STUDY ON THE BOX AND COX POWER TRANSFORMATION TO NORMALITY

By

GWONG-CHAIN SUN

Bachelor of Science
in Engineering
National Taiwan University
Taipei, Taiwan, Republic of China
1968

Master of Science
State University of New York at Stony Brook
Stony Brook, New York
1972

Master of Science
University of Arkansas
Fayetteville, Arkansas
1976

Submitted to the Faculty of the Graduate College
of the Oklahoma State University
in partial fulfillment of the requirements
for the Degree of
DOCTOR OF PHILOSOPHY
May, 1978

STUDY ON THE BOX AND COX POWER TRANSFORMATION

TO NORMALITY

Thesis Approved:

Lyle Broemeling
Thesis Adviser

Ronald W. McNew

Don Holbert

David M. Verhalen

Norman N. Durham
Dean of the Graduate College

ACKNOWLEDGMENTS

I would like to express my sincere appreciation to my adviser, Dr. Lyle D. Broemeling, for his considerable help in solving this problem and in preparing this thesis after my initial adviser, Dr. J. Leroy Folks, went on sabbatical leave in July, 1977. I would like to thank Dr. Folks for his help in initiation, guidance, and counsel in the early stages of solving this problem.

I wish to thank each member of my advisory committee, Dr. Ronald W. McNew, Dr. Donald Holbert, and Dr. Laval M. Verhalen, for their help and constructive suggestions.

Finally, my deep gratitude is extended to my parents, Mr. and Mrs. C. Sun, for their interest, encouragement, and prayers. Likewise, I wish to express sincere appreciation to my wife, Nian-Hwa, and our children, Shiou-Gwo and Nara, for their sacrifices and prayers.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION AND STATEMENT OF THE PROBLEM

The usual techniques of analyzing data are based on such assumptions as additivity of effects, constancy of variance, normality of distribution, and independence of observations. If the collected data does not match the assumptions of the conventional methods of analysis, Tukey (1957) suggested two choices: we may develop new methods of analysis with assumptions which fit the data in its "original" form, or we may transform the data to fit the assumptions which we need. If we can find a satisfactory transformation, it will almost always be easier and simpler rather than developing new methods of analysis.

We know that the distribution of χ^2 tends to normality as the number of degrees of freedom approaches infinity, but for a given number of degrees of freedom, we hope that we have a good normal approximation. For example, Fisher used the 1/2th power and Wilson and Hilferty used the 1/3rd power of χ^2 in seeking a transformation to approximate normality. We are also familiar with two apparently unrelated transformations, $Z=\sqrt{Y}$ and $Z=\log(Y+\lambda)$, which are usually applied to transform a Poisson distribution of average value λ to approximate normal distribution.

Tukey (1957) suggested a family of transformations with an unknown power parameter of a rational number and Box and Cox (1964) modified it. This family of power transformations includs the usual ways of making transformations such as logarithm, square, square root, inverse, and so on. The first task is to estimate this power **from the sample information. Table I provides an example from Box and Cox (1964) illustrating how the power transformation works.**

TABLE I

SURVIVAL TIMES OF ANIMALS IN A 3 x 4 FACTORIAL EXPERIMENT

Poison	Treatment			
	A	B	C	D
I	0.31	0.82	0.43	0.45
	0.45	1.10	0.45	0.71
	0.46	0.88	0.63	0.66
	0.43	0.72	0.76	0.62
II	0.36	0.92	0.44	0.56
	0.29	0.61	0.35	1.02
	0.40	0.49	0.31	0.71
	0.23	1.24	0.40	0.38
III	0.22	0.30	0.23	0.30
	0.21	0.37	0.25	0.36
	0.18	0.38	0.24	0.31
	0.23	0.29	0.22	0.33

We want the model to be such: (1) no interaction terms are needed, **(2) the error variance is constant, and (3) the observations are normally distributed. Here, we fit the model with both the original data and the transformed data and check their residuals against those assumptions. For the original data, the test statistic of Shapiro and Wilk's (1965) normality test is 0.92325 with a corresponding observed** significance level of less than 0.01. When we apply the power transformation, the maximum likelihood estimate of power is -0.75. The normality test statistic of the residuals of the transformed data (with power -0.75) is 0.98411 with a corresponding observed significance level near 0.9. These normality test results tell us that if we want to analyze this example under the above assumptions, the data should be transformed.

Since this family of power transformations has been proposed, most studies have emphasized the maximum likelihood estimate of power and investigated its properties. In addition, Box and Cox (1964) have used the Bayesian approach. Because there is no closed form for the maximum likelihood estimate, one needs a great deal of numerical computations. Other methods of estimation for the univariate case will be considered.

One must also consider the restrictions that the observations are positive and that the range of transformed observations is not from negative infinity to positive infinity, except when the power is equal to zero. Thus, it is invalid to use the full normal distribution as the likelihood function of the transformed observations. Hence, one assumes an approximate normal distribution of the transformed observations for practical situations. Another problem is the degree of approximation. It is valid to use the truncated normal distribution as the likelihood function of the transformed observations, and from the estimate of the truncation error, one can tell the degree of fit of the approximate normal distribution.

In this paper, Chapter II gives a brief review of the literature on this subject. In Chapter III, we present three different methods with which to estimate the power transformation for the univariate case and their properties are investigated. A generalized method of obtaining maximum likelihood estimates based on the truncated normal distribution is explained in Chapter IV. A summary and a brief study of robustness and possibilities for further research are described in Chapter V.

CHAPTER II

LITERATURE REVIEW

Tukey (1957) dubbed the family of power transformation defined by

$$Y_T^{(\lambda)} = \begin{cases} Y^{\lambda}, & \lambda \neq 0, \\ \log Y, & \lambda = 0, \end{cases} \tag{2.1}$$

as the "simple family". He studied their topology and charted their structural features for $|\lambda| \leq 1$.

Box and Cox (1964) altered the definition of the simple family to

$$Z = Y^{(\lambda)} = \begin{cases} (Y^{\lambda} - 1)/\lambda , & \lambda \neq 0, \\ \log Y, & \lambda = 0, \end{cases} \tag{2.2}$$

which has all the features of Tukey's power transformation and in addition is continuous at $\lambda = 0$. Both transformations, (2.1) and (2.2), assume that Y is positive to avoid the inadmissibility of λ.

The fundamental assumption made by Box and Cox was that for some λ, the transformed observations defined by (2.2) can be treated as independent and normally distributed with constant variance σ^2 and with expectations defined by the linear model

$$E\left\{\underline{y}^{(\lambda)}\right\} = A\,\underline{\theta}, \tag{2.3}$$

where $\underline{y}^{(\lambda)}$ is the column vector of transformed observations, A is a known constant matrix, and $\underline{\theta}$ is a vector of unknown parameters associated with the transformed observations. They discussed estimation of the parameter λ from a sampling theory and Bayesian point of view.

The likelihood of the original observations $\underline{y}$ is obtained by multiplying the normal density with the **Jacobian** of the transformation, thus,

$$L(\lambda,\underline{\theta},\sigma^2)=(2\pi\sigma^2)^{-n/2}\exp\left\{-(\underline{y}^{(\lambda)}-A\underline{\theta})'(\underline{y}^{(\lambda)}-A\underline{\theta})/2\sigma^2\right\}J(\lambda:\underline{y}), \qquad (2.4)$$

where

$$J(\lambda:\underline{y})=\prod_1^n\left|\frac{dy_i^{(\lambda)}}{dy_i}\right| = \prod_1^n \; y_i^{\lambda-1}.$$

Finding the maximum likelihood estimate of λ has two steps. First, for a given λ, the maximum likelihood estimates of the $\underline{\theta}$'s and σ^2 are

$$\hat{\underline{\theta}}(\lambda)= (A'A)^{-1}A'\underline{y}^{(\lambda)}, \qquad (2.5)$$

$$\widehat{\sigma^2}(\lambda) = \underline{y}^{(\lambda)'}\left[I- A(A'A)^{-1}A'\right]\underline{y}^{(\lambda)}/n = S(\lambda)/n, \qquad (2.6)$$

if A is of full rank. If A is not of full rank, we may replace $(A'A)^{-1}$ by its generalized inverse (Rao, 1962). Except for a constant, the maximized log likelihood is

$$\log L_{max}(\lambda)=-\tfrac{1}{2}n\log\widehat{\sigma^2}(\lambda) + \log J(\lambda:\underline{y}). \qquad (2.7)$$

Second, plot the maximized log likelihood, $\log L_{max}(\lambda)$, against λ for a trial series of values. From this plot, we choose the value which maximizes the $\log L_{max}(\lambda)$ to be the maximum likelihood estimate of λ. Or, let the first derivative of $\log L_{max}(\lambda)$ with respect to λ equal zero and solve this equation by some numerical methods. Box and Cox (1964) also pointed out that the $100(1-\alpha)$ per cent confidence region of λ can be obtained approximately from

$$\log L_{max}(\hat{\lambda}) - \log L_{max}(\lambda) < \tfrac{1}{2}\chi_1^2(\alpha). \tag{2.8}$$

Draper and Cox (1975) derived expressions for the precision of the maximum likelihood estimate of λ for a simple random sample, i.e., $E(y_i^{(\lambda)}) = \mu$ for $i = 1,2,\ldots,n$. The approximate variance is

$$V(\hat{\lambda}) = \frac{2}{3n\delta^2}\left(1 - \frac{1}{3}\gamma_1^2 + \frac{7}{18}\gamma_2\right)^{-1}, \tag{2.9}$$

where

$$\gamma_1 = \mu_{(3)}/\sigma^3,$$

$$\gamma_2 = \mu_{(4)}/\sigma^4 - 3,$$

$$\delta = \lambda\sigma/(1+\lambda\mu),$$

and σ^2 is the variance **and $\mu_{(i)}$ is the ith central moment of the y's.**

Hinkley (1975) gave an estimate of λ for the power transformations such that the transformed observations have a symmetric distribution and he **termed it a "quick estimate".** Let observations $y_1, y_2, \ldots, y_n$ have the common distribution function $F(y)$ with quantiles ξ_s defined by $F(\xi_s) = s$ $(0 < s < 1)$. If there exists a λ such that the transformed observations have a symmetric distribution, then the p and 1-p quantiles will be symmetrically placed about the median, i.e.,

$$\xi_{\frac{1}{2}}^{\lambda} - \xi_p^{\lambda} = \xi_{1-p}^{\lambda} - \xi_{\frac{1}{2}}^{\lambda}. \tag{2.10}$$

If we denote the ordered values of the observations as $x_1 \leq \ldots \leq x_n$ and define the median $\bar{x}$ in the usual way, then the sample analogue of (2.10) is (unless $x_r = x_{n-r+1} = \bar{x}$)

$$\bar{x}^{\lambda} - x_r^{\lambda} = x_{n-r+1}^{\lambda} - \bar{x}^{\lambda}, \tag{2.11}$$

where $r = [np]$.

From equation (2.11), one may estimate λ associated with the value of p for $\lambda \neq 0$; and the estimate of λ is zero if and only if the following identity is satisfied

$$\frac{\bar{x}}{x_r} = \frac{x_{n-r+1}}{\bar{x}}. \tag{2.12}$$

This estimate is consistent and has a limiting normal distribution as n approaches infinity. If the original distribution function F(y) has the density function f(y), then the asymptotic variance of this estimate of λ is

$$V_{\hat{\lambda}}(\lambda,p) = \frac{\lambda^4}{(\alpha_p^\lambda \log\alpha_p^\lambda + \alpha_q^\lambda \log\alpha_q^\lambda)^2} \Big\{ h_{\frac{1}{2}}^2 + pq(\alpha_p^{2\lambda} h_p^2 + \alpha_q^{2\lambda} h_q^2)$$

$$-2p(\alpha_p^\lambda h_q + \alpha_q^\lambda h_p) h_{\frac{1}{2}} + 2p^2 \alpha_p^\lambda \alpha_q^\lambda h_p h_q \Big\}, \tag{2.13}$$

where $p+q=1$, $\alpha_s = \xi_s / \xi_{\frac{1}{2}}$ and $h_s^{-1} = \xi_s f(\xi_s)$.

Andrews, Gnanadesikan, and Warner (1971) investigated the power transformations of multivariate data. They discussed the three approaches of marginal normality, joint normality, and directional normality for the bivariate situation. Although each approaches used the likelihood method, the three had different objectives and properties.

Schlesselman (1971) suggested some alternative power transformation families with the scale invariance property. Atkinson (1973) devised a test for the power transformation to normality, and other aspects of normal-theory estimation and inferences about λ have been investigated by Fraser (1967), Andrews (1971), and Lindsey (1972, 1975).

CHAPTER III

METHODS OF ESTIMATION OF THE POWER TRANSFORMATION TO AN APPROXIMATE NORMAL DISTRIBUTION FOR THE UNIVARIATE CASE

This chapter describes three methods to estimate the parameter λ of the power transformation to normality from a sample with univariate data (if such transformation exists). Some properties of these estimates are investigated. The first method is called the "quantile estimate" and is based on the p and q quantiles for $0 < p, q < \frac{1}{2}$. The second is termed the "plotting estimate" which utilizes a normal plotting technique, and the last is the "maximum W-statistic estimate" and is based on maximizing the Shapiro-Wilk (1965) W-statistic.

Quantile Estimate

Procedure

Suppose $Y_1, Y_2, \ldots, Y_n$ are continuous, nonnegative, independent, and identically distributed random variables. If there exists a λ such that the Box and Cox power transformed random variables $Y_i^{(\lambda)}$'s are distributed approximately normally with mean μ and variance σ^2, we can write the transformed random variables as

$$Y_i^{(\lambda)} = \mu + \sigma X_i, \tag{3.1}$$

where $X_i \sim N(0, 1)$ for $i=1,2,\dots,n$.

Therefore, we have

$$\begin{cases} Y_i^\lambda = 1+\lambda\mu+\lambda\sigma X_i & (\lambda \neq 0), \\ \log Y_i = \mu+\sigma X_i & (\lambda=0). \end{cases} \tag{3.2}$$

Let $Y_1, Y_2, \dots, Y_n$ have the common distribution function $F(y)$ with quantiles ξ_s defined by $F(\xi_s)=s$ for $0<s<1$. Then,

$$\begin{cases} \xi_s^\lambda = 1+\lambda\mu+\lambda\sigma\eta_s & (\lambda \neq 0), \\ \log \xi_s = \mu+\sigma\eta_s & (\lambda=0), \end{cases} \tag{3.3}$$

where η_s is the s quantile of the standard normal distribution and $\eta_s = -\eta_{1-s}$.

Suppose p and q are chosen such that $p \neq q$ and $0 < p,q < \frac{1}{2}$, then from (3.3) we have the following identities;

$$\begin{cases} \xi_p^\lambda = 1+\lambda\mu+\lambda\sigma\eta_p, \\ \xi_{1-p}^\lambda = 1+\lambda\mu-\lambda\sigma\eta_p, \\ \xi_q^\lambda = 1+\lambda\mu+\lambda\sigma\eta_q, \\ \xi_{1-q}^\lambda = 1+\lambda\mu-\lambda\sigma\eta_q & \text{for } \lambda \neq 0, \end{cases} \tag{3.4}$$

and

$$\begin{cases} \log \xi_p = \mu+\sigma\eta_p, \\ \log \xi_{1-p} = \mu-\sigma\eta_p, \end{cases} \tag{3.5}$$

$$\begin{cases} \log \xi_q = \mu + \sigma \eta_q, \\ \log \xi_{1-q} = \mu - \sigma \eta_q & \text{for } \lambda = 0. \end{cases}$$

The four equations of (3.4) and (3.5) have three unknown parameters. After eliminating the unknown parameters μ and σ, we have two equations with only one unknown parameter λ. Thus,

$$\begin{cases} 2\,\eta_q\, \xi_{1-p}^{\lambda} = (\eta_p + \eta_q)\, \xi_{1-q}^{\lambda} + (\eta_q - \eta_p)\, \xi_q^{\lambda}, \\ \eta_q(\xi_{1-p}^{\lambda} - \xi_p^{\lambda}) = \eta_p(\xi_{1-q}^{\lambda} - \xi_q^{\lambda}) & \text{for } \lambda \neq 0 \end{cases} \tag{3.6}$$

and

$$\begin{cases} \left[\dfrac{\xi_{1-p}}{\xi_p}\right]^{\eta_q} = \left[\dfrac{\xi_{1-q}}{\xi_q}\right]^{\eta_p}, \\ \xi_{1-p}\,\xi_p = \xi_{1-q}\,\xi_q & \text{for } \lambda = 0. \end{cases} \tag{3.7}$$

If we denote the ordered values of the observations by $y_{(1)},\dots, y_{(n)}$, then the estimating equations can be found by substituting for ξ_s the $y_{(i)}$ and for ξ_{1-s} the $y_{(n-i+1)}$ if $s = p$ and q, where i is defined by $i = [ns]$. Therefore, the estimate $\hat{\lambda}$ equals zero if the following identities are satisfied

$$\begin{cases} \left[\dfrac{y_{(n-i+1)}}{y_{(i)}}\right]^{\eta_q} = \left[\dfrac{y_{(n-j+1)}}{y_{(j)}}\right]^{\eta_p}, \\ y_{(n-i+1)}\,y_{(i)} = y_{(n-j+1)}\,y_{(j)}. \end{cases} \tag{3.8}$$

Otherwise, the estimate $\hat{\lambda} \neq 0$ can be found from the following two equations, namely,

$$\begin{cases} 2\eta_q y^{\hat{\lambda}}_{(n-i+1)} = (\eta_p + \eta_q) y^{\hat{\lambda}}_{(n-i+1)} + (\eta_q - \eta_p) y^{\hat{\lambda}}_{(j)}, \\ \eta_q (y^{\hat{\lambda}}_{(n-i+1)} - y^{\hat{\lambda}}_{(i)}) = \eta_p (y^{\hat{\lambda}}_{(n-j+1)} - y^{\hat{\lambda}}_{(j)}). \end{cases} \tag{3.9}$$

We can rewrite the equations (3.8) and (3.9) as

$$\begin{cases} b^a = c^{(a-1)}, \\ d^{\eta_q} = \left(\frac{c}{b}\right)^{\eta_p}, \qquad \text{if } \hat{\lambda} = 0 \end{cases} \tag{3.10}$$

and if $\hat{\lambda} \neq 0$

$$\begin{cases} ab^{\hat{\lambda}} + (1-a)c^{\hat{\lambda}} = 1, \\ \eta_q(1-d^{\hat{\lambda}}) = \eta_p(b^{\hat{\lambda}} - c^{\hat{\lambda}}), \end{cases} \tag{3.11}$$

where $a = (\eta_p + \eta_q)/2\eta_q$, $b = y_{(n-j+1)}/y_{(n-i+1)}$, $c = y_{(j)}/y_{(n-i+1)}$, and $d = y_{(i)}/y_{(n-i+1)}$.

A proof of the existence of a nonzero solution to (3.11) and two suggesting techniques are given in Appendix A. This nonzero solution is positive if $b^a > c^{(a-1)}$ and is negative if $b^a < c^{(a-1)}$. Since there are two equations in (3.11), it is easier to solve $\hat{\lambda}$ from the first and check it against the second. If the checking fails, we can assume there is no power transformation to normality for that particular pair (p,q).

Properties

Definition. An estimator $T = t(x_1, x_2, \ldots, x_n)$ is defined to be scale free if and only if $t(cx_1, cx_2, \ldots, cx_n) = t(x_1, x_2, \ldots, x_n)$ for all values $x_1, x_2, \ldots, x_n$ and all $c > 0$.

1. The quantile estimate is scale free.

$$\begin{cases} \left[\dfrac{cy_{(n-i+1)}}{cy_{(i)}}\right]^{\eta_q} = \left[\dfrac{cy_{(n-j+1)}}{cy_{(j)}}\right]^{\eta_p}, \\ cy_{(n-i+1)}\, cy_{(i)} = cy_{(n-j+1)}\, cy_{(j)} \qquad \text{for } \hat{\lambda}=0, \end{cases}$$

$$\Rightarrow \begin{cases} \left[\dfrac{y_{(n-i+1)}}{y_{(i)}}\right]^{\eta_q} = \left[\dfrac{y_{(n-j+1)}}{y_{(j)}}\right]^{\eta_p}, \\ y_{(n-i+1)}\, y_{(i)} = y_{(n-j+1)}\, y_{(j)}. \qquad \text{for } \hat{\lambda}=0. \end{cases}$$

For $\hat{\lambda} \neq 0$, from (3.9)

$$\begin{cases} 2\eta_q[cy_{(n-i+1)}]^{\hat{\lambda}} = (\eta_p+\eta_q)[cy_{(n-j+1)}]^{\hat{\lambda}} + (\eta_q-\eta_p)[cy_{(j)}]^{\hat{\lambda}}, \\ \eta_q\left\{[cy_{(n-i+1)}]^{\hat{\lambda}} - [cy_{(i)}]^{\hat{\lambda}}\right\} = \eta_p\left\{[cy_{(n-j+1)}]^{\hat{\lambda}} - [cy_{(j)}]^{\hat{\lambda}}\right\} \end{cases}$$

$$\Rightarrow \begin{cases} 2\eta_q\, y^{\hat{\lambda}}_{(n-i+1)} = (\eta_p+\eta_q)\, y^{\hat{\lambda}}_{(n-j+1)} + (\eta_q-\eta_p)\, y^{\hat{\lambda}}_{(j)}, \\ \eta_q\left[y^{\hat{\lambda}}_{(n-i+1)} - y^{\hat{\lambda}}_{(i)}\right] = \eta_p\left[y^{\hat{\lambda}}_{(n-j+1)} - y^{\hat{\lambda}}_{(j)}\right]. \end{cases}$$

The estimating equations based on the $(cy_1, cy_2, \ldots, cy_n)$ are identical to the estimating equations based on the $(y_1, y_2, \ldots, y_n)$.

2. The quantile estimate is consistent.

Let us rewrite (3.6) to be, for $\lambda \neq 0$,

$$\begin{cases} g_1(\lambda, \xi_{1-p}, \xi_q, \xi_{1-q}) = 2\eta_p\xi^{\lambda}_{1-p} - (\eta_p+\eta_q)\xi^{\lambda}_{1-q} - (\eta_q-\eta_p)\xi^{\lambda}_q = 0, \\ g_2(\lambda, \xi_p, \xi_{1-p}, \xi_q, \xi_{1-q}) = \eta_q(\xi^{\lambda}_{1-p} - \xi^{\lambda}_p) - (\xi^{\lambda}_{1-q} - \xi^{\lambda}_q)\eta_p = 0. \end{cases} \tag{3.12}$$

Since the function g_1 and g_2 are differentiable for $\lambda \neq 0$, there exists continuous functions h_1 and h_2, for $\lambda \neq 0$, such that $\lambda = h_1(\xi_{1-p}, \xi_q, \xi_{1-q})$ and $\lambda = h_2(\xi_p, \xi_{1-p}, \xi_q, \xi_{1-q})$ are the solutions by letting the functions

g_1 and g_2 equal zero, respectively. Similarly, from (3.9) we can get $\hat{\lambda} = h_1(y_{(n-i+1)}, y_{(j)}, y_{(n-j+1)})$ and $\hat{\lambda} = h_2(y_{(i)}, y_{(n-i+1)}, y_{(j)}, y_{(n-j+1)})$.

We know that $y_{(i)}$, $y_{(n-i+1)}$, $y_{(j)}$, $y_{(n-j+1)}$ converge in probability to ξ_p, ξ_{1-p}, ξ_q, ξ_{1-q}, respectively. These convergences in probability imply that $\hat{\lambda}$ converges to λ in probability because of the continuity of the functions h_1 and h_2 (Rao, 1973).

3. The quantile estimate has an asymptotic normal distribution. We use the joint asymptotic normality of the order statistics $Y_{(r_1)}, \ldots, Y_{(r_n)}$ for $r_j = [np_j]$, $0 < p_1 < \ldots < p_m < 1$. Specifically, if the original distribution function $F(y)$ has density $f(y)$ and quantiles $\xi_s \doteq F^{-1}(s)$, then the vector $(Y_{(r_1)}, Y_{(r_2)}, \ldots, Y_{(r_n)})$ has a limiting multivariate normal distribution with mean vector $(\xi_{p_1}, \xi_{p_2}, \ldots, \xi_{p_n})$ and variance covariance determined by (David, 1970)

$$\text{cov}\,(Y_{(r_i)}, Y_{(r_j)}) = \frac{p_i(1-p_j)}{nf(\xi_{p_i})f(\xi_{p_j})} \quad \text{for } i \leq j. \tag{3.13}$$

From the consistency of $\hat{\lambda}$, we can say $\hat{\lambda} = \lambda + o(1)$, where $o(x)$ is called little o notation (Olmasted 1959). Suppose $\lambda \neq 0$ and with $i = [np]$, $j = [nq]$ for $0 < p < q < \frac{1}{2}$, we define random variables W_p, W_{1-p}, W_q, W_{1-q} by

$$\begin{cases} Y_{(i)} = \xi_p(1 + n^{-\frac{1}{2}} W_p), \\ Y_{(j)} = \xi_q(1 + n^{-\frac{1}{2}} W_q), \\ Y_{(n-i+1)} = \xi_{1-p}(1 + n^{-\frac{1}{2}} W_{1-p}), \\ Y_{(n-j+1)} = \xi_{1-q}(1 + n^{-\frac{1}{2}} W_{1-q}). \end{cases} \tag{3.14}$$

By the joint asymptotic normality of $Y_{(k)}$'s, we have the joint asymptotic normality of $W_{(k)}$'s with means zeros and variance covariance matrix determined by

$$\text{cov}\,(W_{p_s},\, W_{p_t}) = \frac{p_s\,(1-p_t)}{\xi_{p_s}\,\xi_{p_t}\,f(\xi_{p_s})\,f(\xi_{p_t})} \quad \text{for } p_s \leq p_t. \qquad (3.15)$$

The first equation of (3.9) can be rewritten as

$$2\eta_q\left[\xi_{1-p}(1+n^{-\frac{1}{2}}W_{1-p})\right]^{\hat{\lambda}} - (\eta_p+\eta_q)\left[\xi_{1-q}(1+n^{-\frac{1}{2}}W_{1-q})\right]^{\hat{\lambda}}$$
$$-(\eta_q-\eta_p)\left[\xi_q(1+n^{-\frac{1}{2}}W_q)\right]^{\hat{\lambda}} = 0. \qquad (3.16)$$

Expand it about $\hat{\lambda}=\lambda$ and after simplied by using (3.6), we get

$$2\eta_q\left[\xi_{1-p}\lambda n^{-\frac{1}{2}}W_{1-p} + (\hat{\lambda}-\lambda)\xi_{1-p}\log\xi_{1-p} + o(n^{-\frac{1}{2}}) + o(\hat{\lambda}-\lambda)\right]$$
$$-(\eta_p+\eta_q)\left[\xi_{1-q}\lambda n^{-\frac{1}{2}}W_{1-q} + (\hat{\lambda}-\lambda)\xi_{1-q}\log\xi_{1-q} + o(n^{-\frac{1}{2}}) + o(\hat{\lambda}-\lambda)\right]$$
$$-(\eta_q-\eta_p)\left[\xi_q\lambda n^{-\frac{1}{2}}W_q + (\hat{\lambda}-\lambda)\xi_q\log\xi_q + o(n^{-\frac{1}{2}}) + o(\hat{\lambda}-\lambda)\right] = 0. \qquad (3.17)$$

To the first order, (3.17) becomes

$$\sqrt{n}\,\frac{(\hat{\lambda}-\lambda)}{\lambda} = \frac{2\eta_q\xi_{1-p}W_{1-p} - (\eta_p+\eta_q)\xi_{1-q}W_{1-q} - (\eta_q-\eta_p)\xi_qW_q}{(\eta_p+\eta_q)\xi_{1-q}\log\xi_{1-q} + (\eta_q-\eta_p)\xi_q\log\xi_q - 2\eta_q\xi_{1-p}\log\xi_{1-p}}. \qquad (3.18)$$

By the joint asymptotic normality of W_{1-p}, W_{1-q}, and W_q, we have that $\sqrt{n}\,\frac{\hat{\lambda}-\lambda}{\lambda}$ is distributed asymptotically normal with mean zero and variance V, as n approaches infinity, where

$$V = \frac{1}{(\eta_p+\eta_q)\xi_{1-q}\log\xi_{1-q} + (\eta_q-\eta_p)\xi_q\log\xi_q - 2\eta_q\xi_{1-p}\log\xi_{1-p}}$$

$$4\eta_q^2 \frac{p(1-p)}{f^2(\xi_{1-p})} - 4\eta_q(\eta_p+\eta_q)\frac{p(1-q)}{f(\xi_{1-p})f(\xi_{1-q})}$$

$$+(\eta_p+\eta_q)^2\frac{q(1-q)}{f^2(\xi_{1-q})} - 4\eta_q(\eta_q-\eta_p)\frac{q(1-p)}{f(\xi_{1-p})f(\xi_q)}$$

$$+(\eta_q-\eta_p)^2\frac{q(1-q)}{f^2(\xi_q)} + 2(\eta_q^2-\eta_p^2)\frac{q^2}{f(\xi_{1-q})f(\xi_q)} \Bigg\} .$$

Example

Table II includes a random sample of 50 observations of y's generated by $Y^{-0.64} \sim N(0.5,\ 0.15^2)$ from the IBM normal generator subroutine. Table III gives the estimates of λ for each pair (p,q) if $0 < p < q < \frac{1}{2}$ by the iteration method (Appendix A).

TABLE II

RANDOM SAMPLE OF $Y^{-0.64} \sim N(0.5,\ 0.15^2)$

1.36	1.86	1.89	1.93	2.16	2.25	2.35	2.40	2.48	2.49
2.65	2.71	2.77	2.83	2.86	2.89	2.92	3.02	3.16	3.45
3.68	3.70	3.81	3.81	3.90	4.10	4.13	4.31	4.35	4.39
4.49	4.52	4.61	4.76	4.77	5.64	5.83	5.94	6.37	7.00
7.67	7.88	8.48	10.62	12.03	14.50	15.51	18.04	41.71	42.42

TABLE III

ESTIMATES OF λ FOR SELECTED PAIRS (p,q)

p	q	$\hat{\lambda}$	Check	p	q	$\hat{\lambda}$	Check	
0.05	0.10	-0.737	-0.066	0.15	0.35	-1.055	-0.010	
0.05	0.15	-0.777	-0.002	0.15	0.40	-0.736	-0.012	
0.05	0.20	-0.810	-0.011	0.15	0.45	-0.834	0.000	
0.05	0.25	-0.881	0.003	0.20	0.25	-1.225	0.007	
0.05	0.30	-0.751	-0.018	0.20	0.30	-0.610	0.007	
0.05	0.35	-0.848	-0.013	0.20	0.35	-1.116	0.006	
0.05	0.40	-0.706	-0.018	0.20	0.40	-0.638	-0.008	
0.05	0.45	-0.797	-0.000	0.20	0.45	-0.681	0.001	
0.10	0.15	-0.745	0.043	0.25	0.30	0.394	0.045	
0.10	0.20	-0.850	0.029	0.25	0.35	0.000	3.364	#
0.10	0.25	-0.962	0.034	0.25	0.40	-0.401	-0.030	
0.10	0.30	-0.749	0.007	0.25	0.45	-0.273	-0.004	
0.10	0.35	-0.922	0.006	0.30	0.35	-2.281	-0.004	
0.10	0.40	-0.697	0.004	0.30	0.40	-0.916	-0.015	
0.10	0.45	-0.802	0.005	0.30	0.45	-0.573	0.011	
0.15	0.20	-1.027	-0.008	0.35	0.40	3.129	11.702	#
0.15	0.25	-1.135	0.002	0.35	0.45	0.000	-2.196	#
0.15	0.30	-0.793	-0.012	0.40	0.45	8.340	4960.740	#

The # denotes that the check equation failed to correspond (check > 0.05 or check < -0.05) for that particular pair (p,q).

Plotting Estimate

Procedure

If there exists a power transformation such that the transformed observations $y_i^{(\lambda)}$'s are distributed independently, identically, and approximately normal with mean μ and variance σ^2, we can write the transformed observations as

$$y_i^{(\lambda)} = \mu + \sigma x_i \qquad \text{for } i = 1,2,\ldots,n, \tag{3.19}$$

where x_i's are distributed independently $N(0, 1)$.
Rewrite (3.19) as

$$\begin{cases} y_i^{\lambda} = 1 + \lambda\mu + \lambda\sigma x_i, & \lambda \neq 0, \\ \log y_i = \mu + \sigma x_i, & \lambda = 0 \ (i = 1,2,\ldots,n). \end{cases} \tag{3.20}$$

The equality of (3.20) holds for the order statistics, thus,

$$\begin{cases} y_{(i)}^{\lambda} = 1 + \lambda\mu + \lambda\sigma x_{(i)}, & \lambda \neq 0, \\ \log y_{(i)} = \mu + \sigma x_{(i)}, & \lambda = 0 \ (i = 1,2,\ldots,n). \end{cases} \tag{3.21}$$

Taking expectations on both sides, we have

$$\begin{cases} E\left[y_{(i)}^{\lambda}\right] = 1 + \lambda\mu + \lambda\sigma m_i, & \lambda \neq 0, \\ E\left[\log y_{(i)}\right] = \mu + \sigma m_i, & \lambda = 0 \ (i = 1,2,\ldots,n), \end{cases} \tag{3.22}$$

where m_i is the expected value of the ith order statistic of the standard normal distribution.

Let $y_{(1)}, y_{(2)}, \ldots, y_{(n)}$ denote the ordered values of observations $y_1, y_2, \ldots, y_n$. If we know the true λ and plot the n points $(m_i, y_{(i)}^{\lambda})$ for $\lambda \neq 0$ and $(m_i, \log y_{(i)})$ for $\lambda = 0$, they should lie on a straight line. From (3.20) replacing the expected values by observations, we have the estimating equations

$$\begin{cases} y_{(i)}^{\hat{\lambda}} = 1 + \hat{\lambda}\hat{\mu} + \hat{\lambda}\hat{\sigma} m_i, & \hat{\lambda} \neq 0, \\ \log y_{(i)} = \hat{\mu} + \hat{\sigma} m_i, & \hat{\lambda} = 0 \ (i = 1, 2, \ldots, n). \end{cases} \tag{3.23}$$

If we find some value of $\hat{\lambda}$ such that the n points $(m_i, y_{(i)}^{\hat{\lambda}})$ for $\hat{\lambda} \neq 0$ and $(m_i, \log y_{(i)})$ for $\hat{\lambda} = 0$ fit a straight line, then this value is called the plotting estimate of λ.

We know that if $y^{\lambda} = a + bx$ is the true equation, then the plot of this equation is a straight line with the scale of vertical axis y^{λ}. However, if we change the scale of the vertical axis y^{λ} to y^{λ_0}, then the plot is not a straight line. Let us rewrite this as $(y^{\lambda_0})^{\lambda/\lambda_0} = a + bx$; thus, we can see that its plot will be a monotone concave curve if $\lambda/\lambda_0 > 1$ and monotone convex curve if $\lambda/\lambda_0 < 1$ with the restriction of $y > 0$. From this fact, the suggested procedure for finding the plotting estimate is given as follow:

1. Guess a value, say λ_0, and plot the n points $(m_i, y_{(i)}^{\lambda_0})$ for $\lambda_0 \neq 0$ and $(m_i, \log y_{(i)})$ for $\lambda_0 = 0$ on regular plotting paper. If these points fit a straight line, then this value λ_0 will be the estimate of λ.

2. If the fitting line is concave or nearly concave, then we need to try another guess $\lambda_1 > \lambda_0$ for $\lambda_0 > 0$ and $\lambda_1 < \lambda_0$ for $\lambda_0 < 0$.

3. If the fitting line is convex or nearly convex, then we

need to try another guess $\lambda_1 < \lambda_o$ for $\lambda_o > 0$ and $\lambda_1 > \lambda_o$ for $\lambda_o < 0$.

4. Continue the same procedure until the fitting line is almost a straight line and use the last guess as the plotting estimate of λ.

Note 1. We can use the quantiles $\Phi^{-1}(\frac{2i-1}{2n})$ for $i = 1,2,\ldots,n$ from the standard normal distribution instead of the m_i's.

Note 2. If we use the full normal plotting paper, the abscissas will be $(\frac{2i-1}{2n})$ for $i = 1,2,\ldots,n$.

Note 3. If we can not determine whether the fitted curve is convex or concave, it is suggested that one use two guesses, say $\lambda_1 > \lambda_o$ and $\lambda_2 < \lambda_o$, and see which direction produces the most improvement.

Note 4. After we have an estimate of λ, say $\hat{\lambda}$, then the estimate of μ will be $(b-1)/\hat{\lambda}$ where b is the vertical axis intercept of the fitted line and the estimate of σ will be the slope of the fitted line divided by $\hat{\lambda}$.

Note 5. The plotting estimate of λ is scale free. If $y^{\hat{\lambda}} = 1 + \hat{\lambda}\mu + \hat{\lambda}\sigma x$ is a straight line on the axes of x and $y^{\hat{\lambda}}$, then $(cy)^{\hat{\lambda}} = 1 + \hat{\lambda}\mu + \hat{\lambda}\sigma x$ will also be a straight line on the axes of x and $(cy)^{\hat{\lambda}}$.

Note 6. This method can quickly detect outliers.

<u>Example</u>

Here we are using the same data as presented in Table II. There are eight figures, Figure 1 through Figure 8, for our various guesses of $\hat{\lambda}$, namely -1.5, -1.0, -0.75, -0.65, -0.5, -0.25, 0.0, 0.5, respectively. Each figure has two plots of the n points $(m_i, y_{(i)}^{\hat{\lambda}})$ and $(\Phi^{-1}(\frac{2i-1}{2n}), y_{(i)}^{\hat{\lambda}})$ separately, where $y_{(i)}$ is the ith ordered observation of the data.

Suppose our first guess λ_o is 0.5 and we use the $\Phi^{-1}(\frac{2i-1}{2n})$,

$i = 1, 2, \ldots, n$, to be the abscissa. From Figure 8, we see that the plotting is convex, so we need to guess again. Because λ_0 is positive, hence, our next guess should be less than λ_0, say $\lambda_1 = -1.0$. From Figure 2, the plotting is convex too, but this λ_1 is negative. Consequently, the third guess should be greater than λ_1 and in addition be less than λ_0. Let $\lambda_2 = -0.5$ and the plotting is shown on Figure 5. The plot is closer to a straight line but is slighty concave. Suppose we try one more guess which is less than λ_2 because the plot is concave with a negative value of λ_2. Therefore, choose $\lambda_3 = -0.65$ and the corresponding plot which fits a straight line quite closely is in Figure 4. Finally, we can say that the plotting estimate of λ is -0.65 for the data in Table II.

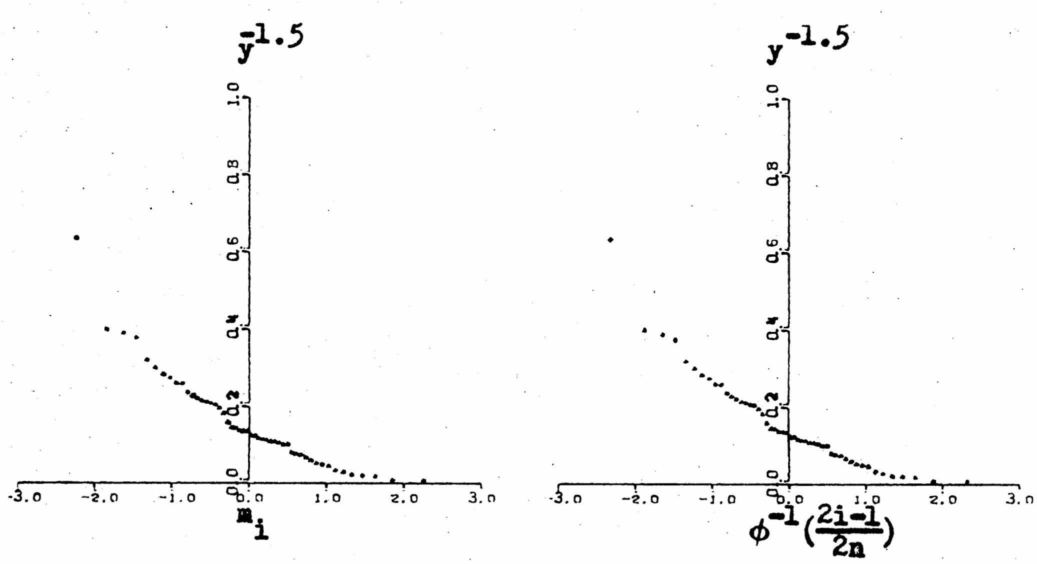

Figure 1. Plotting of $\hat{\lambda} = -1.5$

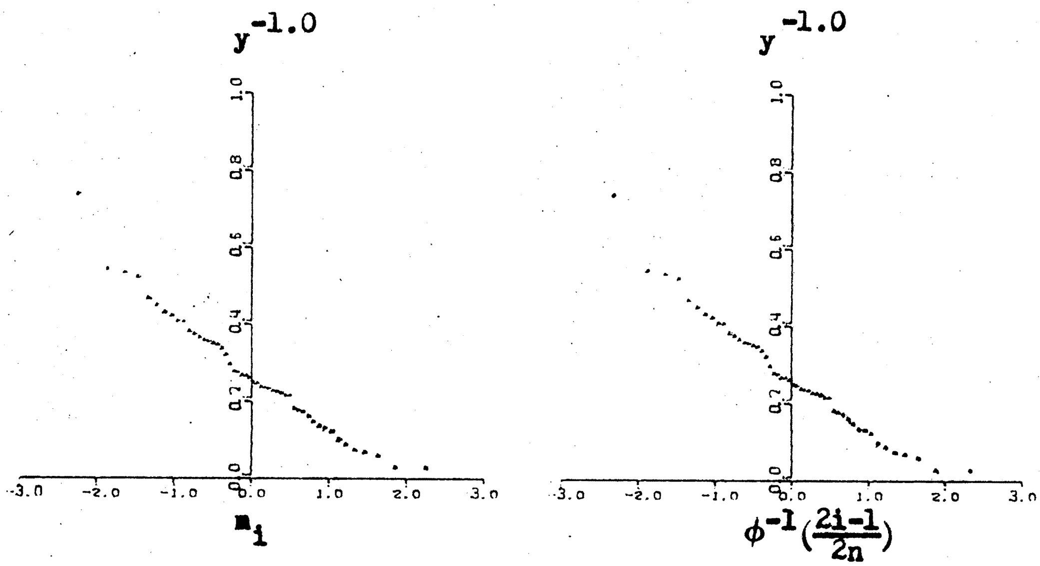

Figure 2. Plotting of $\hat{\lambda}=-1.0$

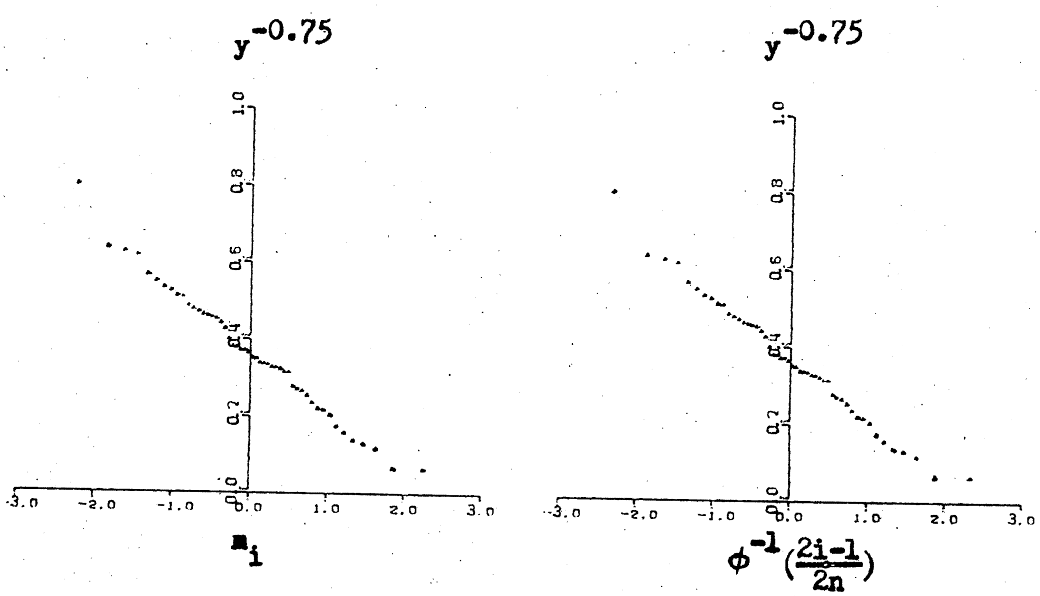

Figure 3. Plotting of $\hat{\lambda}=-0.75$

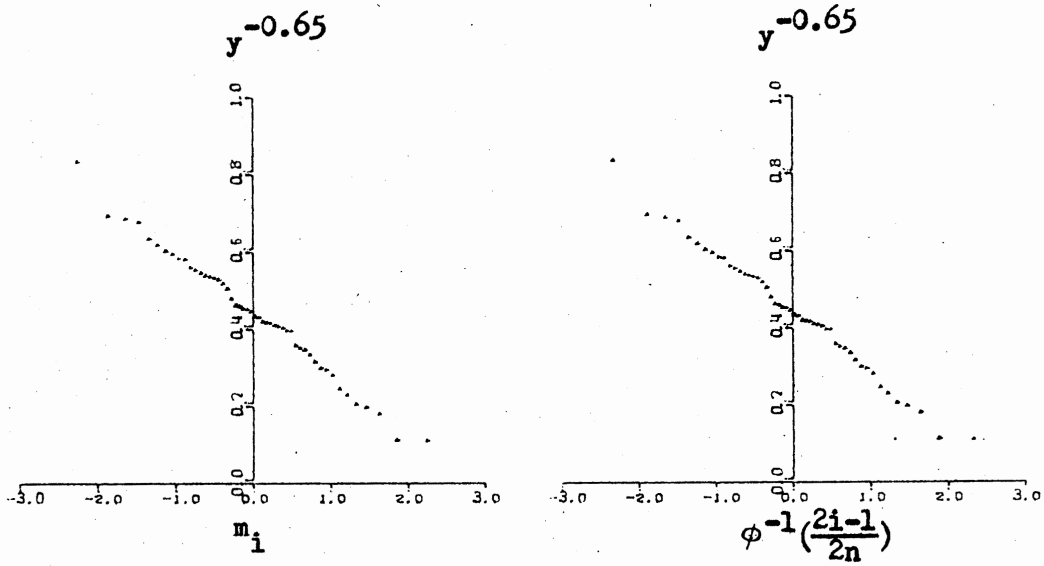

Figure 4. Plotting of $\hat{\lambda}=-0.65$

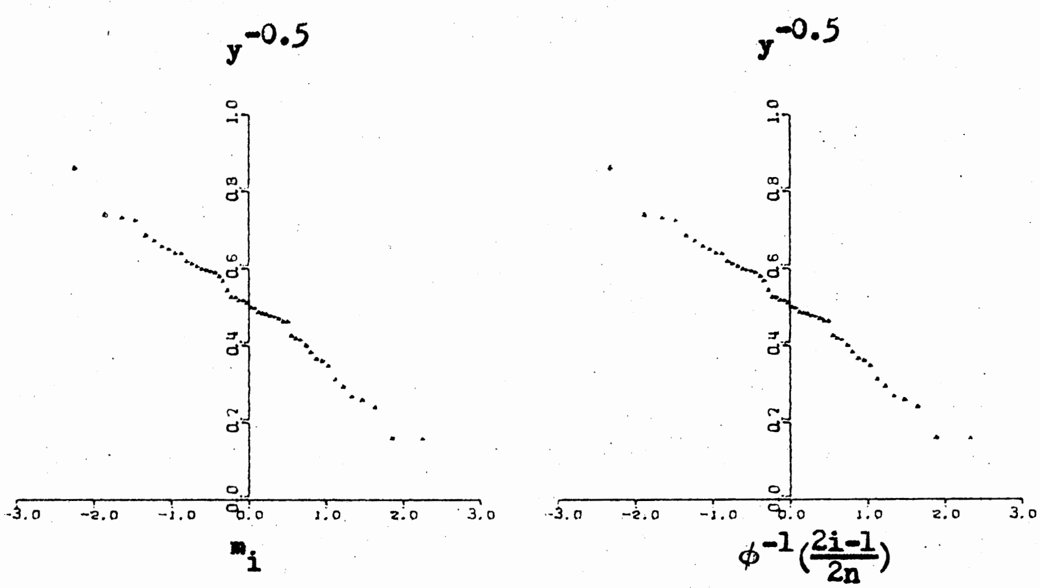

Figure 5. Plotting of $\hat{\lambda}=-0.5$

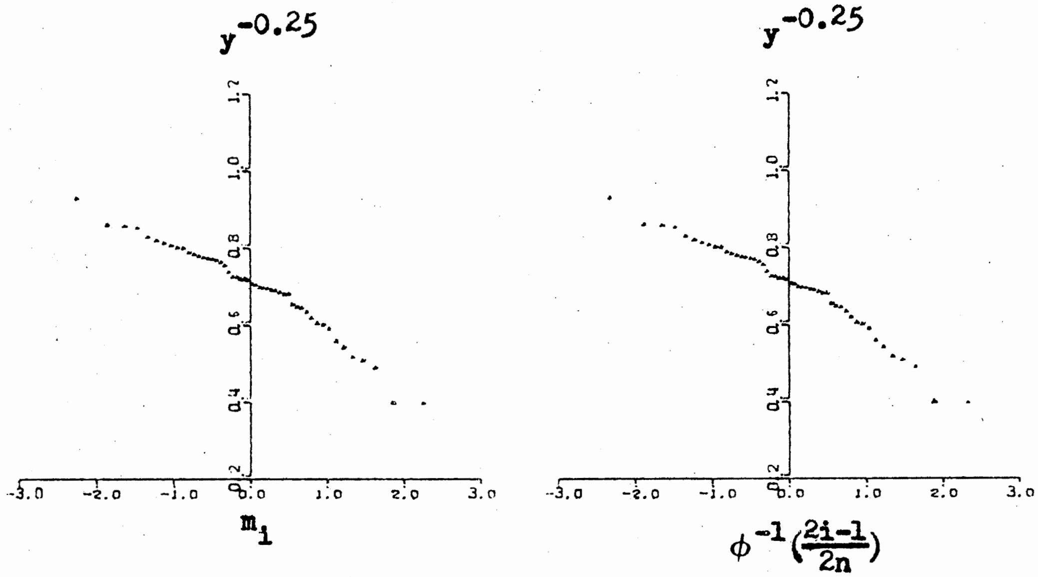

Figure 6. Plotting of $\hat{\lambda}=-0.25$

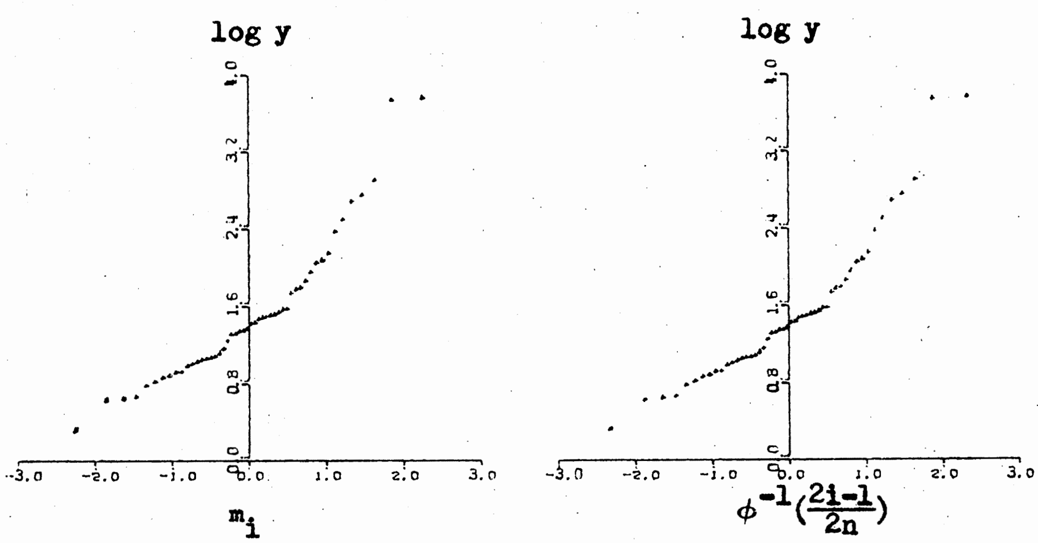

Figure 7. Plotting of $\hat{\lambda}=0.0$

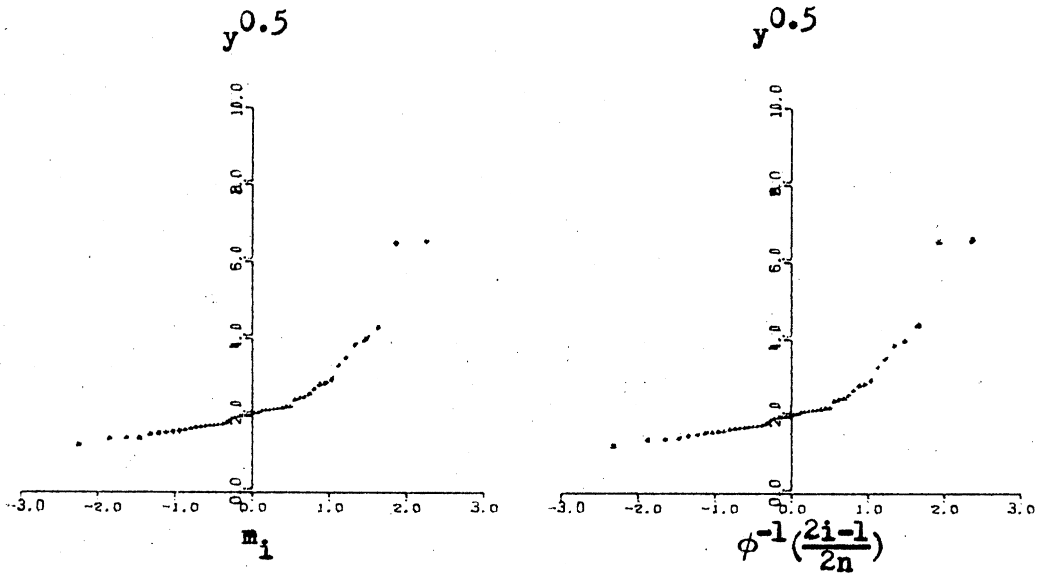

Figure 8. Plotting of $\hat{\lambda} = 0.5$

Maximum W-Statistic Estimate

Procedure

If there exists a λ such that the power transformed observations are distributed independently and are approximately normally distributed with mean μ and variance σ^2, the maximum W-statistic estimate is that value which maximizes the Shapiro and Wilk (1965) W-test statistic, i.e., maximizes the observed significance level, of the transformed observations. The W-statistic of the transformed observations for given λ is

$$W(\lambda) = \frac{\left(\sum_1^n a_i \, y_{(i)}^{(\lambda)} \right)^2}{\sum_1^n \left(y_{(i)}^{(\lambda)} - \frac{1}{n} \sum_1^n y_{(i)}^{(\lambda)} \right)^2}$$

$$
= \begin{cases} \dfrac{(\sum_1^n a_i\, y_{(i)}^{\lambda})^2}{\sum_1^n (y_i^{\lambda} - \frac{1}{n}\sum_1^n y_i^{\lambda})^2}, & \lambda \neq 0, \\[2ex] \dfrac{(\sum_1^n a_i \log y_{(i)})^2}{\sum_1^n (\log y_i - \frac{1}{n}\sum_1^n \log y_i)^2}, & \lambda = 0, \end{cases} \tag{3.24}
$$

where a_i's are the coefficients of the W-test for normality (Shapiro and Wilk, 1965) and $\sum_1^n a_i = 0$.

There are two proposals for finding an estimate from the given observations. The first is to plot the $W(\lambda)$ from (3.24) against λ for a trial series of values. From this plot, the value which maximizes $W(\lambda)$ is the estimate of λ. The second is to solve the nonlinear equation by letting the first derivative of $W(\lambda)$ with respect to λ equal zero. Thus, for $\lambda \neq 0$

$$
\sum_1^n (a_i y_{(i)}^{\hat{\lambda}} \log y_{(i)}) \cdot \sum_1^n (y_i^{\hat{\lambda}} - \frac{1}{n}\sum_1^n y_i^{\hat{\lambda}})
$$

$$
- \sum_1^n (a_i y_{(i)}^{\hat{\lambda}}) \cdot \sum_1^n (y_i^{\hat{\lambda}} \log y_i - \frac{1}{n}\sum_1^n y_i^{\hat{\lambda}} \log y_i) = 0, \tag{3.25}
$$

and the estimate is zero if the following identity is satisfied

$$
\sum_1^n (a_i \log^2 y_{(i)}) \cdot \sum_1^n (\log y_i - \frac{1}{n}\sum_1^n \log y_i)
$$

$$
- \sum_1^n (a_i \log y_{(i)}) \cdot \sum_1^n (\log^2 y_i - \frac{1}{n}\sum_1^n \log^2 y_i) = 0. \tag{3.26}
$$

Because there is no closed form for the solution, we need to use numerical iteration method to find the estimate $\hat{\lambda}$.

Properties

1. This estimate is scale free.

Because

$$W(\lambda \mid cy_1, cy_2, \ldots, cy_n) = \frac{\left[\sum_1^n a_i (cy_{(i)})^{\lambda}\right]^2}{\sum_1^n \left[(cy_i)^{\lambda} - \frac{1}{n}\sum_1^n (cy_i)^{\lambda}\right]^2}$$

$$= \frac{c^2\left[\sum_1^n a_i y_{(i)}^{\lambda}\right]^2}{c^2 \sum_1^n \left[y_i^{\lambda} - \frac{1}{n}\sum_1^n y_i^{\lambda}\right]^2} = W(\lambda \mid y_1, y_2, \ldots, y_n),$$

the value which maximizes $W(\lambda \mid cy_1, cy_2, \ldots, cy_n)$ will also maximize $W(\lambda \mid y_1, y_2, \ldots, y_n)$.

2. The $100(1-\alpha)$ per cent confidence interval (λ_1, λ_2) for λ can be obtained from

$$W(\lambda_1) = W(\lambda_2) = W_\alpha, \tag{3.27}$$

where W_α is the α percentage point of the Shapiro and Wilk (1965) W-test for normality.

3. We can find the observed significance level for testing normality of the transformed observations by the corresponding percentage of the Shapiro and Wilk (1965) W-test statistics.

Example

We again use the same data as presented in Table II. Table IV shows $W(\lambda)$ calculated over the interesting ranges of λ and the results are plotted in Figure 9. The optimal value for the power transformation parameter is $\hat{\lambda}=-0.65$. The critical point of 0.05 percentage of Shapiro and Wilk's (1965) normality test is 0.947. Using (3.26), the curve of $W(\lambda)$ gives a 95% confidence interval of λ extending from about -1.15 to -0.22.

TABLE IV

CALCULATIONS OF $W(\lambda)$ BASED ON THE NORMALITY OF TRANSFORMED OBSERVATIONS

λ	$W(\lambda)$	λ	$W(\lambda)$	λ	$W(\lambda)$
-4.0	0.4385	-0.8	0.9821	-0.3	0.9600
-3.0	0.5967	-0.75	0.9840	-0.22	0.9486
-2.5	0.6944	-0.7	0.9852	-0.1	0.9241
-2.0	0.7978	-0.65	0.9856	0.0	0.8988
-1.5	0.8954	-0.6	0.9850	0.25	0.8185
-1.3	0.9287	-0.55	0.9836	0.5	0.7218
-1.15	0.9500	-0.5	0.9812	1.0	0.5297
-1.1	0.9563	-0.45	0.9779	1.5	0.3917
-1.0	0.9672	-0.4	0.9735	2.0	0.3113
-0.9	0.9759	-0.35	0.9680	3.0	0.2433

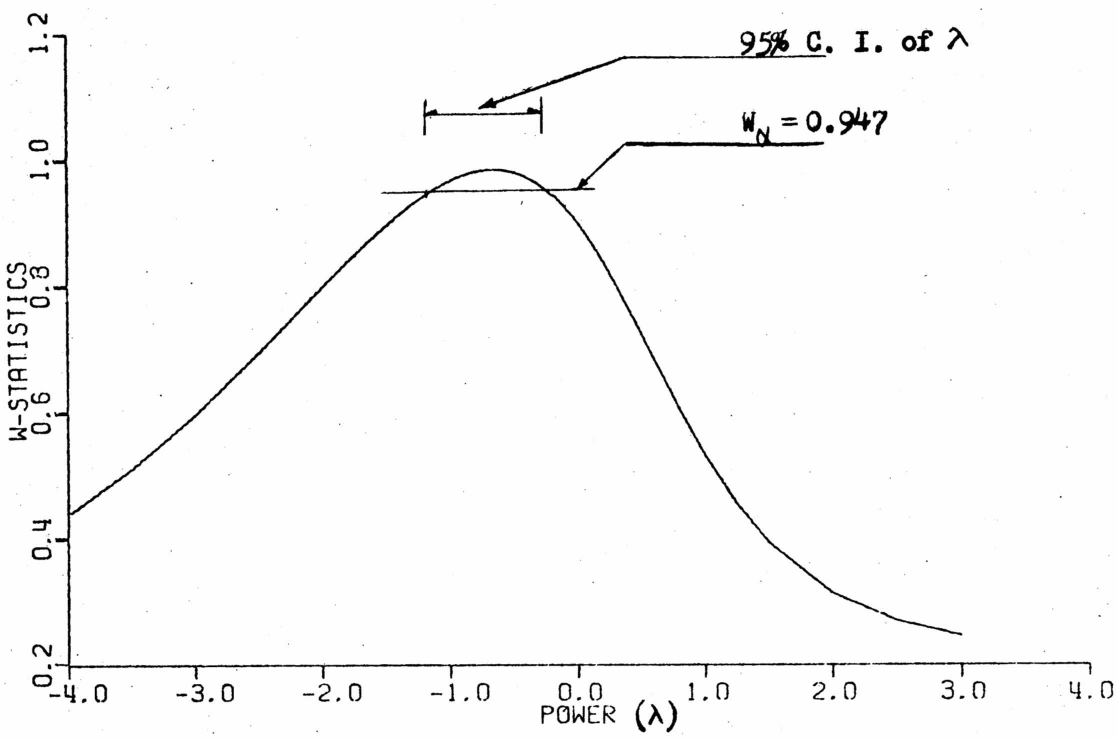

Figure 9. Function W(λ) Against λ of Random Sample

Comparison of Maximum W-Statistic Estimate With Maximum Likelihood Estimate

1. The maximum W-statistic method requires $6n+1$ operations while the maximum likelihood method requires $6n+7$ operations in the calculations for each given λ. Both have formulas to find the $100(1-\alpha)$ per ent confidence region of λ.

2. It is easier to find the observed significance level of the transformation to normality for the maximum W-statistic method.

3. A set of 1,000 random samples were generated for five different values of λ and two sample sizes (25, 50). The procedure was to

generate the observations y_i's such that $(y_i^{\lambda} - 1)/\lambda \backsim N(0, 0.1^2)$ independently for $i = 1, 2, \ldots, n$ with $n = 25, 50$ in each sample. We estimated the transformation parameter λ using both methods, and the results are shown in Table V.

TABLE V

SUMMARY OF COMPARISON BETWEEN MAXIMUM W-STATISTIC ESTIMATE AND MAXIMUM LIKELIHOOD ESTIMATE

Sample Size		Method	λ=3.0	λ=1.0	λ=0.0	λ=-1.0	λ=-3.0
50	$\hat{\lambda}$	M.W.E.	2.813	0.994	0.014	-0.970	-2.789
		M.L.E.	2.732	0.973	0.042	-0.960	-2.707
	Bias	M.W.E.	-0.187	-0.006	0.014	0.030	0.211
		M.L.E.	-0.268	-0.027	0.042	0.040	0.293
	$s^2(\hat{\lambda})$	M.W.E.	1.358	1.614	1.601	1.687	1.350
		M.L.E.	1.145	1.453	1.475	1.546	1.116
25	$\hat{\lambda}$	M.W.E.	2.779	0.894	-0.144	-0.927	-2.825
		M.L.E.	2.541	0.842	-0.173	-0.984	-2.594
	Bias	M.W.E.	-0.221	-0.106	-0.144	0.079	0.175
		M.L.E.	-0.459	-0.158	-0.173	0.016	0.406
	$s^2(\hat{\lambda})$	M.W.E.	3.794	4.143	3.948	4.629	3.859
		M.L.E.	2.793	3.481	3.255	3.666	2.996

From the above results, we find the maximum W-statistic estimate has smaller bias but larger variance and that the bias increases and the sample variance of both estimates decreases as the absolute value of λ increases.

CHAPTER IV

POWER TRANSFORMATION TO A TRUNCATED NORMAL DISTRIBUTION

To avoid the inadmissibility of λ of the power transformation, we must assume that the observations are positive. This assumption will cause the range of the transformed observations $y_i^{(\lambda)}$'s to be $(-\infty, -\frac{1}{\lambda})$ for $\lambda < 0$ and $(-\frac{1}{\lambda}, \infty)$ for $\lambda > 0$. It is clear that there is no λ, except $\lambda = 0$, such that the transformed observations have a normal distribution. That is, for $\lambda \neq 0$, we only can transform the positive random variable to a truncated normal distribution, if one exists. But in practice, we would like to find a transformation for observations from an unknown distribution such that the transformed observations are distributed approximately normally. If the truncation error is small enough, then the truncated normal distribution may be considered as a good approximate normal distribution.

In the maximum likelihood method of finding the estimate of λ of Box and Cox (1964), the incorrect likelihood function of a full normal distribution was used. Nevertheless, this method is adequate sometimes for practical purposes. In other words, if there exists a transformation to a truncated normal distribution with a truncation error near zero, the maximum likelihood estimate of Box and Cox is very nearly correct. But if the truncation error is not close to zero yet is small, for example 0.05, it will produce misleading estimates of the mean, variance, and $\hat{\lambda}$.

This chapter considers a more general and accurate method based on maximizing the exact likelihood function of a truncated normal distribution with a varying truncation error. Hence, we can find a maximum likelihood estimate among all truncated normal distributions and full normal distributions. If the estimate of the truncation error is approximately zero, both methods produce the same estimate of λ. And if the estimate of the truncation error is not small enough, then we will claim that there is no power transformation such that the transformed observations are approximately normally distributed.

The Truncated Normal Distribution

<u>Definition</u>. The random variable X is said to have a truncated normal distribution with parameters μ and σ^2 if and only if its density function is given by

$$f_X(x) = \begin{cases} \dfrac{1}{\sqrt{2\pi}\sigma \displaystyle\int_a^b \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt} e^{-\frac{(x-\mu)^2}{2\sigma^2}} & \text{for } a < x < b, \\ 0 & \text{otherwise.} \end{cases} \tag{4.1}$$

Note 1. We use the symbol $X \sim N^*(\mu, \sigma^2)$ for $a < x < b$ to denote that X has a truncated normal distribution.

Note 2. For convenience, we let

$$g(t) = \frac{1}{\sqrt{2\pi}\ \sigma} e^{-\frac{(t-\mu)^2}{2\sigma^2}},$$

$$\alpha = \int_{-\infty}^{a} g(t)\ dt,$$

$$\beta = \int_{b}^{\infty} g(t)\ dt,$$

where α and β are called left and right truncation errors, respectively.

From this definition, we have the following lemmas (the proofs are shown in Appendix B).

Lemma 1. If $X \sim N^*(\mu, \sigma^2)$ for $a < x < b$ and $Z = (X-\mu)/\sigma$, then the random variable $Z \sim N^*(0, 1)$ for $(a-\mu)/\sigma < z < (b-\mu)/\sigma$.

Lemma 2. Suppose $X \sim N^*(0, 1)$ for $a < x < \infty$, then $F(x) = (\Phi(x) - \alpha)/(1-\alpha)$; and if $F(x) = k$, then the k quantile $x_k = F^{-1}(k) = \Phi^{-1}(k - k\alpha + \alpha)$, where $\Phi(t)$ is the distribution function of standard normal distribution.

Lemma 3. Suppose $X \sim N^*(0, 1)$ for $-\infty < x < b$, then $F(x) = \Phi(x)/(1-\beta)$, and for $F(x) = k$ the k quantile $x_k = F^{-1}(k) = \Phi^{-1}(k - k\beta)$.

Lemma 4. Suppose random variable $X \sim N^*(\mu, \sigma^2)$ for $a < x < b$, then

$$E(X) = \mu - \frac{g(b) - g(a)}{1 - \alpha - \beta}\ \sigma^2. \tag{4.2}$$

Lemma 5. Let $X_1, X_2, \ldots, X_n$ be a random sample from $N^*(\mu, \sigma^2)$ for $a < x_i < b$, where $i = 1, 2, \ldots, n$. The maximum likelihood estimate of μ and σ^2 are the solutions of the following two equations

$$\hat{\mu} = \frac{1}{n}\sum_{1}^{n} x_i + \frac{1}{1-\hat{\alpha}-\hat{\beta}}\hat{\sigma}^2\ (\hat{g}(b) - \hat{g}(a)), \tag{4.3}$$

$$\hat{\sigma}^2 = \frac{\sum_1^n (x_i - \hat{\mu})^2/n}{1 - \frac{(b-\hat{\mu})\hat{g}(b) - (a-\hat{\mu})\hat{g}(a)}{1-\hat{\alpha}-\hat{\beta}}}, \tag{4.4}$$

where

$$\hat{g}(t) = \frac{1}{\sqrt{2\pi}\,\hat{\sigma}} e^{-\frac{(t-\hat{\mu})^2}{2\hat{\sigma}^2}},$$

$$\hat{\alpha} = \int_{-\infty}^{a} \hat{g}(t)\, dt,$$

$$\hat{\beta} = \int_{b}^{\infty} \hat{g}(t)\, dt.$$

Note 3. There is no closed form for the solutions of $\hat{\mu}$ and $\hat{\sigma}^2$, but it is quite simple to obtain them by a numerical iteration method with the initial guesses as

$$\hat{\mu}(1) = \bar{x},$$

$$\hat{\sigma}^2(1) = \sum_1^n (x_i - \hat{\mu}(1))^2/ n.$$

Existence of Power Transformation to a Truncated Normal Distribution

Since the range of the transformed variable $Y^{(\lambda)}$ is bounded by one side for $\lambda \neq 0$, we only consider transformations to one-side truncated normal distributions. Suppose there exists a power transformation with $\lambda \neq 0$ such that the transformed variable $Y^{(\lambda)}$ is distributed $N^*(\mu, \sigma^2)$ with a truncation error α, then the probability density function of the original variable Y is

$$f_Y(y) = \begin{cases} 0 & \text{for } y \leqslant 0 \end{cases} \tag{4.5}$$

$$\left\{ \frac{y^{\lambda-1}}{\sqrt{2\pi}\,\sigma(1-\alpha)} e^{-\frac{\left(\frac{y^{\lambda}-1}{\lambda}-\mu\right)^2}{2\sigma^2}} \right. \qquad \text{for } y > 0,$$

where

$$\alpha = \int_{-\infty}^{-1/\lambda} \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt \qquad \text{for } \lambda > 0,$$

$$\alpha = \int_{-1/\lambda}^{\infty} \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt \qquad \text{for } \lambda < 0.$$

Theorem. There is only one distribution with $y > 0$ that can be transformed to a normal distribution by the power transformation, namely the log normal distribution with transformation parameter $\lambda=0$.

Proof. For $y > 0$, there is only one power transformation with the range $(-\infty, \infty)$. That is $Y^{(\lambda)} = \log Y$ for $\lambda = 0$. If $Y^{(\lambda)}$ is distributed normally, then the random variable Y has a log normal distribution.

Theorem. The family of gamma distribution with the parameter $(\frac{1}{2}, \gamma)$ can be transformed to a truncated normal distribution with truncation error 0.5 by the power transformation with $\lambda = \frac{1}{2}$.

Proof. Let random variable $Y \sim Ga\ (\frac{1}{2}, \gamma)$, i.e., the density function of Y is

$$f_Y(y) = \begin{cases} \dfrac{1}{\Gamma(\frac{1}{2})\,\gamma^{\frac{1}{2}}}\, y^{\frac{1}{2}-1}\, e^{-\frac{y}{\gamma}} & \text{for } y > 0, \\ 0 & \text{for } y \leqslant 0. \end{cases}$$

We can rewrite it as

$$f_Y(y)=\begin{cases} \dfrac{y^{\frac{1}{2}-1}}{\sqrt{2\pi}\,(\sqrt{2}\,\gamma^{\frac{1}{2}})(1-\frac{1}{2})}\, e^{-\dfrac{(\frac{y^{\frac{1}{2}}-1}{\frac{1}{2}}+2)^2}{2\,(\sqrt{2\gamma}\,)^2}} & \text{for } y>0, \\ 0 & \text{for } y\leq 0. \end{cases}$$

Now compare this to (4.5), where we see that the transformed variable $Y^{(\lambda)}$ has a truncated normal distribution with $\mu=-2$, $\sigma=\sqrt{2\gamma}$, and truncation error 0.5.

There are an infinite number of distributions with positive support which can be transformed to a truncated normal distribution with parameters μ, σ^2, and truncation error d. Some of them are shown in Figure 10 through Figure 21 with truncation errors (0.01, 0.05, 0.1) and standard deviations (0.5, 1.0, 1.5, 2.0). Each figure has several plots of the density function which can be transformed to a truncated normal distribution from the power transformation with the value of λ indicated. The parameter μ can be determined correspondingly by

$$\begin{cases} \mu = -\frac{1}{\lambda} - \sigma\phi^{-1}(d) & \text{for } \lambda > 0, \\ \mu = -\frac{1}{\lambda} - \sigma\phi^{-1}(1-d) & \text{for } \lambda < 0, \end{cases} \tag{4.6}$$

where ϕ is the distribution of standard normal and d is the truncation error.

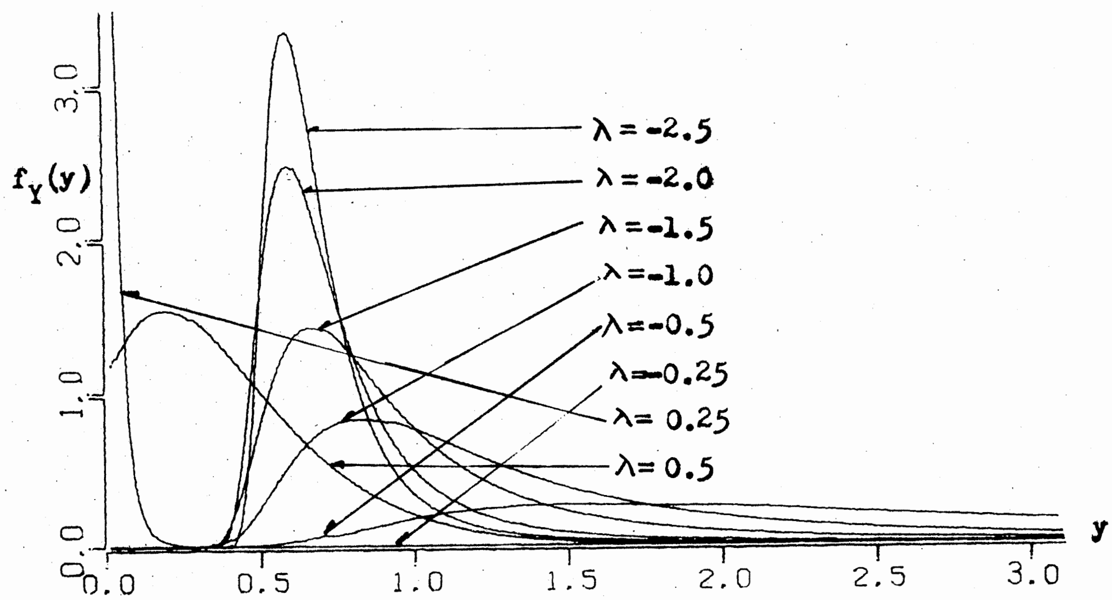

Figure 10. Density Function of Original Variable (α=0.01, σ=0.5)

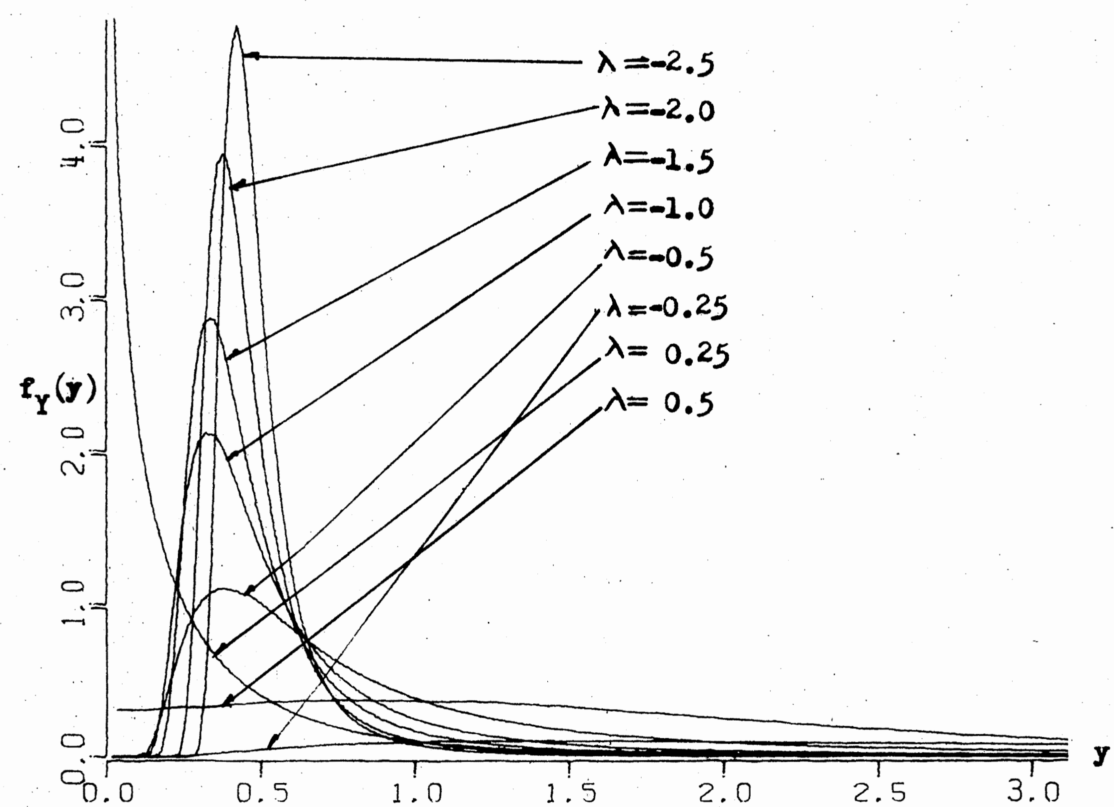

Figure 11. Density Function of Original Variable (α=0.01, σ=1.0)

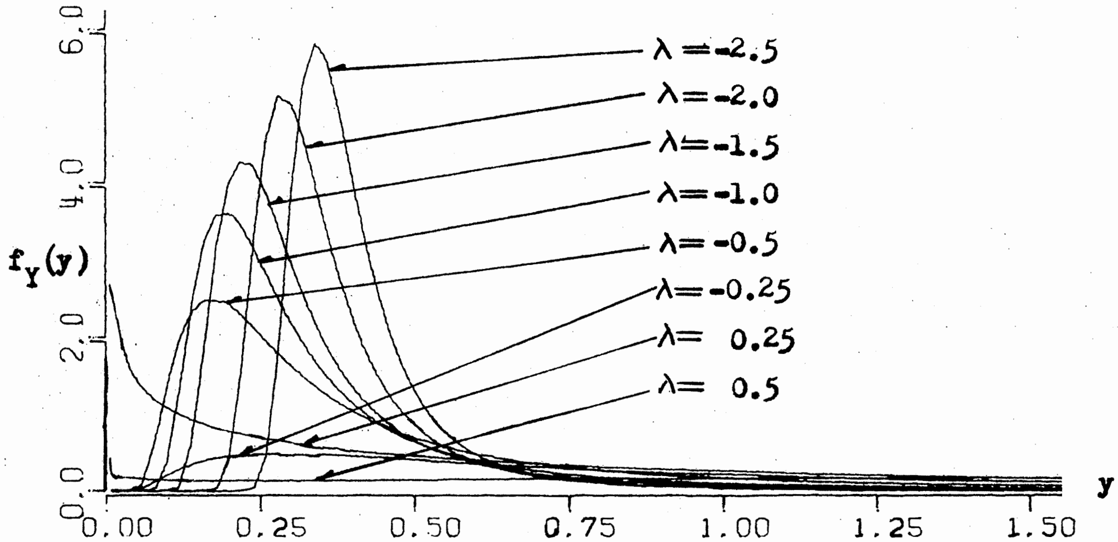

Figure 12. Density Function of Original Variable ($\alpha = 0.01$, $\sigma = 1.5$)

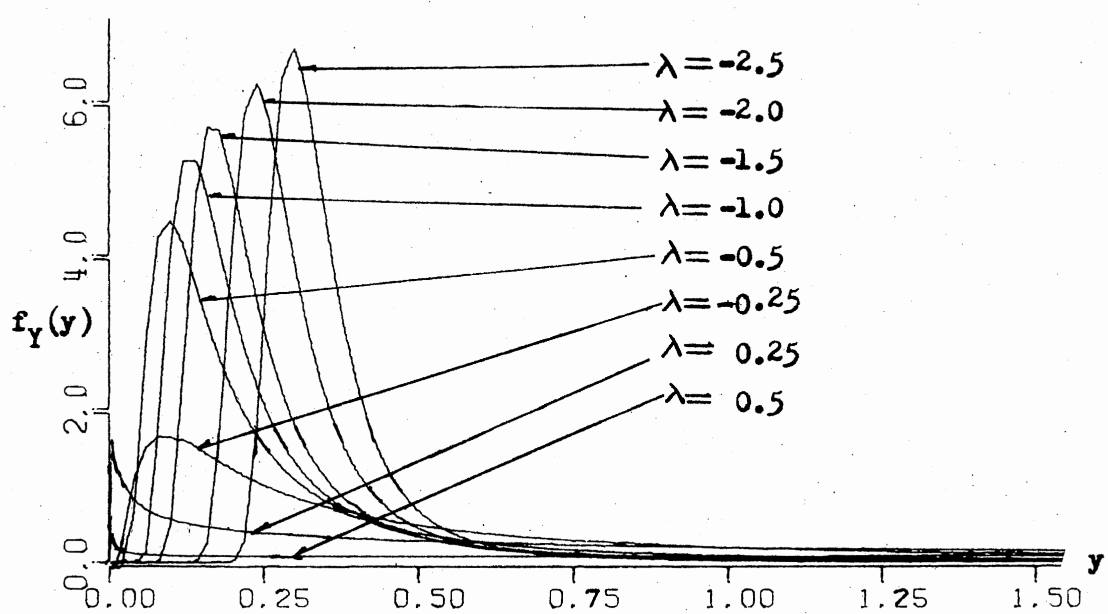

Figure 13. Density Function of Original Variable ($\alpha = 0.01$, $\sigma = 2.0$)

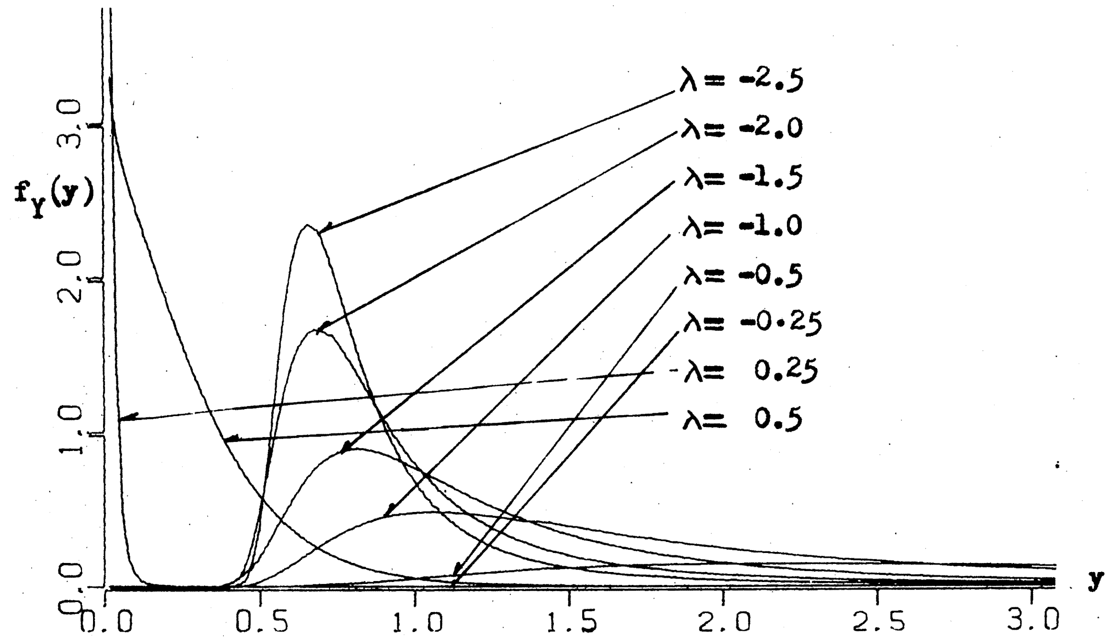

Figure 14. Density Function of Original Variable (α=0.05, σ=0.5)

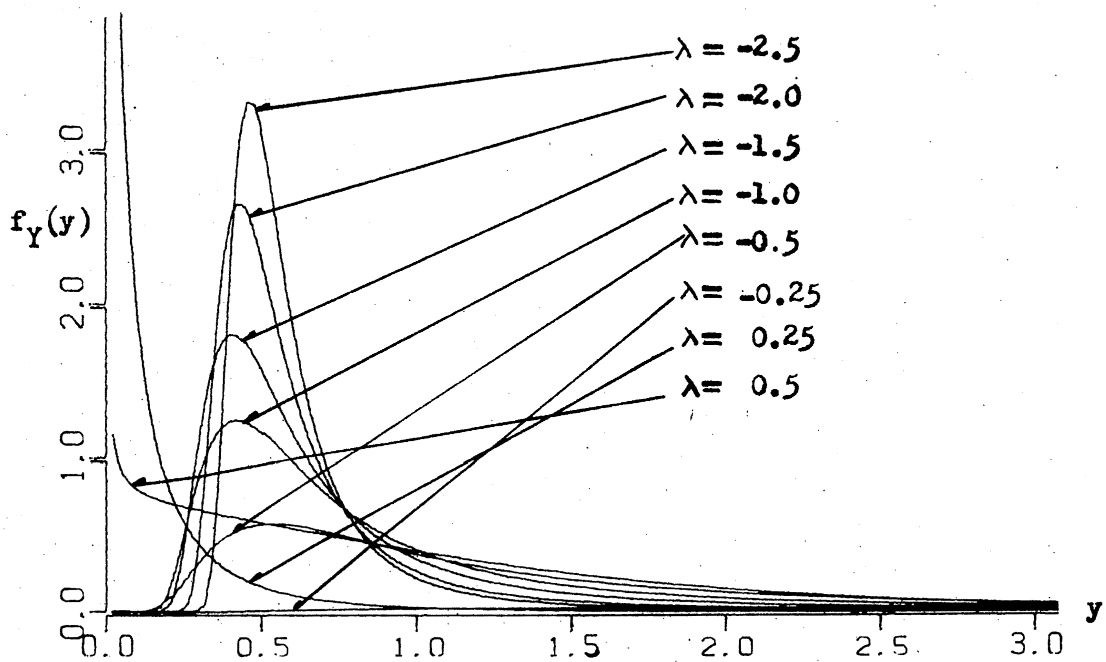

Figure 15. Density Function of Original Variable (α=0.05, σ=1.0)

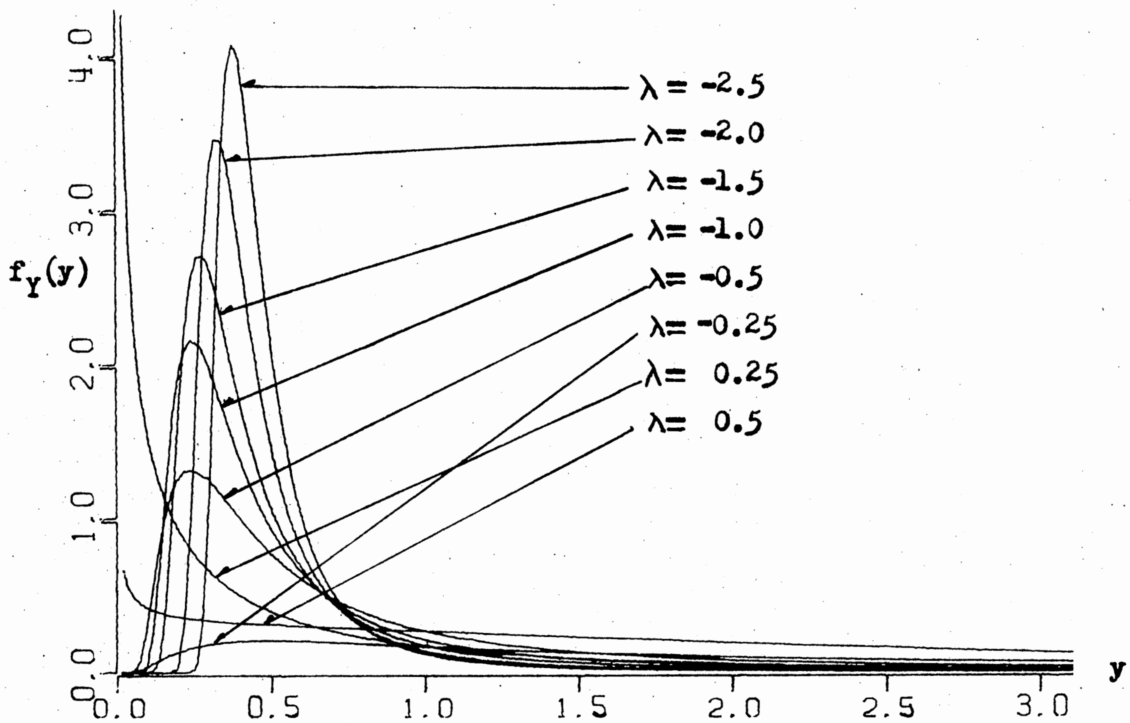

Figure 16. Density Function of Original Variable (α=0.05, σ=1.5)

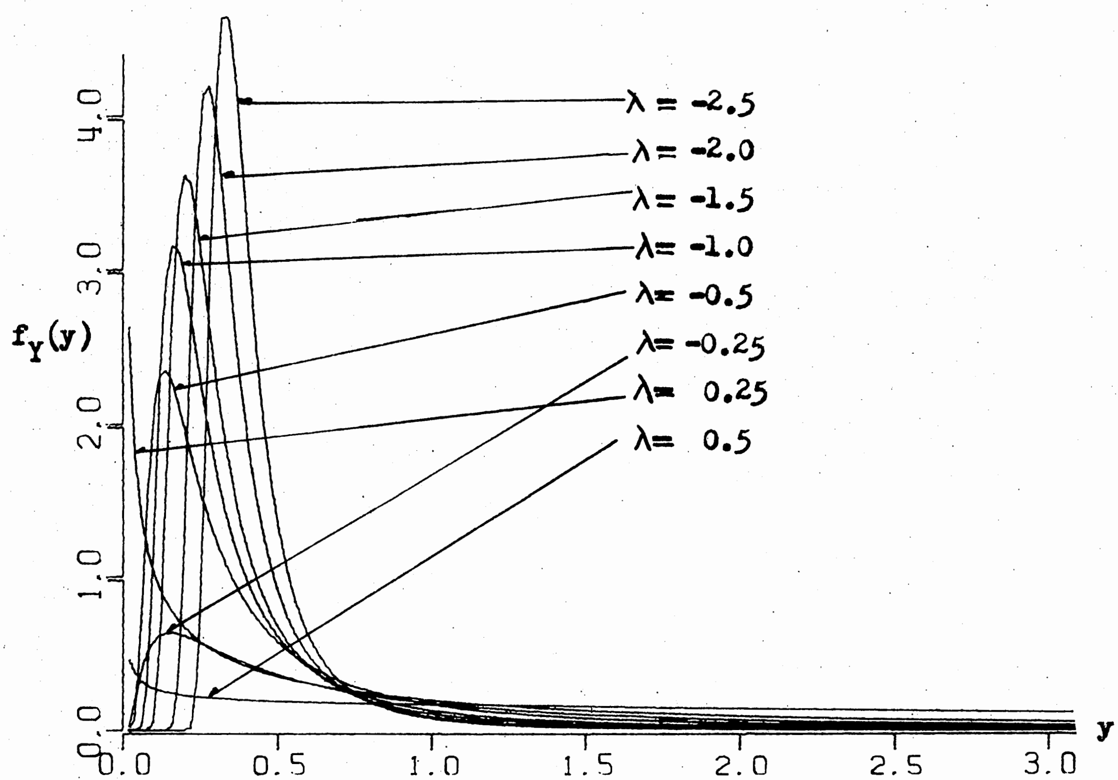

Figure 17. Density Function of Original Variable (α=0.05, σ=2.0)

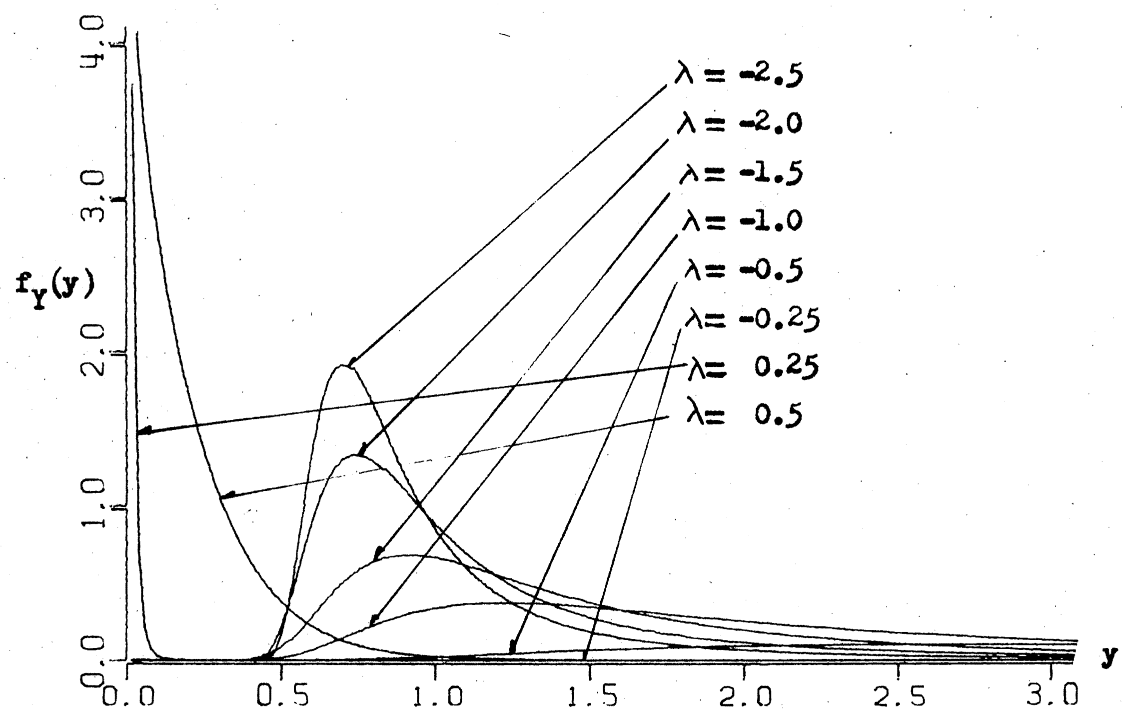

Figure 18. Density Function of Original Variable (α=0.1, σ=0.5)

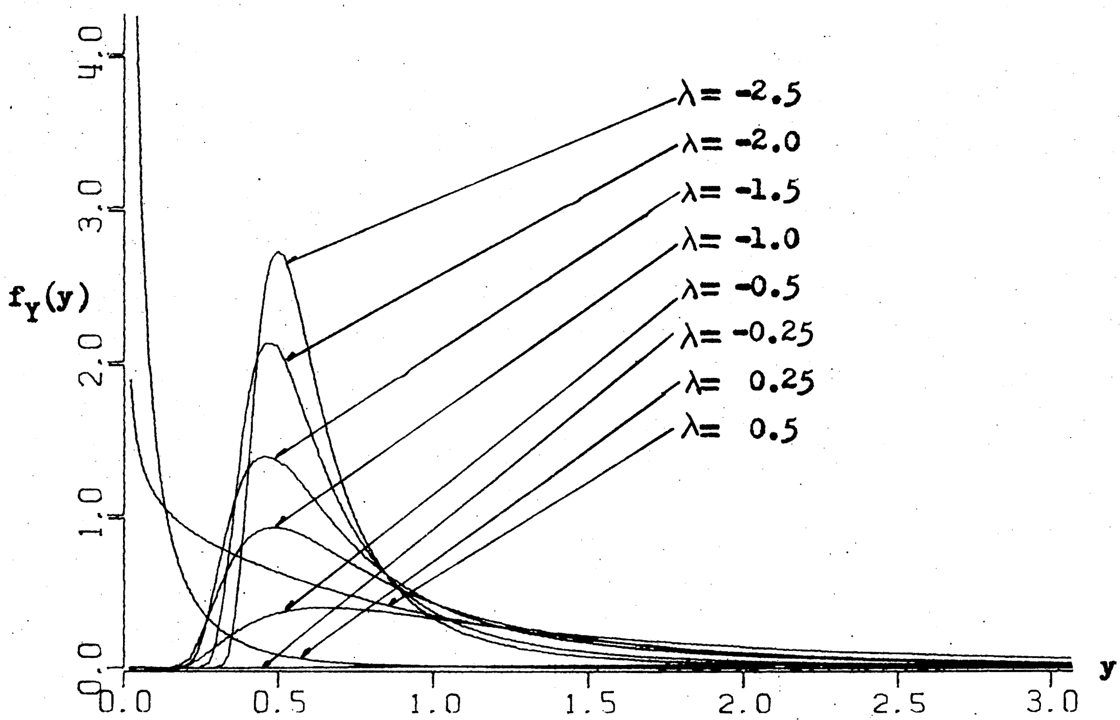

Figure 19. Density Function of Original Variable (α=0.1, σ=1.0)

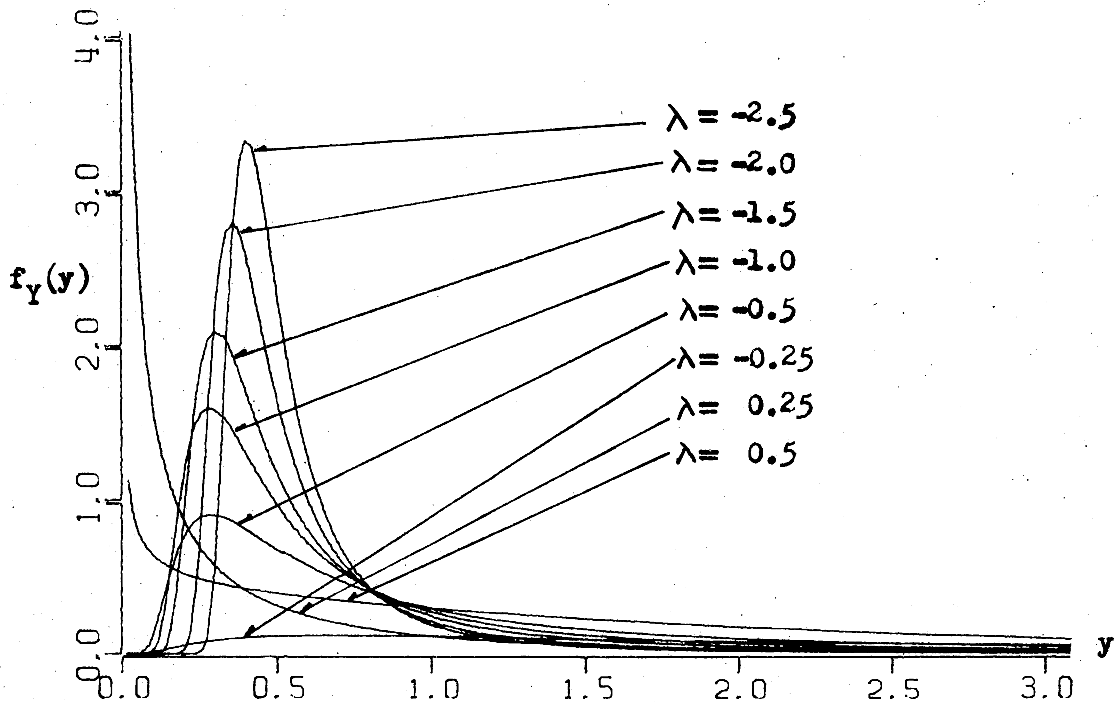

Figure 20. Density Function of Original Variable (α =0.1, σ=1.5)

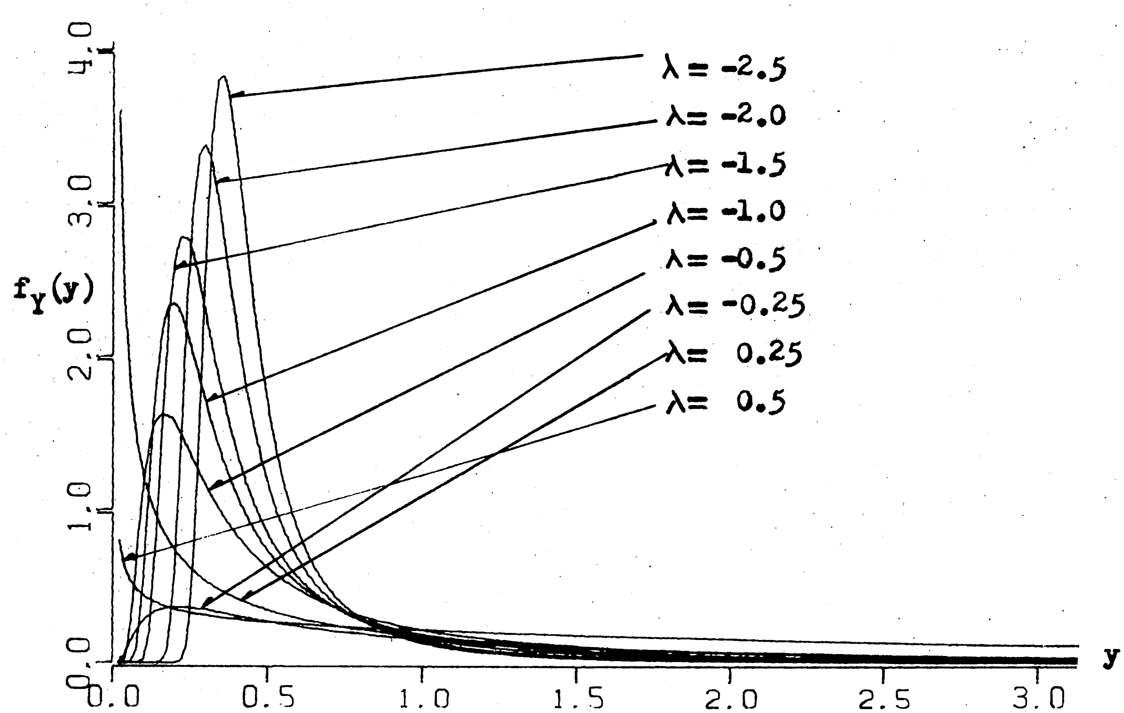

Figure 21. Density Function of Original Variable (α =0.1, σ=2.0)

Maximum Likelihood Estimation

Suppose we observe an n x 1 vector of observations $\underline{y} = [y_1, y_2, \ldots, y_n]$ and the appropriate linear model for the problem is specified by

$$\underline{z} = A\underline{\theta} + \underline{e}, \tag{4.7}$$

where $\underline{z}$ is the n x 1 vector of transformed observations, A is an n x p known matrix, $\underline{\theta}$ is a p x 1 vector of unknown parameters associated with the transformed observations, and $\underline{e}$ is an n x 1 vector of random error. Since, for positive observations y_i's, the transformed observations z_i's are bounded by one side, we only can assume that for some unknown λ the random errors e_i $(i = 1,2,\ldots,n)$ are independently distributed $N^*(0, \sigma^2)$ associated with an unknown truncation error α. The probability density for the untransformed observations is obtained by multiplying the truncated normal density by the Jacobian of the transformation.

We find the maximum likelihood estimate in two steps. First, for given λ, find the maximum likelihood estimates of μ, σ^2, and α. Then, we plot the maximum likelihood function against λ for a trial series of values and inspect the value of λ which maximizes the maximum likelihood function to be the estimate of λ. We now discuss the three cases $\lambda > 0$, $\lambda = 0$, and $\lambda < 0$.

1. $\lambda > 0$

Since $y_i > 0$ $(i = 1,2,\ldots,n)$, the transformed observations z_i's $(i = 1,2,\ldots,n)$ are greater than $-1/\lambda$ and the random error e_i $(i = 1,2,\ldots, n)$ will greater than $-1/\lambda$ - minimum of $\{\mu_i\}$, where μ_i is the ith element of $A\underline{\theta}$. The likelihood function of the original observations $\underline{y}$ is

$$L(\lambda) = \frac{J(\lambda : \underline{y})}{(2\pi)^{n/2}\,\sigma^n (1-\alpha)^n}\, e^{-\frac{\left[\frac{1}{\lambda}(\underline{y}^{\lambda} - \underline{1}) - A\underline{\theta}\right]'\left[\frac{1}{\lambda}(\underline{y}^{\lambda} - \underline{1}) - A\underline{\theta}\right]}{2\sigma^2}}, \qquad (4.8)$$

where $\underline{1}$ is an n x 1 vector with elements of 1,

$$\alpha = \int_{-\infty}^{-\frac{1}{\lambda} - \text{min. of}\{\mu_i\}} \frac{1}{\sqrt{2\pi}\,\sigma}\, e^{-\frac{t^2}{2\sigma^2}}\, dt,$$

and

$$J(\lambda : \underline{y}) = \prod_{1}^{n} \left| \frac{dz_i}{dy_i} \right| = \prod_{1}^{n} y_i^{\lambda - 1}.$$

By letting the first partial derivative of log $L(\lambda)$ with respect to $\underline{\theta}$ and σ^2 be zero, the maximum likelihood estimates of $\underline{\theta}$ and σ^2 for given $\lambda > 0$ are the solutions of the following equations (Appendix C),

$$\begin{aligned} \hat{\underline{\theta}} &= (A'A)^{-1}\left[A'\frac{1}{\lambda}(\underline{y}^{\lambda} - \underline{1}) - \frac{n\hat{\sigma}^2}{1-\hat{\alpha}}\, \hat{g}_k\left(-\frac{1}{\lambda}\right) \underline{a}_k\right], \\ \hat{\sigma}^2 &= \frac{\left[\frac{1}{\lambda}(\underline{y}^{\lambda} - \underline{1}) - A\hat{\underline{\theta}}\right]'\left[\frac{1}{\lambda}(\underline{y}^{\lambda} - \underline{1}) - A\hat{\underline{\theta}}\right]}{n\left[1 + \frac{\hat{g}_k(-\frac{1}{\lambda})}{1-\hat{\alpha}}\left(-\frac{1}{\lambda} - \hat{\mu}_k\right)\right]}, \end{aligned} \qquad (4.9)$$

where $\hat{\mu}_k$ is the estimate of minimum of $\{\mu_i\}$, $\underline{a}_k$ is the p x 1 vector of the kth row of A,

$$\hat{g}_k\left(-\frac{1}{\lambda}\right) = \frac{1}{\sqrt{2\pi}\,\hat{\sigma}}\, e^{-\frac{(-\frac{1}{\lambda} - \hat{\mu}_k)^2}{2\hat{\sigma}^2}},$$

and

$$\hat{\alpha} = \int_{-\infty}^{-\frac{1}{\lambda}-\hat{\mu}_k} \frac{1}{\sqrt{2\pi}\,\hat{\sigma}} e^{-\frac{t^2}{2\hat{\sigma}^2}} dt.$$

Then the maximized log likelihood function is

$$\log L_{max}(\lambda) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log \hat{\sigma}^2 - n\log(1-\hat{\alpha})$$

$$-\frac{n}{2}\left[1 + \frac{(-\frac{1}{\lambda}-\hat{\mu}_k)}{1-\hat{\alpha}}\,\hat{g}_k(-\frac{1}{\lambda})\right] + (\lambda-1)\sum_1^n \log y_i. \qquad (4.10)$$

2. $\lambda = 0$

The likelihood function of the original observations $\underline{y}$ is

$$L(\lambda) = \frac{\prod_1^n y_i^{-1}}{(2\pi)^{n/2}\sigma^n} e^{-\frac{\left[\log \underline{y} - A\underline{\theta}\right]'\left[\log \underline{y} - A\underline{\theta}\right]}{2\sigma^2}}. \qquad (4.11)$$

The maximum likelihood estimates of $\underline{\theta}$ and σ^2 are

$$\hat{\underline{\theta}} = (A'A)^{-1}A'\,\underline{\log y},$$
$$\hat{\sigma}^2 = \frac{(\underline{\log y} - A\hat{\underline{\theta}})'(\underline{\log y} - A\hat{\underline{\theta}})}{n}. \qquad (4.12)$$

Then, the maximized log likelihood function is

$$\log L_{max}(\lambda) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log \hat{\sigma}^2 - \frac{n}{2} - \sum_1^n \log y_i. \qquad (4.13)$$

3. $\lambda < 0$

For $y_i > 0$ $(i = 1,2,\ldots,n)$, the transformed observations are less than $-1/\lambda$ and the random errors e_i $(i = 1,2,\ldots,n)$ are less than $-1/\lambda$

- maximum of $\{\mu_i\}$. The likelihood function of the original observations $\underline{y}$ is

$$L(\lambda) = \frac{\prod_{1}^{n} y_i^{\lambda-1}}{(2\pi)^{n/2}\sigma^n(1-\alpha)^n} e^{-\frac{\left[\frac{1}{\lambda}(\underline{y}^{\lambda}-\underline{1})-A\underline{\theta}\right]'\left[\frac{1}{\lambda}(\underline{y}^{\lambda}-\underline{1})-A\underline{\theta}\right]}{2\sigma^2}}, \qquad (4.14)$$

where

$$\alpha = \int_{-\frac{1}{\lambda}-\text{max. of }\{\mu_i\}}^{\infty} \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{t^2}{2\sigma^2}} dt.$$

The maximum likelihood estimates of $\underline{\theta}$ and σ^2 are the solutions of the following equations (Appendix C)

$$\hat{\underline{\theta}} = (A'A)^{-1}\left[A'\frac{1}{\lambda}(\underline{y}^{\lambda}-\underline{1}) + \frac{n\hat{\sigma}^2}{1-\hat{\alpha}}\hat{g}_k(-\tfrac{1}{\lambda})\,\underline{a}_k\right],$$

$$\hat{\sigma}^2 = \frac{\left[\frac{1}{\lambda}(\underline{y}^{\lambda}-\underline{1}) - A\hat{\underline{\theta}}\right]'\left[\frac{1}{\lambda}(\underline{y}^{\lambda}-\underline{1}) - A\hat{\underline{\theta}}\right]}{n\left[1 - \frac{\hat{g}_k(-\frac{1}{\lambda})}{1-\hat{\alpha}}(-\frac{1}{\lambda}-\hat{\mu}_k)\right]}, \qquad (4.15)$$

where $\hat{\mu}_k$ is the estimate of maximum of $\{\mu_i\}$, $\underline{a}_k$ is the p x 1 vector of the kth row of A,

$$\hat{g}_k(-\tfrac{1}{\lambda}) = \frac{1}{\sqrt{2\pi}\,\hat{\sigma}} e^{-\frac{(-\frac{1}{\lambda}-\hat{\mu}_k)^2}{2\hat{\sigma}^2}},$$

and

$$\hat{\alpha} = \int_{\frac{-1}{\lambda} - \hat{\mu}_k}^{\infty} \frac{1}{\sqrt{2\pi}\,\hat{\sigma}} e^{-\frac{t^2}{2\hat{\sigma}^2}} dt.$$

Then, the maximized log likelihood function is

$$\log L_{max}(\lambda) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log \hat{\sigma}^2 - n\log(1-\hat{\alpha})$$

$$-\frac{n}{2}\left[1+\frac{(-\frac{1}{\lambda}-\hat{\mu}_k)}{1-\hat{\alpha}}\,\hat{g}_k(-\frac{1}{\lambda})\right] + (\lambda-1)\sum_1^n \log y_i. \qquad (4.16)$$

If we are considering the univariate case, i.e., A is an n x 1 vector of elements 1, then the $\hat{\mu}_k$ will be $\hat{\mu}$. There is no closed form of $\hat{\beta}$ and $\hat{\sigma}^2$ for $\lambda > 0$ and $\lambda < 0$, but they can be determined numerically. If the $\hat{g}_k(-\frac{1}{\lambda})$ is close to zero for a given λ, then $\hat{\beta}$ and $\hat{\sigma}^2$ will be the least squares estimates.

Example

Table VI shows data taken from page 339 of Steel and Torrie (1960) which will be used for illustrative purposes.

We assume that the power transformed observations have a simple regression model

$$y_i^{(\lambda)} = \frac{y_i^{\lambda} - 1}{\lambda} = b_o + b_1\, x_i + e_i. \qquad (4.17)$$

If the transformed observations have a truncated normal distribution, then the random errors e_i's are independently distributed $N^*(0, \sigma^2)$ and have a truncation error α.

TABLE VI

ALASKA PEAS GROWN AT MADISON, WISCONSIN, 1953

y:	24.0	22.0	26.5	22.0	25.0	37.5	36.0	39.5	32.0
x:	76.2	76.8	77.3	79.2	80.0	87.8	93.2	93.5	94.3
y:	26.5	55.5	49.5	56.0	55.5	58.0	61.5	69.0	71.5
x:	96.8	97.5	99.5	104.2	106.3	106.7	119.0	119.7	119.8
y:	73.0	76.5	78.5	74.0	71.5	77.0	85.5		
x:	119.8	123.5	141.0	142.3	145.5	149.0	150.0		

y — yield in pounds per plot

x — tenderometer reading

After applying the proposed method, the maximum likelihood estimates are

$$\hat{\lambda} = 1.29,$$

$$\hat{b}_0 = -150.40,$$

$$\hat{b}_1 = 2.59,$$

$$\hat{\sigma}^2 = 552.38,$$

$$\hat{\alpha} = 0.021.$$

The regression line will be

$$\frac{y^{1.29} - 1}{1.29} = -150.40 + 2.59\, x$$

which can be rewritten as

$$y^{1.29} = -195.17 + 3.35\ x. \qquad (4.18)$$

Since the estimate of truncation error is 0.021, we can say that this truncated normal distribution is an approximate normal distribution.

But, if we apply the maximum likelihood method of Box and Cox (1964) with the full normal distribution for the e_i's, the maximum likelihood estimates are

$$\hat{\lambda} = 1.58,$$
$$\hat{b}_0 = -538.44,$$
$$\hat{b}_1 = 8.29,$$
$$\hat{\sigma}^2 = 4674.76,$$

and the regression line is

$$\frac{y^{1.58}-1}{1.58} = -538.44 + 8.29\ x,$$

which can be rewritten as

$$y^{1.58} = -854.42 + 13.14\ x. \qquad (4.19)$$

Therefore, we have two estimates of λ and want to compare their goodness of fit to the assumed distributions by residuals. We computed the maximum distance between the emperical distribution of residuals and the estimated theoretical distribution of random errors. The emperical distribution, $S(x)$, is a step function of x for $-\infty < x < \infty$, where each step is of height $1/n$ and occurs only at the sample values. If we let x_i denote the ith ordered value of residuals, then the emperical distribution is: $S(x) = 0$ for $x < x_1$, $S(x) = i/25$ for $x_i \leqslant x < x_{i-1}$ and

$i = 1, 2, \ldots, 25$, and $S(x) = 1$ for $x \geq x_{25}$. The estimated theoretical distribution of the random error is $N^*(0, 575.899)$ with left truncation error 0.021 for $\hat{\lambda} = 1.29$; and the estimated theoretical distribution of the random error is $N(0, 4873.77)$ for $\hat{\lambda} = 1.58$. The estimate of σ^2 are calculated as $\frac{n}{n-1}\hat{\sigma}^2$ for both estimates of λ, where $\hat{\sigma}^2$ is the maximum likelihood estimate of σ^2.

For the power transformation of $\hat{\lambda} = 1.29$, the maximum distance between the emperical distribution of residuals and $N^*(0, 575.899)$ with left truncation error 0.021 is 0.07134, while the maximum distance is 0.09195 for the power transformation of $\hat{\lambda} = 1.58$. Thus, we can see that the author's proposed method, based on the truncated normal distribution, gives a closer fit to the observed data than the maximum likelihood method of Box and Cox (1964).

CHAPTER V

ROBUSTNESS STUDY AND SUMMARY

Robustness Study

The simple regression line is used here as an example to investigate the robustness of the transformation to normality. Two cases are discussed in this chapter.

Case 1. Suppose that we have the true regression model of

$$y_i = b_o + b_1 x_i + e_i \qquad \text{for } i = 1,2,\ldots,n, \qquad (5.1)$$

where y_i is the dependent variable, x_i is the independent variable, and the random errors e_i $(i=1,2,\ldots,n)$ are independently distributed $N(0, \sigma^2)$. If we apply the power transformation to the dependent variable y_i with the normality assumption of random errors, what will happen? The transformed model is

$$\frac{y_i^\lambda - 1}{\lambda} = c_o + c_1 x_i + e_i \qquad \text{for } i = 1,2,\ldots,n. \qquad (5.2)$$

Case 2. Suppose for some value of λ, the true regression model is

$$\frac{y_i^\lambda - 1}{\lambda} = c_o + c_1 x_i + e_i \qquad \text{for } i = 1,2,\ldots,n, \qquad (5.3)$$

where the e_i $(i=1,2,\ldots,n)$ are independently distributed $N(0, \sigma^2)$.

But, suppose we analyze the data and assume that the regression model is

$$y_i = b_o + b_1 x_i + e_i \qquad \text{for } i = 1,2,\ldots,n, \tag{5.4}$$

where e_i's are independently distributed $N(0, \sigma^2)$, what then will happen?

A set with 1,000 samples of 25 observations was generated for each of five values (2, 1, 0, -1, -2) of λ. Suppose that the range of dependent variable y is from 10 to 100, the range of independent variable x is from 1 to 10, and y increases as x increases. It is reasonable that we allow a maximum deviation of nine on y which is 10% of the range of y. The values of x_i's are chosen uniformly between 1 and 10. The standard deviations of the random errors were calculated from that a maximum deviation of y with a value of nine was caused by three standard deviations.

The generating formulas were:

1. For $\lambda = 2$,

$$\frac{y_i^2 - 1}{2} = -500.5 + 550\, x_i + e_i, \tag{5.5}$$

where e_i i.i.d. $N(0, 16.5^2)$.

2. For $\lambda = 1$,

$$y_i = 10\, x_i + e_i, \tag{5.6}$$

where e_i i.i.d. $N(0, 3^2)$.

3. For $\lambda = 0$,

$$\log y_i = \frac{8}{9} \log 10 + \frac{\log 10}{9}\, x_i + e_i, \tag{5.7}$$

where e_i i.i.d. $N(0, 0.03^2)$.

4. For $\lambda = -1$,

$$\frac{y_i^{-1}-1}{-1} = 0.89 + 0.01\ x_i + e_i, \tag{5.8}$$

where e_i i.i.d. $N(0,\ 0.0003^2)$.

5. For $\lambda = -2$,

$$\frac{y_i^{-2}-1}{-2} = 0.49445 + 0.00055\ x_i + e_i, \tag{5.9}$$

where e_i i.i.d. $N(0,\ 0.000003^2)$.

In comparisons between the untransformed vs. transformed model, we deal with two considerations:

1. The average of 1,000 values of mean squares of residuals, $\sum_1^{25} (y_i - \hat{y}_i)^2/25$, for both models.

2. In how many cases is the value of the mean square of residuals of the transformed model less than the corresponding value for the untransformed model?

The results are shown in Table VII. For the case where $\lambda = 1$, 722 of the 1,000 samples have a mean square of residuals of the transformed model which was less than that of the untransformed model. Also, the average of the mean squares of the residuals of the transformed model of 1,000 samples was less than the average of the mean squares of the untransformed model. For the other cases, the transformed model was always better than the untransformed model.

TABLE VII

SUMMARY OF ROBUSTNESS STUDY

λ	-2	-1	0	1	2
a	-2.001	-1.000	0.032	0.982	1.982
b	0.003	0.010	0.023	0.089	0.014
c	0.028	0.413	2.749	8.683	2.829
d	57.479	93.897	50.779	8.957	17.611
e	1,000	1,000	1,000	722	1,000

a. The average of $\hat{\lambda}$ of 1,000 samples for the transformed model.

b. The sample standard deviation of $\hat{\lambda}$ of 1,000 samples for the transformed model.

c. The average of the mean squares of residuals of 1,000 samples for the transformed model.

d. The average of the mean squares of residuals of 1,000 samples for the untransformed model.

e. In how many samples (of 1,000) is the mean squares of residuals of the transformed model less than that of the untransformed model?

Summary

Since the time that Box and Cox (1964) proposed the power transformation to approximate normality, it has been evident that the maximum likelihood method involves a large number of calculations. We presented the quantile estimate method in Chapter III in which a close estimate is possible to obtain even using a desk calculator. The asymptotic distribution of this estimate was shown to be normally distributed. From the example illustrated, when the value of p and q are near 0.5, either the check procedure fails or the estimate deviates too much from the true value for small sample size. We would suggest that the value of p and q should be chosen as far away as possible from 0.5.

The plotting method in Chapter III also simplifies the calculations and it is most convenient for small sample size. Because the estimate is obtained graphically, one can not develop the statistical properties of the estimate.

The last method presented in Chapter III of the power transformation to approximate normality is the maximum W-statistic technique. Although it does not substantially reduce the calculations, it is useful with the observed significance level of normality test of the transformed observations. We tried to develop the approximate variance for this estimate, but did not succeed in doing so.

In Chapter IV we generalize the transformation to truncated normality. In practical cases, we are interested in finding a transformation to normality. As we mentioned before, there is no exact power transformation to normality except where $\lambda=0$. Instead, a transformation to approximste normality can be used. If the truncation error is small, this truncated normal distribution can be treated as an approximate normal distribution. The maximum likelihood method of transformation to a truncated normal distribution is a generalization of the Box and Cox (1964) maximum likelihood method. As proved by the example in Chapter IV, the residuals from fitting the linear model of transformed observations (with the truncated normal distribution) are less than those from applying the Box and Cox maximum likelihood method.

We can not evade the fact that this method requires more calculations to find the maximum likelihood estimates of $\underline{\theta}$ and σ^2; but with the aid of a computer, the increase should not be a major problem. The precision of this estimate is too complicated to determine at present, but with a small truncation error, the precision of the Box and Cox (1964) maximum likelihood estimate could be used. A Bayesian approach toward the power transformation to a truncated normal distribution might be considered.

A SELECTED BIBLIOGRAPHY

Andrews, D. F. "A Note on the Selection of Data Transformation." Biometrika, Vol. 58 (1971), 249-254.

Andrews, D. F., Gnanadesikan, R. and Warner, J. L. "Transformations of Multivariate Data." Biometrika, Vol. 27 (1971), 825-840

Atkinson, A. C. "Testing Transformations to Normality." Journal of the Royal Statistical Soceity, Series B, Vol. 35 (1973), 473-479.

Box, G. E. P. and Cox, D. R. "An Analysis of Transformations (with discussion)." Journal of the Royal Statistical Soceity, Series B, Vol. 26 (1964), 211-252.

David, H. A. Order Statistics. New York: Wiley, 1970.

Draper, N. R. and Cox, D. R. "On Distribution and Their Transformation to Normality." Journal of the Royal Statistical Soceity, Series B, Vol. 31 (1975), 472-476.

Fraser, D. A. S. "Data Transformations and the Linear Model." Annals of Mathematical Statistics, Vol. 38 (1967), 1456-1465.

Hinkley, D. V. "On Power Transformations to Symmetry." Biometrika, Vol. 62 (1975), 101-111.

Lindsey, J. K. "Fitting Response Surfaces With Power Transformation." Journal of the Royal Statistical Soceity, Series C, Vol. 21 (1972), 234-237.

Lindsey, J. K. "The Roles of Transformations to Normality." Biometrics, Vol. 31 (1975), 247-249.

Olmasted, J. M. H. Advanced Calculus. New York: Appleton Cintury Crofts, 1959.

Rao, C. R. "A Note on Generalized Inverse of a Matrix With Application to Problems in Mathematical Statistics." Journal of the Royal Statistical Soceity, Series B, Vol. 24 (1962), 152-158.

Rao, C. R. Linear Statistical Inference and its Application (2nd Ed.). New York: Wiley, 1973.

Schlesselman, J. "Power Families: A Note on the Box and Cox Transformation." Journal of the Royal Statistical Soceity, Series B, Vol. 33 (1971), 307-311.

Shapiro, S. S. and Wilk, M. B. "Analysis of Variance Test for Normality (complete samples)." Biometrika, Vol. 52 (1965), 591-611.

Steel, R. G. D. and Torrie, J. H. Principles and Procedures of Statistics. New York: McGraw-Hill, 1960.

Tukey, J.W. "On the Comparative Anatomy of Transformation." Annals of Mathematical Statistics, Vol. 28 (1957), 602-632.

APPENDIX A

NUMERICAL METHOD FOR FINDING QUANTILE ESTIMATE

Recall the equation (3.11) in Chapter III

$$\begin{cases} a\, b^{\hat{\lambda}} + (1-a)\, c^{\hat{\lambda}} = 1, \\ \eta_q(1 - d^{\hat{\lambda}}) = \eta_p(b^{\hat{\lambda}} - c^{\hat{\lambda}}) \qquad \text{for } \hat{\lambda} \neq 0, \end{cases} \tag{A.1}$$

where $a = (\eta_p + \eta_q)/2\eta_q$, $b = y_{(n-j+1)}/y_{(n-i+1)}$, $c = y_{(j)}/y_{(n-i+1)}$, and $d = y_{(i)}/y_{(n-i+1)}$.

Since $0 < p < q < \frac{1}{2}$, $i = [np]$ and $j = [nq]$, we have $y_{(i)} \leqslant y_{(j)} < y_{(n-j+1)} \leqslant y_{(n-i+1)}$ and $\eta_p < \eta_q < 0$. Therefore, $1 \geqslant b > c \geqslant d > 0$ and $a > 1$.
Let function $g(\hat{\lambda})$ be

$$g(\hat{\lambda}) = a\, b^{\hat{\lambda}} + (1-a)\, c^{\hat{\lambda}} - 1. \tag{A.2}$$

Then the limits of the function $g(\hat{\lambda})$ are

$$\lim_{\hat{\lambda}\to\infty} g(\hat{\lambda}) = a \lim_{\hat{\lambda}\to\infty} b^{\hat{\lambda}} + (1-a) \lim_{\hat{\lambda}\to\infty} c^{\hat{\lambda}} - 1 = -1, \tag{A.3}$$

$$\lim_{\hat{\lambda}\to-\infty} g(\hat{\lambda}) = a \lim_{\hat{\lambda}\to-\infty} b^{\hat{\lambda}} + (1-a) \lim_{\hat{\lambda}\to-\infty} c^{\hat{\lambda}} - 1 = (1-a) \lim_{\hat{\lambda}\to-\infty} c^{\hat{\lambda}} = -\infty. \tag{A.4}$$

The limits of $g(\hat{\lambda})$ tell us that the equation $g(\hat{\lambda}) = 0$ has an even number of roots or none, but we know that it also has a trivial root zero. Therefore, we can say that this equation $g(\hat{\lambda}) = 0$ has at least one nonzero root or multiple zero roots.

Let us separate $g(\hat{\lambda})=0$ into two equations with a parameter k.

Thus,

$$\begin{cases} a\, b^{\hat{\lambda}} = k, \\ (1-a)c^{\hat{\lambda}} = 1-k. \end{cases} \tag{A.5}$$

It can be rewritten to

$$\begin{cases} \hat{\lambda} = \log_b \frac{k}{a}, \\ \hat{\lambda} = \log_c \frac{1-k}{1-a}, \end{cases} \tag{A.6}$$

and these two logarithmic equations have at most two intersections of which the value of $\hat{\lambda}$ is the root of $g(\hat{\lambda})=0$ including the intersection at $\hat{\lambda}=0$ and $k=a$. Hence, it simplifies the number of roots to two possible cases; a nonzero root and a zero root or double zero roots. If the latter is true, then the solution of $g'(\hat{\lambda})=0$ should be located at zero. Thus

$$g'(\hat{\lambda}) = ab^{\hat{\lambda}} \log b + (1-a)c^{\hat{\lambda}} \log c = 0. \tag{A.7}$$

$$\Longrightarrow \qquad ab^{\hat{\lambda}} \log b = (a-1)c^{\hat{\lambda}} \log c.$$

$$\Longrightarrow \qquad \hat{\lambda} = \log_{\frac{b}{c}} \frac{\log c^{a-1}}{\log b^{a}}.$$

Because of $c^{a-1} \neq b^{a}$, the equation, $g'(\hat{\lambda})=0$, does not have a root at zero. Hence, the equation $g(\hat{\lambda})=0$ for $\hat{\lambda}\neq 0$ has only one nonzero root.

Two methods are suggested to find the nonzero solution of $g(\hat{\lambda})=0$.

1. Use the numerical iteration method called the "Modified Newton-Raphson Method" which is

$$\hat{\lambda}_{i+1} = \hat{\lambda}_i - \frac{U(\hat{\lambda}_i)}{U'(\hat{\lambda}_i)}, \tag{A.8}$$

where

$$U(\hat{\lambda}_i) = g(\hat{\lambda}_i)/\ g'(\hat{\lambda}_i),$$

$$U'(\hat{\lambda}_i) = 1 - g(\hat{\lambda}_i)g''(\hat{\lambda}_i)/[g'(\hat{\lambda}_i)]^2.$$

The initial guess will be

$$\hat{\lambda}_1 = \log_{\frac{b}{c}} \left[\frac{\log c^{a-1}}{\log b^a}\right] \pm \delta \tag{A.9}$$

where the sign will associate with the sign of the logarithmic term and δ is a reasonably small positive real number.

2. Plot the functions of (A.6) against a trial values of k, the nonzero value of $\hat{\lambda}$ at the two intersections will be the answer.

APPENDIX B

PROOFS OF LEMMAS OF TRUNCATED NORMAL DISTRIBUTION

Lemma 1 The truncated normal density of random variable X is

$$f_X(x) = \begin{cases} \dfrac{1}{\sqrt{2\pi}\,\sigma(1-\alpha-\beta)} e^{-\frac{(x-\mu)^2}{2\sigma^2}} & \text{for } a < x < b, \\ 0 & \text{otherwise.} \end{cases} \tag{B.1}$$

For the transformation $Z = (X - \mu)/\sigma$, the Jacobian is

$$J = \left| \frac{dz}{dx} \right| = \sigma. \tag{B.2}$$

Hence, the density function of the random variable Z is

$$f_Z(z) = f_X(\mu + \sigma z)\cdot\sigma$$

$$= \begin{cases} \dfrac{1}{\sqrt{2\pi}(1-\alpha-\beta)} e^{-\frac{z^2}{2\sigma^2}} & \text{for } \frac{a-\mu}{\sigma} < z < \frac{b-\mu}{\sigma}, \\ 0 & \text{otherwise.} \end{cases} \tag{B.3}$$

Lemma 2 For random variable $X \sim N^*(0, 1)$ with the domain $a < x < \infty$, the distribution will be

$$F_X(x) = \int_a^x \frac{1}{\sqrt{2\pi}(1-\alpha)} e^{-t^2/2}\, dt = \frac{1}{1-\alpha} \int_a^x \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt$$

$$= \frac{1}{1-\alpha}\,(\phi(x) - \phi(x)) = \frac{1}{1-\alpha}\,(\phi(x) - \alpha) \qquad \text{for } a < x < \infty, \qquad (B.4)$$

and $F_X(x) = 0$ for $x \leqslant a$.

If $F_X(x) = k$, then

$$(1-\alpha)k = \phi(x) - \alpha$$

$$\Rightarrow \quad \phi(x) = (1-\alpha)k + \alpha$$

$$\Rightarrow \quad x = \phi^{-1}\left[(1-\alpha)k + \alpha\right]. \qquad (B.5)$$

<u>Lemma</u> 3 For random variable $X \backsim N^*(0, 1)$ with the domain $-\infty < x < b$, the distribution function will be

$$F_X(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}(1-\beta)}\, e^{-t^2/2}\, dt$$

$$= \frac{\phi(x)}{1-\beta} \qquad \text{for } -\infty < x < b, \qquad (B.6)$$

and $F_X(x) = 1$ for $b \leqslant x$.

If $F_X(x) = k$, then

$$(1-\beta)k = \phi(x)$$

$$\Rightarrow \quad x = \phi^{-1}\left[(1-\beta)k\right]. \qquad (B.7)$$

<u>Lemma</u> 4 For random variable $X \backsim N^*(\mu, \sigma^2)$ with the domain $a < x < b$, then the expected value of X is

$$E(X) = \int_a^b \frac{x}{\sqrt{2\pi}\sigma(1-\alpha-\beta)}\, e^{-(x-\mu)^2/2\sigma^2}\, dx$$

$$= \int_a^b \frac{-\sigma^2}{\sqrt{2\pi}\sigma(1-\alpha-\beta)}\, e^{-(x-\mu)^2/2\sigma^2}\, d\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] + \mu$$

$$= \frac{\sigma^2}{\sqrt{2\pi}(1-\alpha-\beta)} e^{-(x-\mu)^2/2\sigma^2}\Big|_a^b + \mu$$

$$= \mu - \frac{g(b)-g(a)}{1-\alpha-\beta} \sigma^2, \tag{B.8}$$

where

$$g(t) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-(t-\mu)^2/2\sigma^2}.$$

<u>Lemma</u> 5 Let $X_1, X_2, \ldots, X_n$ be independently distributed $N^*(\mu, \sigma^2)$ for $a < x_i < b$, then the likelihood function is

$$L(\mu, \sigma^2 | x_1, x_2, \ldots, x_n) = \frac{1}{(2\pi)^{n/2}\sigma^n(1-\alpha-\beta)^n} e^{-\frac{\sum_1^n (x_i-\mu)^2}{2\sigma^2}}. \tag{B.9}$$

Hence,

$$\log L = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log \sigma^2 - n\log(1-\alpha-\beta) - \sum_1^n (x_i-\mu)^2/2\sigma^2. \tag{B.10}$$

Therefore,

$$\frac{\partial \log L}{\partial \mu} = \frac{\sum_1^n (x_i-\mu)}{\sigma^2} - \frac{n}{(1-\alpha-\beta)} \frac{\partial}{\partial \mu}\int_a^b \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt$$

$$= \frac{\sum_1^n (x_i-\mu)}{\sigma^2} - \frac{n}{1-\alpha-\beta}\int_a^b \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(t-\mu)^2}{2\sigma^2}} \frac{t-\mu}{\sigma^2} dt$$

$$= \frac{\sum_1^n (x_i-\mu)}{2} + \frac{n}{1-\alpha-\beta}\int_a^b \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(t-\mu)^2}{2\sigma^2}} d\left[-\frac{(t-\mu)^2}{2\sigma^2}\right]$$

$$= \sum_1^n (x_i-\mu)/\sigma^2 + \frac{n}{1-\alpha-\beta} \frac{1}{\sqrt{2\pi}\,\sigma} e^{-(t-\mu)^2/2\sigma^2}\Big|_a^b$$

$$= \frac{\sum_1^n (x_i - \mu)}{\sigma^2} - \frac{n}{1-\alpha-\beta}\Big[g(b) - g(a)\Big], \tag{B.11}$$

and

$$\frac{\partial \log L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{\sum_1^n (x_i-\mu)^2}{2\sigma^4} - \frac{n}{1-\alpha-\beta}\frac{\partial}{\partial \sigma^2}\int_a^b \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt$$

$$= -\frac{n}{2\sigma^2} + \frac{\sum_1^n (x_i-\mu)^2}{2\sigma^4} - \frac{n}{1-\alpha-\beta}\Bigg\{\int_a^b \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(t-\mu)^2}{2\sigma^2}} \frac{(t-\mu)^2}{2\,\sigma^4} dt$$

$$- \frac{1}{2\sigma^2}\int_a^b \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt\Bigg\}$$

$$= -\frac{n}{2\sigma^2} + \frac{\sum_1^n (x_i-\mu)^2}{2\sigma^4} - \frac{n}{1-\alpha-\beta}\Bigg\{-\frac{(t-\mu)}{2\sigma^2}\frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(t-\mu)^2}{2\sigma^2}}\Bigg|_a^b$$

$$+ \frac{1}{2\sigma^2}\int_a^b \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt - \frac{1}{2\sigma^2}\int_a^b \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt\Bigg\}$$

$$= -\frac{n}{2\sigma^2} + \frac{\sum_1^n (x_i-\mu)^2}{2\sigma^4} + n\,\frac{(b-\mu)g(b) - (a-\mu)g(a)}{2\sigma^2}. \tag{B.12}$$

Let $\frac{\partial \log L}{\partial \mu} = 0$ and $\frac{\partial \log L}{\partial \sigma^2} = 0$, then the maximum likelihood estimates of μ and σ^2 are the solutions of the following

$$\hat{\mu} = \frac{\sum_1^n x_i}{n} + \frac{\hat{\sigma}^2\big[\hat{g}(b) - \hat{g}(a)\big]}{1 - \hat{\alpha} - \hat{\beta}}, \tag{B.13}$$

$$\hat{\sigma}^2 = \frac{\sum_1^n (x_i - \hat{\mu})^2}{n} \frac{1}{1 - \frac{(b-\hat{\mu})\hat{g}(b) - (a-\hat{\mu})\hat{g}(a)}{1 - \hat{\alpha} - \hat{\beta}}} \tag{B.14}$$

where

$$\hat{\alpha} = \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}\,\hat{\sigma}} e^{-\frac{(t-\hat{\mu})^2}{2\hat{\sigma}^2}} dt,$$

$$\hat{\beta} = \int_{b}^{\infty} \frac{1}{\sqrt{2\pi}\,\hat{\sigma}} e^{-\frac{(t-\hat{\mu})^2}{2\hat{\sigma}^2}} dt,$$

$$\hat{g}(a) = \frac{1}{\sqrt{2\pi}\,\hat{\sigma}} e^{-\frac{(a-\hat{\mu})^2}{2\hat{\sigma}^2}},$$

and

$$\hat{g}(b) = \frac{1}{\sqrt{2\pi}\,\hat{\sigma}} e^{-\frac{(b-\hat{\mu})^2}{2\hat{\sigma}^2}}.$$

APPENDIX C

DERIVATION OF MAXIMUM LIKELIHOOD ESTIMATES OF POWER TRANSFORMATION TO A TRUNCATED NORMAL DISTRIBUTION

1. $\lambda > 0$

Take the logarithm of (4.9), then

$$\log L(\lambda) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log \sigma^2 - n\log(1-\alpha) + (\lambda-1)\sum_{1}^{n}\log y_i$$
$$- \frac{\left[\frac{1}{\lambda}(\underline{y}^\lambda - \underline{1}) - A\underline{\theta}\right]'\left[\frac{1}{\lambda}(\underline{y}^\lambda - \underline{1}) - A\underline{\theta}\right]}{2\sigma^2}. \qquad (C.1)$$

For convenience, let μ_j = minimum of $\{\mu_i\}$, then

$$\alpha = \int_{-\infty}^{-1/\lambda \, - \, \text{min. of } \{\mu_i\}} \frac{1}{\sqrt{2\pi}\,\sigma} e^{-t^2/2\sigma^2} dt$$
$$= \int_{-\infty}^{-1/\lambda} \frac{1}{\sqrt{2\pi}\,\sigma} e^{-(t-\mu_j)^2/2\sigma^2} dt.$$

The first partial derivative of log $L(\lambda)$ with respect to $\underline{\theta}$ is

$$\frac{\partial \log L(\lambda)}{\partial \underline{\theta}} = -\frac{A'A\underline{\theta} - A'\frac{1}{\lambda}(\underline{y}^\lambda - \underline{1})}{\sigma^2} - n\begin{bmatrix} \frac{\partial}{\partial\theta_1}\log(1-\alpha) \\ \vdots \\ \frac{\partial}{\partial\theta_n}\log(1-\alpha) \end{bmatrix}, \qquad (C.2)$$

where

$$\frac{\partial}{\partial\theta_i}\log(1-\alpha) = \frac{1}{1-\alpha}\frac{\partial}{\partial\theta_i}\int_{-1/\lambda}^{\infty}\frac{1}{\sqrt{2\pi}\,\sigma}\, e^{-(t-\mu_j)^2/2\sigma^2}\,dt$$

$$= \frac{1}{1-\alpha}\int_{-1/\lambda}^{\infty}\frac{1}{\sqrt{2\pi}\,\sigma}\, e^{-(t-\mu_j)^2/2\sigma^2}\cdot\frac{(t-\mu_j)}{2}\cdot\frac{\partial\mu_j}{\partial\theta_i}\,dt \quad \text{for } i=1,2,\ldots,p.$$

Since $\mu_j = \sum_{i=1}^{p} a_{ji}\theta_i$, where a_{ji} is the jith element of matrix A, then

$$\frac{\partial\mu_j}{\partial\theta_i} = a_{ji}\ . \tag{C.3}$$

Therefore,

$$\frac{\partial}{\partial\theta_i}\log L(\lambda) = \frac{a_{ji}}{1-\alpha}\int_{-\frac{1}{\lambda}}^{\infty}\frac{1}{\sqrt{2\pi}\,\sigma}\, e^{-(t-\mu_j)^2/2\sigma^2}\, d\,\frac{(t-\mu_j)^2}{2\sigma^2}$$

$$= -\frac{a_{ji}}{1-\alpha}\frac{1}{\sqrt{2\pi}\,\sigma}\, e^{-(t-\mu_j)^2/2\sigma^2}\Bigg|_{-1/\lambda}^{\infty}$$

$$= \frac{a_{ji}}{1-\alpha}\, g_j(-\tfrac{1}{\lambda}). \tag{C.4}$$

Hence,

$$\frac{\partial}{\partial\underline{\theta}}\log L(\lambda) = -\frac{A'A\underline{\theta} - A'\frac{1}{\lambda}(\underline{y}^{\lambda}-\underline{1})}{\sigma^2} - \frac{n}{1-\alpha}\, g_j(-\tfrac{1}{\lambda})\,\underline{a}_j, \tag{C.5}$$

where

$$\underline{a}_j = \begin{bmatrix} a_{j1} \\ a_{j2} \\ \vdots \\ a_{jp} \end{bmatrix}.$$

The first partial derivative of log $L(\lambda)$ with respect to σ^2 is

$$\frac{\partial}{\partial\sigma^2}\log L(\lambda) = -\frac{n}{2\sigma^2} + \frac{\left[\frac{1}{\lambda}(\underline{y}^\lambda - \underline{1}) - A\underline{\theta}\right]'\left[\frac{1}{\lambda}(\underline{y}^\lambda - \underline{1}) - A\underline{\theta}\right]}{2\sigma^4}$$

$$-\frac{n}{1-\alpha}\frac{\partial}{\partial\sigma^2}\int_{-\frac{1}{\lambda}}^{\infty}\frac{1}{\sqrt{2\pi}\sigma}e^{-(t-\mu_j)^2/2\sigma^2}dt. \qquad (C.6)$$

The partial derivative in the last term of (C.6) will be

$$\frac{\partial}{\partial\sigma^2}\int_{-\frac{1}{\lambda}}^{\infty}\frac{1}{\sqrt{2\pi}\sigma}e^{-(t-\mu_j)^2/2\sigma^2}dt$$

$$=\int_{-\frac{1}{\lambda}}^{\infty}\frac{1}{\sqrt{2\pi}\sigma}e^{-(t-\mu_j)^2/2\sigma^2}\frac{(t-\mu_j)^2}{2\sigma^4}dt - \frac{1}{2\sigma^2}\int_{-\frac{1}{\lambda}}^{\infty}\frac{1}{\sqrt{2\pi}\sigma}e^{-(t-\mu_j)^2/2\sigma^2}dt$$

$$= -\frac{1}{2\sigma^2}(t-\mu_j)\, e^{-(t-\mu_j)^2/2\sigma^2}\Big|_{-\frac{1}{\lambda}}^{\infty}$$

$$= \frac{1}{2\sigma^2}\left(-\frac{1}{\lambda} - \mu_j\right) g_j\left(-\frac{1}{\lambda}\right). \qquad (C.7)$$

Therefore,

$$\frac{\partial}{\partial\sigma^2}\log L(\lambda) = -\frac{n}{2\sigma^2} + \frac{\left[\frac{1}{\lambda}(\underline{y}^\lambda - \underline{1}) - A\underline{\theta}\right]'\left[\frac{1}{\lambda}(\underline{y}^\lambda - \underline{1}) - A\underline{\theta}\right]}{2\sigma^4}$$

$$-\frac{n}{2\sigma^2(1-\alpha)}\left(-\frac{1}{\lambda} - \mu_j\right) g_j\left(-\frac{1}{\lambda}\right). \qquad (C.8)$$

Let (C.5) and (C.8) equal zero, then the maximum likelihood estimates of $\underline{\theta}$ and σ^2 are the solutions of the following

$$\hat{\underline{\theta}} = (A'A)^{-1}\left[A'\,\frac{1}{\lambda}(\underline{y}^\lambda - \underline{1}) - \frac{n\hat{\sigma}^2}{1-\hat{\alpha}}\,\hat{g}_k\left(-\frac{1}{\lambda}\right)\underline{a}_k\right], \qquad (C.9)$$

$$\hat{\sigma}^2 = \frac{\left[\frac{1}{\lambda}(\underline{y}^\lambda - \underline{1}) - A\hat{\underline{\theta}}\right]' \left[\frac{1}{\lambda}(\underline{y}^\lambda - \underline{1}) - A\hat{\underline{\theta}}\right]}{n\left\{1 + \frac{\hat{g}_k(-\frac{1}{\lambda})}{1-\hat{\alpha}}\left(-\frac{1}{\lambda} - \hat{\mu}_k\right)\right\}}, \tag{C.10}$$

where k, the estimate of j, is the number of the row associating with the minimum value of elements of $A\hat{\underline{\theta}}$, $\underline{a}_k$ is the p x 1 vector of the kth row of A,

$$\hat{g}_k(-\frac{1}{\lambda}) = \frac{1}{\sqrt{2\pi}\,\hat{\sigma}}\, e^{-\frac{(-\frac{1}{\lambda} - \hat{\mu}_k)^2}{2\hat{\sigma}^2}},$$

and

$$\hat{\alpha} = \int_{-\infty}^{-1/\lambda - \hat{\mu}_k} \frac{1}{\sqrt{2\pi}\,\hat{\sigma}}\, e^{-\frac{t^2}{2\hat{\sigma}^2}} dt.$$

2. $\lambda < 0$

Take the logarithm of (4.14)

$$\log L(\lambda) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - n\log(1-\alpha) - (\lambda-1)\sum_1^n \log y_i - \frac{\left[\frac{1}{\lambda}(\underline{y}^\lambda - \underline{1}) - A\underline{\theta}\right]' \left[\frac{1}{\lambda}(\underline{y}^\lambda - \underline{1}) - A\underline{\theta}\right]}{2\sigma^2}. \tag{C.11}$$

For the same reason of convenience, let μ_j = maximum of $\{\mu_i\}$. The first derivatives of log $L(\lambda)$ with respect to $\underline{\theta}$ and σ^2 are just the same as those of $\lambda > 0$ excepting the term of $\log(1-\alpha)$. Thus

$$\frac{\partial}{\partial\theta_i}\log(1-\alpha) = -\frac{a_{ji}}{1-\alpha}\frac{1}{\sqrt{2\pi}\,\sigma}\, e^{-(t-\mu_j)^2/2\sigma^2}\Big|_{-\infty}^{-\frac{1}{\lambda}} = \frac{-a_{ji}}{1-\alpha}g_j(\frac{-1}{\lambda}), \tag{C.12}$$

and

$$\frac{\partial}{\partial \sigma^2} \log(1-\alpha) = \frac{-1}{2\sigma^2}(-\frac{1}{\lambda} - \mu_j)\ g_j(-\frac{1}{\lambda}). \tag{C.13}$$

Therefore,

$$\frac{\partial}{\partial \underline{\theta}} \log L(\lambda) = -\frac{A'A\underline{\theta} - A'\ \frac{1}{\lambda}(\underline{Y}^{\lambda}-\underline{1})}{\sigma^2} + \frac{n}{1-\alpha} g_j(-\frac{1}{\lambda})\ \underline{a}_j \tag{C.14}$$

and

$$\frac{\partial}{\partial \sigma^2} \log L(\lambda) = \frac{-n}{2\sigma^2} + \frac{\left[\frac{1}{\lambda}(\underline{Y}^{\lambda}-\underline{1})-A\underline{\theta}\right]'\left[\frac{1}{\lambda}(\underline{Y}^{\lambda}-\underline{1})-A\underline{\theta}\right]}{2\sigma^4}$$

$$+ \frac{n}{2\sigma^2(1-\alpha)}(-\frac{1}{\lambda} - \mu_j)\ g_j(-\frac{1}{\lambda}). \tag{C.15}$$

Let (C.14) and (C.15) equal zero, then the maximum likelihood estimates of $\underline{\theta}$ and σ^2 are the solutions of the following

$$\hat{\underline{\theta}} = (A'A)^{-1}\left[A'\ \frac{1}{\lambda}(\underline{Y}^{\lambda}-\underline{1}) + \frac{n\hat{\sigma}^2}{1-\hat{\alpha}}\hat{g}_k(-\frac{1}{\lambda})\ \underline{a}_k\right], \tag{C.16}$$

$$\hat{\sigma}^2 = \frac{\left[\frac{1}{\lambda}(\underline{Y}^{\lambda}-\underline{1})- A\hat{\underline{\theta}}\right]'\left[\frac{1}{\lambda}(\underline{Y}^{\lambda}-\underline{1})- A\hat{\underline{\theta}}\right]}{n\left(1 - \frac{\hat{g}_k(-\frac{1}{\lambda})}{1-\hat{\alpha}}(-\frac{1}{\lambda} - \hat{\mu}_k)\right)}, \tag{C.17}$$

where k, the estimate of j, is the number of the row associating with the maximum value of elements of $A\hat{\underline{\theta}}$, $\underline{a}_k$ is the p x 1 vector of the kth row of A,

$$\hat{g}_k(-\frac{1}{\lambda}) = \frac{1}{\sqrt{2\pi}\,\hat{\sigma}}\, e^{-(-1/\lambda - \hat{\mu}_k)^2/2\hat{\sigma}^2}$$

and

$$\hat{\alpha} = \int_{-1/\lambda - \hat{\mu}_k}^{\infty} \frac{1}{\sqrt{2\pi}\,\hat{\sigma}} e^{-\frac{t^2}{2\hat{\sigma}^2}} \, dt.$$

VITA

Gwong-Chain Sun

Candidate for the Degree of

Doctor of Philosophy

Thesis: STUDY ON THE BOX AND COX POWER TRANSFORMATION TO NORMALITY

Major Field: Statistics

Biographical:

Personal Data: Born in Chungking, China, July 1, 1944, the third son of Chiang and Fang-Jung Sun.

Education: Graduated from Taichung High School, Taiwan, China, in 1963; received Bachelor of Science degree in Electrical Engineering from National Taiwan University, Taiwan, China, in 1968; received Master of Science degree in Electrical Engineering from State University of New York at Stony Brook in 1972; received Master of Science degree in Statistics from University of Arkansas in 1976; completed requirements for the Doctor of Philosophy degree at Oklahoma State University in May, 1978.

Professional Experience: Production Engineer, General Instruments Company, Taiwan, China, 1969-1970; Teaching Assistant in Electrical Engineering Department, National Taiwan University, Taiwan, China, 1970-1971; Graduate Teaching Assistant in the Department of Mathematics, University of Arkansas, Fayetteville, Arkansas, 1972-1975; Graduate Teaching Assistant in The Department of Statistics, Oklahoma State University, Stillwater, Oklahoma, 1976-1978.

Professional Organizations: American Statistical Association, The Institute of Electrical and Electronics Engineers.