for values of \( \phi \) and \( \theta \).

Consider

\[
f(Y, \phi) = \begin{cases} \frac{Y^\phi - 1}{\phi}, \phi \neq 0 \\ \log Y, \phi = 0 \end{cases} \quad \text{(2.3.6)}
\]

\[
g(Y, \theta) = \begin{cases} \frac{Y^\theta - 1}{\theta}, \theta \neq 0 \\ \log Y, \theta = 0 \end{cases} \quad \text{(2.3.7)}
\]

If \( \phi = \theta \), this is the first example considered by Box and Cox.

Suppose that for a particular set of data past experience indicates that often \( \phi = 0 \).

To test this hypothesis consider

\[
L_R = 2 \left[ L_{\text{max}}(\hat{\phi}, \hat{\theta}) - L_{\text{max}}(0, \hat{\theta}_0) \right] \quad \text{(2.3.8)}
\]

Where \( \hat{\phi}, \hat{\theta} \) are the maximum likelihood estimates of \( \phi \) and \( \theta \),

and \( \hat{\theta}_0 \) is the maximum likelihood estimates of \( \theta \), given \( \phi = 0 \).
\( L_R \) is asymptotically distributed as \( \chi^2 \) with one degree of freedom.

\[
L_R = -n \left\{ \log \left( \frac{S(\hat{\phi}, \hat{\theta})}{n} \right) - \log \left( \frac{S(0, \hat{\theta}_0)}{n} \right) \right\} \quad \ldots (2.3.9)
\]

\[
L_R = n \log \left\{ 1 + \frac{S(0, \hat{\theta}_0) - S(\hat{\phi}, \hat{\theta})}{S(\hat{\phi}, \hat{\theta})} \right\} \quad \ldots (2.3.10)
\]

Which is a function of \( \frac{S(0, \hat{\theta}_0) - S(\hat{\phi}, \hat{\theta})}{S(\hat{\phi}, \hat{\theta})} \) so the likelihood-ratio test is analogous to the standard F-test of the analysis of the variance.

The examination of a sequence of composite hypothesis using log likelihoods has been discussed by Good (1967) and Nelder and Wedderburn (1972). For example, if the hypothesis are \((\phi = 0, \theta = -1), (\phi = 0, \theta \text{ unknown})\) and \((\phi, \theta \text{ both unknown})\),

\[
2L_{\text{max}}(\phi, \theta) = 2L_{\text{max}}(\hat{\phi}, \hat{\theta}) - L_{\text{max}}(0, \hat{\theta}_0) + 2L_{\text{max}}(0, \hat{\theta}_0) - L_{\text{max}}(0, -1) + 2L_{\text{max}}(0, -1) \quad \ldots (2.3.11)
\]

The first two terms are asymptotically independent and distributed as \( \chi^2 \) on one degree of freedom.
2.4 ESTIMATION OF PARAMETERS OF REGRESSION MODELS WITH BOX-COX TRANSFORMATION

The power transformation introduced by Box and Cox (1964), and given by

\[ y^{(\lambda)} = \begin{cases} 
\frac{Y^\lambda - 1}{\lambda}, & \lambda \neq 0 \\
\ln Y, & \lambda = 0
\end{cases} \quad \ldots (2.4.1) \]

has been extensively used in recent years. Transformed variables can be included in a "linear" function so that generalized models of the form

\[ y^{(\lambda)} = \beta_1 + \beta_2 x_2^{(\lambda_2)} + \ldots + \beta_k x_k^{(\lambda_k)} + \epsilon \quad \ldots (2.4.2) \]

can be specified and estimated.

The likelihood function of the sample and an estimator of the variance-covariance matrix of the estimators are derived as follows; initially, it is assumed that (2.4.2) is the model to be estimated. Under the assumption that there exists some \( \lambda \) for which \( \epsilon \) in (2.4.2) approximately normally distributed with mean zero and variance \( \sigma^2 \), the density function of the \( i^{th} \) observation of \( \epsilon \) is given by

\[ f(\epsilon_i) = \left(2\pi\sigma^2\right)^{-\frac{1}{2}} e^{-\frac{\epsilon_i^2}{2\sigma^2}} \quad \ldots (2.4.3) \]
The likelihood function is then

\[ L'(\beta, \lambda, \sigma^2; X, Y) = \prod_{i=1}^{n} f(\epsilon_i) Y_i^{4-1} \]  

... (2.4.4)

Where

the last term is the Jacobian of the transformation From $\epsilon$ to $Y$. The logarithm of (2.4.4) is given by

\[ L = \ln L' = K_1 - \frac{3}{2} \ln \sigma^2 - (2\sigma^4)^{\frac{3}{2}} \sum_{i=1}^{n} (Y_i^{4}) - \beta_i - \beta_j X_i^{4-1} - \ldots - \beta_i X_i^{4-1} \sigma^2 + (\lambda_i - 1) \sum_{i=1}^{n} \ln Y \]  

... (2.4.5)

Where

$K_1$ is a constant. Since (2.4.5) is a monotonic transformation of (2.4.4), both functions will be maximized for the same parameter values. Maximization of (2.4.5) under the assumption of approximate normality of the $\epsilon_i$ obtains estimators $\beta_i, \lambda_i$ and $\sigma^2$ which are Best Asymptotically Normal (BAN) under general regularity conditions. Further more, the asymptotic covariance matrix of the parameter estimates is given by

\[ -E \left( \frac{\partial^2 L}{\partial \theta \partial \theta'} \right)^{-1} \]  

... (2.4.6)

Where

\[ \theta = \{\beta_1, \beta_2, \lambda_1, \lambda_2, \sigma^2\} \]

This matrix is the Cramer-Rao lower bound Irrespective of the approach taken to estimation of the parameters of (2.4.2) a consistent
estimate from (2.4.6) may be obtained. The expected value of the second derivative matrix cannot be evaluated for functions which contain the Box-Cox transformation, the expectations in (2.4.6) are too complex to evaluate. Instead, the observed matrix

$$\left( \frac{\partial^2 L}{\partial \theta \partial \theta^\prime} \right)^{-1}$$

is used. ... (2.4.7)

Goldfeld and Quandt (1972) indicate that (2.4.6) is consistently estimated by (2.4.7) if the estimators are sufficient.

Simplex Models with Box-Cox Transformation

Several different versions of (2.4.4) are possible. Each representation places certain restrictions on some of the $\lambda$ making the estimation problem some what simpler.

$$Y^{(\lambda)} = \beta_1 + \beta_2 X_2 + ... + \beta_k X_k + \epsilon$$  

... (2.4.8)

$$Y^{(\lambda')} = \beta_1 + \beta_2 X_2^{(\lambda')} + ... + \beta_k X_k^{(\lambda')} + \epsilon$$  

... (2.4.9)

$$Y^{(\lambda_1)} = \beta_1 + \beta_2 X_2^{(\lambda_2)} + ... + \beta_k X_k^{(\lambda_k)} + \epsilon$$  

... (2.4.10)

$$Y^{(\lambda_1)} = \beta_1 + \beta_2 X_2^{(\lambda_2)} + ... + \beta_k X_k^{(\lambda_k)} + \epsilon$$  

... (2.4.11)
are all possible combinations. In (2.4.8), only the y values are power transformed. In (2.4.9), all X’s and the y value are transformed by the same value of λ.

In (2.4.10), all X’s are transformed in the same way, but the Y value is subject to a different transformation. Lastly (2.4.11), which is the most general case, allows for different Box-Cox transformations on all the variables in the model. Equation (2.4.11) is identical to (2.4.2).

Using the more compact notation of matrix algebra, equation (2.4.8) is

\[ Y^{(\lambda)} = X\beta + \varepsilon \]  \hspace{1cm} ... (2.4.12)

Where

\[ Y^{(\lambda)} \] is a nx1 vector of transformed observations on the dependent variable;

\[ X \] is a nxk matrix of observations on the independent variables, where the first column is a vector of ones.

The likelihood function to be maximized for model (2.4.8) is given by

\[ L(\beta, \lambda, \sigma^2; X, Y) = K_1 - \frac{n}{2} \ln \sigma^2 + (\lambda - 1)q^1 \ln Y - (Y^{(\lambda)} - X\beta)'(Y^{(\lambda)} - X\beta)2\sigma^2 \]  \hspace{1cm} ... (2.4.13)
Where

\[ q^l = [111111...1] \] The third term on the right hand side of (2.4.13) is the log of the Jacobian.

Different Approaches to the Estimation of Regression Models

Four approaches to the estimation of (2.4.8) presented below.

Each is equivalent to maximum likelihood (ML) estimation.

1. Maximizing the full log-likelihood function (13).
2. Maximizing the concentrated log-likelihood function.
3. Maximizing a function of the transformed sum of squares function. This method is identical to nonlinear least squares.
4. Minimizing the transformed sum of squares function by repeated use of ordinary squares [Iterated OLS].

Maximizing the Full Log-Likelihood Function

Equation (2.4.13) is to be maximized with respect to \( \beta, \lambda \) and \( \sigma^2 \).

Let \[ \epsilon_\lambda = \frac{\partial \epsilon}{\partial \lambda} \text{ and } \epsilon_\mu = \frac{\partial^2 \epsilon}{\partial \lambda^2} \].

The first order conditions for a maximum require that

\[ \frac{\partial L}{\partial \beta} = \sigma^{-2} X^l \epsilon = 0 \quad \ldots (2.4.14) \]
\[
\frac{\partial L}{\partial \lambda} = -\sigma^{-2} \epsilon' \epsilon + q' \ln Y = 0 \quad \ldots (2.4.15)
\]

\[
\frac{\partial L}{\partial \sigma^2} = -\left(\frac{T}{2}\right) \sigma^{-2} + \frac{\epsilon' \epsilon}{2\sigma^4} = 0 \quad \ldots (2.4.16)
\]

The second order conditions for a local maximum require that the matrix of second derivatives be negative definite.

Let
\[
\theta' = [\beta' \lambda \sigma^2].
\]

Then the second derivative matrix is a \((K+2) \times (K+2)\) matrix given by

\[
\frac{\partial^2 L}{\partial \theta \partial \theta'} = \begin{pmatrix} X'X & -X' \\ \epsilon' \epsilon_{\lambda\lambda} + \epsilon'_{\lambda} \epsilon_{\lambda} & -q' \ln Y \\
\frac{\epsilon' \epsilon}{(\sigma^2)^2} - \frac{T}{(2\sigma^2)} & \frac{X'X}{(\sigma^2)} \end{pmatrix}
\]

\[
\ldots (2.4.17)
\]

Use has been made of the first order computations in the derivation of (2.4.17). The \((k\times1)\) vector of zeros in (2.4.17) results from the vector

\[
\frac{\partial^2 L}{\partial \beta \partial \sigma^2} = -(\sigma^{-2})^2 X' \epsilon = -\sigma^{-2} \frac{\partial L}{\partial \beta} = 0.
\]
The substitution of \( q^1 \ln Y \) in (2.4.17) is obtained from the first order condition given in (2.4.15). The inverse of the negative of (2.4.17), evaluated at \( L_{\text{max}} \) is the estimated covariance matrix of the parameter estimates.

**Maximizing the concentrated log-likelihood Function**

The parameter \( \sigma^2 \) can be solved from the equation (2.4.16) as

\[
\sigma^2 = \frac{\varepsilon_1^\prime \varepsilon}{(Y^{(4)} - \bar{X}\beta)(Y^{(4)} - \bar{X}\beta)}
\]  

...(2.4.18)

By substituting (2.4.18) in to (2.4.13) gives the concentrated log-likelihood function as

\[
L(\beta, \lambda; X, Y) = K_2 - \frac{n}{2} \ln \sigma^2 + (\lambda - 1)q^1 \ln Y
\]  

...(2.4.19)

Where,

\( K_2 \) is a constant contained the last term of (13) as \(-\frac{n}{2}\).

First order conditions for the maximization of (2.4.19) gives

\[
\frac{\partial L}{\partial \beta} = -\frac{n}{2\sigma^2} \frac{\partial \sigma^2}{\partial \beta} - \frac{X^1 \varepsilon}{\sigma^2} = 0
\]  

...(2.4.20)

\[
\frac{\partial L}{\partial \lambda} = \frac{n}{2\sigma^2} \frac{\partial \sigma^2}{\partial \lambda} + q^1 \ln Y = 0
\]

...(2.4.21)
The second order conditions require the matrix

\[
\frac{\partial^2 L}{\partial \theta \partial \theta^\prime} = \hat{\sigma}^2 \begin{bmatrix} X'X & -X'\epsilon_X \\ \epsilon_X'X & \epsilon_X'\epsilon_X + \epsilon_X'\epsilon_X - 2(q'\ln Y)\hat{\sigma}^2 \end{bmatrix} \quad \text{(2.4.22)}
\]

to be negative definite. The inverse of the negative of (2.4.22) gives the estimated covariance matrix of \( \hat{\theta} = [\hat{\beta}, \hat{\lambda}] \).

Spitzer (1982) has shown that the covariance matrix obtained from (2.4.22) is identical to the covariance matrix for \( \hat{\beta} \) and \( \hat{\lambda} \) obtained from (2.4.17).

Spitzer (1982) has also explained the nonlinear least squares method by taking transformation through the geometric mean of the sample Y's. He has shown that the IOLS estimation does not provide correct ML covariance estimates directly.

There are several approaches to the estimation of parameters in the models which contain the Box-Cox transformation. All the methods can make to obtain the same parameter estimates and the same covariance matrix estimates.
2.5 BOX-COX TRANSFORMATION FOR NONNORMAL DISTURBANCES

Some times, in the case of some statistical models, the dependent variable may not be normally distributed, but there may exist a transformation such that the transformed observations are normally distributed. For instance, consider the nonlinear model

\[ Y_i = e^{X_i' \beta + \varepsilon_i}, \quad i = 1, 2, \ldots, n. \] ... (2.5.1)

Where,

- \( Y_i \) is the \( i \)th observation on a dependent variable;
- \( X_i \) is a (kx1) vector containing the \( i \)th observation on explanatory variables;
- \( \beta \) is a (kx1) vector of parameters;
- \( \varepsilon_i \) are i.i.d. \( N(0, \sigma_i^2) \)

Here,

- \( Y_i \) are log normally distributed and heteroscedastic with variances

\[ V(Y_i) = \left[ e^{X_i' \beta} \right]^2 \left[ e^{\sigma_i^2} - 1 \right] \quad i = 1, 2, \ldots, n. \] ... (2.5.2)

By taking logarithms on both sides of (2.5.1) gives

\[ \ln Y_i = X_i' \beta + \varepsilon_i, \] ... (2.5.3)

Where \( \ln Y_i \) is normally distributed, homoscedastic and a linear function of \( \beta \). Application of OLS to (2) yields the best linear unbiased estimator (BLUE) for \( \beta \).
It is a particular case of a class of transformations given by Box and Cox (1964). For this, they assumed that there exists a value $\lambda$ such that

\[
\left[ \frac{Y_i^\lambda - 1}{\lambda} \right] = Y_i^\lambda = X_i^\prime \beta + \epsilon_i \quad \ldots (2.5.4)
\]

Where, $\epsilon_i$'s are i.i.d. $N(0, \sigma^2)$.

The transformed dependent variable $Y^{(\lambda)}$ is

i. normally distributed,

ii. homoscedastic

and iii. has an expectation that is linear in $\beta$.

It can be shown that

\[
\lim_{\lambda \to 0} \left[ (Y_i^\lambda - 1) \right] = \ln Y_i \quad \ldots (2.5.5)
\]

So equation (2.5.3) is regarded as a particular case of (2.5.4),

Where $\lambda = 0$

$\lambda = 1$ yields the familiar linear model $Y_i = X_i^\prime \beta + \epsilon_i$.

Thus when observations on the dependent variable come from a skewed distribution, the Box–Cox transformation may be useful; however, because it also makes assumptions about the functional form and the homoscedasticity of the transformed observations, it should be used cautiously.