

A Brief Review of Transformation

Xiaohui Cai
Department of Mathematical Sciences
Montana State University

December 2, 1998

A writing project submitted in partial fulfillment
of the requirements for the degree

Master of Science in Statistics

APPROVAL

of a writing project submitted by

XIAOHUI CAI

This writing project has been read by the writing project director and has been found to be satisfactory regarding content, English usage, format, citations, bibliographic style, and consistency, and is ready for submission to the Statistics Faculty.

12/16/98
Date

Robert J. Boik
Robert J. Boik
Writing Project Director

1 Abstract

This article briefly reviews the major transformation methods in statistical regression. Both Maximum Likelihood Estimation and Bayesian Methods in Box-Cox transformation are explained in detail. One example is given for demonstration.

2 Introduction

In linear model analysis, the following assumptions usually are required:

- (a) Additivity — error terms are added to the main effect to “explain” observation.
- (b) Constant variance (homogeneity) — error terms are treated as homogeneous.
- (c) Normality — observations are normally distributed.
- (d) Independence — observations are independent of each other.

In practice, one or more of above assumptions may be violated. Statisticians have several options for analysis. One is to transform the data so that those assumptions are tenable.

One common method to satisfy homogeneity assumption is based on the empirical relation between the variance and the mean. The theory behind this is as follows. For function $f(y)$, if the first order derivative $f'(y)$ is continuous and the second order derivative $f''(y)$ is finite, we can apply the Taylor series expansion and get $f(y_i) - f(\eta_i) \approx f'(\eta_i)(y_i - \eta_i)$, where $\eta_i = E(y_i)$. Introducing $\sigma_i^2(\eta_i)$ to be a function of $E(y_i)$, $\sigma_i^2(\eta_i) = E[y_i - E(y_i)]^2$, then write $\text{var}(f(y_i))$ as $\sigma_i^2(\eta_i)$. Now squaring both sides and taking expectations, we will have

$$\text{var}(f(y_i)) \approx (f'(\eta_i))^2 \sigma_i^2.$$

Therefore, in order to keep $\text{var}(f(y_i))$ approximately constant, we need to find a proper transformation $f(y_i)$.

$$\text{var}(f(y_i)) = \text{constant}$$

$$\rightarrow f'(\eta_i) \sigma_i(\eta_i) = c$$

$$\rightarrow f'(\eta_i) = c / \sigma_i(\eta_i)$$

$$\rightarrow f(\eta_i) = \int [c / \sigma_i(\eta_i)] d\eta_i$$

Integrate the right-hand-side and $f(\eta_i)$ is the transformation formula.

Let's look at several examples. For the binomial distribution, the pmf is $p(Y=y|n,p) = \binom{n}{y} p^y (1-p)^{n-y}$, $EY = np$, $\text{Var}Y = np(1-p)$. Let $\eta = EY$, $\text{Var}Y = \eta(1-\eta/n)$.

$$\rightarrow f'(\eta) (\text{Var}Y)^{1/2} = c$$

$$\rightarrow f'(\eta) = c \times [\eta(1-\eta/n)]^{-1/2}$$

$$\rightarrow f(\eta) = \int c \times [\eta(1-\eta/n)]^{-1/2} d\eta$$

$$\rightarrow f(\eta) = c^* \arcsin[\text{sqrt}(\eta/n)]$$

Here we treat η as a continuous variable. It is a good approximate as $n \rightarrow \infty$. This is called the Angular transformation and further discussions are continued in section 3.4.

If we have $\sigma(\eta) = \eta^2$, then $f(\eta) = \int [c/\sigma(\eta)]d\eta = \int c\eta^{-2} d\eta = c^*/\eta$. This is the reciprocal transformation.

If $\sigma(\eta) = \eta^{1/2}$, similarly, we can get $f(\eta) = \eta^{1/2}$, which is square root transformation. More details about it are listed in section 3.3.

3 Some Commonly Used Transformations

This part focuses on dependent variable transformation.

3.1 Power transformation and some alternative versions:

Turkey (1957) introduced a family of power transformation:

$$\mathbf{y}^{(\lambda)} = \begin{cases} \mathbf{y}^\lambda, & \lambda \neq 0; \\ \log \mathbf{y}, & \lambda = 0. \end{cases} \quad (1)$$

Box and Cox (1964) made some improvement so that this function is continuous at $\lambda=0$.

$$\mathbf{y}^{(\lambda)} = \begin{cases} (\mathbf{y}^\lambda - 1)/\lambda, & \lambda \neq 0; \\ \log \mathbf{y}, & \lambda = 0. \end{cases} \quad (2)$$

The model here is $\mathbf{y}^{(\lambda)} = (\mathbf{y}_1^{(\lambda)}, \mathbf{y}_2^{(\lambda)}, \dots, \mathbf{y}_n^{(\lambda)})' = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\boldsymbol{\beta}$ is a vector of unknown parameters and $\boldsymbol{\varepsilon} \sim \text{MVN}(0, \sigma^2 \mathbf{I}_n)$ is the error term.

However, the above transformation can only be made when $y_i > 0$. With negative observations, the following form may be used:

$$\mathbf{y}^{(\lambda)} = \begin{cases} \{(\mathbf{y} + \lambda_2)^{\lambda_1} - 1\} / \lambda_1, & \lambda_1 \neq 0; \\ \log(\mathbf{y} + \lambda_2), & \lambda_1 = 0. \end{cases} \quad (3)$$

The transformation introduced by Manly in 1976 is capable not only of taking care of negative observations, but also of changing skewed unimodal distributions into nearly symmetric normal-like distributions:

$$\mathbf{y}^{(\lambda)} = \begin{cases} (e^{\lambda y} - 1) / \lambda, & \lambda \neq 0; \\ \mathbf{y}, & \lambda = 0. \end{cases} \quad (4)$$

For nearly symmetric distributions, John & Draper (1980) suggested the modulus transformation:

$$y^{(\lambda)} = \begin{cases} \text{sign}(y) * \{(|y|^\lambda - 1) / \lambda, & \lambda \neq 0; \\ \log y, & \lambda = 0. \end{cases} \quad (5)$$

In order to include those distributions of $y^{(\lambda)}$ with unbounded support, Bickel & Doksum (1981) suggested:

$$y_i^{(\lambda)} = \frac{|y_i|^\lambda \text{sign}(y_i) - 1}{\lambda} \quad \text{for } \lambda > 0.$$

We should note that range of (1) (2) (3) & (5) is restricted. So the transformed values do not cover the entire range $(-\infty, +\infty)$. Accordingly, only approximate normality is possible.

3.2 Reciprocal Transformation:

$$y^{(\lambda)} = y^{-1}.$$

This can reduce skewness. When y^{-1} has a physical meaning, and we needn't worry about zero or negative values, this method is suitable.

3.3 Square Root and Cube Root Transformation:

$$y^{(\lambda)} = (y + \lambda)^{1/2}.$$

When observations follow a Poisson distribution, then the square root transformation is used. Anscombe (1948) showed $\lambda = 3/8$ will give the best result if we want to get constant-variance. Later, Kihlberg, Herson and Schotz (1967) concluded $\lambda = 0.386$ is the best.

Freeman & Tukey (1950) suggested the Chordal transformation for the Poisson distribution if the expectation is small, then

$$y_i^{(\lambda)} = \sqrt{y_i} + \sqrt{y_i + 1}.$$

When y_i obeys a $\chi^2_{(\nu)}$ distribution with ν degrees of freedom, then $y_i^{(\lambda)} = \sqrt{2 y_i}$ is used. Fisher (1925) showed that the transformed distribution is approximately $N(\sqrt{2\nu-1}, 1)$. For the same distribution, Wilson & Hilferty (1931) used the cube root transformation:

$$y^{(\lambda)} = (y / \nu)^{1/3}.$$

which is approximately $N(1 - 2/(9\nu), 2/(9\nu))$. And for $\chi^2_{(\nu)}$, it's better than the previous one.

3.4 Binomial Distribution:

The Angular transformation is considered the best. It was first mentioned by Fisher (1954). Eisenhert, Hastay and Wellis refined it later:

$$y_i^{(\lambda)} = \arcsin \left\{ \sqrt{[(y_i + \lambda_1) / (n + \lambda_2)]} \right\}.$$

Early users used $\lambda_1 = \lambda_2 = 0$, Bartlett (1936) suggested $\lambda_1 = 0.5$, $\lambda_2 = 0$. Anscombe (1948) preferred $\lambda_1 = 0.375$, $\lambda_2 = 0.75$.

Fisher (1922, 1930) also gave:

$$y_i^{(\lambda)} = \arcsin (y_i).$$

which is known as Arc Sin transformation. However, it quite depends on sample size and is unstable for variance.

For a negative binomial, the following transformation was first introduced by Beall in 1942:

$$y_i^{(\lambda)} = \operatorname{arcsinh} \left\{ \sqrt{[(y_i + \lambda_1) / (k + \lambda_2)]} \right\}.$$

He recommended zero for both λ values. Anscombe (1948) suggested $\lambda_1 = 0.375$, $\lambda_2 = -0.75$. Additionally, he further simplified it to $y_i^{(\lambda)} = \log (y_i + 0.5k)$ which is as good as Beall's.

3.5 Normal Scores and Exponential Scores:

In psychological preferences, usually only the order of data is known. Those data can be replaced with their corresponding expected values by applying order statistics to a size n standard normal distribution. This was first introduced by Fisher and Yates (1938).

Exponential scores are basically the same, except that this method uses the unit exponential distribution. And thus we can calculate the usual exponential theory test statistics. This was developed by Cox (1964).

4 Parameter Estimation

The assumption here is that $\mathbf{y}^{(\lambda)} = \mathbf{X}\beta + \varepsilon$, where $\varepsilon \sim N(0, \sigma^2)$; \mathbf{X} is a known matrix and β is the parameter vector that needs to be evaluated. Here are some methods to find out λ and its confidence interval.

One of them is power (Box-Cox) transformation, of which the simplest form is (2). Box and Cox have done a lot of analysis on it (See reference 1). We assume $\mathbf{y}^{(\lambda)}$ is normally distributed, λ is unknown, the likelihood *in relation to these original observations* \mathbf{y} is:

$$(6) \quad (2\pi)^{-n/2} \sigma^{-n} \exp\{-(2\sigma^2)^{-1} (\mathbf{y}^{(\lambda)} - \mathbf{X}\beta)'(\mathbf{y}^{(\lambda)} - \mathbf{X}\beta)\} J(\lambda; \mathbf{y}),$$

where $J(\lambda; \mathbf{y}) = \prod_1^n |d y_i^{(\lambda)} / d y_i| = \prod_1^n y_i^{\lambda-1}$
is called Jacobian.

There are two ways to get the inference about the parameter λ in formula (6). ① “orthodox” large-sample maximum-likelihood theory. This will lead to point estimates of parameters and chi-squared distribution of confidence intervals. ② Bayes’s theorem. Here prior distribution of β and $\log \sigma$ are assumed uniformly distributed in the region where likelihood is appreciable. The posterior distribution of λ can be found by integrating over the parameters.

4.1 MLE method:

In the first method, the MLE can be found in two steps:

(1) For given λ , (6) is the likelihood for a standard least-squares problem. Thus the estimator of σ^2 , is $\sigma^2(\lambda)$.

$$\hat{\sigma}^2 = \mathbf{y}^{(\lambda)'} (\mathbf{I} - \mathbf{H}) \mathbf{y}^{(\lambda)} / n = S(\lambda) / n,$$

where \mathbf{H} is the perpendicular projection operator onto the column space of \mathbf{X} . That is $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

Let $L(\beta, \sigma^2, \lambda | \mathbf{y})$ be the log likelihood evaluated at β, σ^2 , and λ . That is

$$\begin{aligned} & L(\beta, \sigma^2, \lambda | \mathbf{y}) \\ &= -(2\sigma^2)^{-1} (\mathbf{y}^{(\lambda)} - \mathbf{X}\beta)' (\mathbf{y}^{(\lambda)} - \mathbf{X}\beta) - 0.5n \ln(\sigma^2) + \ln J(\lambda; \mathbf{y}). \end{aligned}$$

Let $\hat{\beta}_\lambda$, and $\hat{\sigma}_\lambda^2$ be the maximizers of L for fixed λ . That is, $\hat{\beta}_\lambda = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}^{(\lambda)}$ and $\hat{\sigma}_\lambda^2 = \mathbf{y}^{(\lambda)'} (\mathbf{I} - \mathbf{H}) \mathbf{y}^{(\lambda)} / n$, where \mathbf{H} is defined as above.

Denote $L(\hat{\beta}_{\lambda_0}, \hat{\sigma}_{\lambda_0}^2, \lambda_0 | \mathbf{y})$ by $L_{\max}(\lambda_0)$. That is,

$$L_{\max}(\lambda_0) = -0.5n - 0.5n \log (\hat{\sigma}_{\lambda_0}^2) + \ln J(\lambda_0; \mathbf{y}).$$

(2) Now we can get the maximizer $\hat{\lambda}$ by plotting $D(\lambda_0)$ vs λ , where $D(\lambda_0) = 2 L_{\max}(\hat{\lambda}) - 2 L_{\max}(\lambda_0)$. The minimum value of $D(\lambda_0)$ corresponding to $\hat{\lambda}$. A large sample 95% confidence interval for λ can be obtained by solving $2 L_{\max}(\hat{\lambda}) - 2 L_{\max}(\lambda_0) = 3.84$.

We can also calculate $\hat{\lambda}$ precisely by setting the derivative with respect to λ to zero. With transformation (2), we can get

$$\frac{d L_{\max}(\lambda)}{d \lambda} = \frac{-n \mathbf{y}^{(\lambda)'} (\mathbf{I}-\mathbf{H}) \mathbf{u}^{(\lambda)}}{\mathbf{y}^{(\lambda)'} (\mathbf{I}-\mathbf{H}) \mathbf{y}^{(\lambda)}} + \frac{n}{\lambda} + \sum \log y_i = 0. \quad (7)$$

Now, solve for λ .

Here $\mathbf{u}(\lambda)$ is the vector of $\{\lambda^{-1} y_i^\lambda \log y_i\}$. In (7), the numerator is the residual sum of products from the analysis of covariance of $\mathbf{y}^{(\lambda)}$ and $\mathbf{u}^{(\lambda)}$.

If we use normalized transformation, $\mathbf{z}^{(\lambda)} = \mathbf{y}^{(\lambda)} / \{J(\lambda; \mathbf{y})\}^{1/n}$, then

$$L_{\max}(\lambda) = -0.5 n \log \hat{\sigma}^2(\lambda; \mathbf{z})$$

and

$$\hat{\sigma}^2(\lambda; \mathbf{z}) = \mathbf{z}^{(\lambda)'} (\mathbf{I}-\mathbf{H}) \mathbf{z}^{(\lambda)} / n = S(\lambda; \mathbf{z}) / n.$$

$S(\lambda; \mathbf{z})$ is called the residual sum of squares of $\mathbf{z}^{(\lambda)}$. Because the maximized likelihood is proportional to $\{S(\lambda; \mathbf{z})\}^{-n}$, so the MLE is gained by minimizing $S(\lambda; \mathbf{z})$ with respect to λ .

For the transformation in (2), $\mathbf{z}^{(\lambda)} = \mathbf{y}^{\lambda-1} / \lambda m(\mathbf{y})^{\lambda-1}$, where $m(\mathbf{y})^{\lambda-1}$ is the geometric mean of \mathbf{y} 's.

For the transformation with shifted location, (3),

$$\mathbf{z}^{(\lambda)} = \{ (\mathbf{y} + \lambda_2)^{\lambda_1 - 1} \} / \lambda \{ \text{gm}(\mathbf{y} + \lambda_2) \}^{\lambda_1 - 1},$$

where $\text{gm}(\mathbf{y} + \lambda_2)$ is the sample geometric mean of $(\mathbf{y} + \lambda_2)$'s.

4.2 Bayesian Method:

Now let's focus on the second method, Bayesian analysis. Use $\nu = n - \text{rank}(X)$ to represent the degrees of freedom for residual. When conducting an analysis of variance on $\mathbf{y}^{(\lambda)}$, use

$$s_{\lambda}^2 = \mathbf{y}^{(\lambda)'} (\mathbf{I} - \mathbf{H}) \mathbf{y}^{(\lambda)} / \nu$$

to represent the residual mean square. Given β , σ^2 , λ , the conditional pdf of likelihood is

$$p(\mathbf{y} | \theta, \sigma^2, \lambda) =$$

$$(2\pi)^{-n/2} \sigma^{-n} \exp\{-(2\sigma^2)^{-1} * [\nu s_{\lambda}^2 + (\hat{\beta}_{\lambda} - \beta)' X' X (\hat{\beta}_{\lambda} - \beta)]\} J(\lambda; \mathbf{y}).$$

where $\hat{\beta}_{\lambda}$ is the least-square estimator of β for a certain λ .

Let $m(\mathbf{y})$ be the geometric mean: $m(\mathbf{y}) = (\prod_{i=1}^n y_i)^{1/n}$. Then $J(\mathbf{y}, \lambda) = m(\mathbf{y})^{n(\lambda-1)}$. Box and Cox (1964) suggest the following prior density for β, σ, λ is

$$h(\beta, \sigma, \lambda) = p_0(\lambda) / \sigma m(\mathbf{y})^{k(\lambda-1)},$$

where k is the dimension of β . This is an "empirical Bayes" type prior, because it is a function of the data.

The joint distribution of \mathbf{y} , β , σ , λ is

$$\begin{aligned} p(\mathbf{y}, \beta, \sigma, \lambda) &= p(\mathbf{y} | \beta, \sigma, \lambda) h(\beta, \sigma, \lambda) \\ &= (2\pi)^{-n/2} \sigma^{-(n+1)} \exp\{-(2\sigma^2)^{-1} \times [\nu s^2(\lambda) + (\hat{\beta}_{\lambda} - \beta)' X' X (\hat{\beta}_{\lambda} - \beta)]\} \\ &\quad \times p_0(\lambda) m(\mathbf{y})^{\nu(\lambda-1)}. \end{aligned}$$

To obtain the posterior distribution of λ given \mathbf{y} , first find the joint marginal distribution of \mathbf{y} and λ .

$$p(\mathbf{y}, \lambda) = p_0(\lambda) \int_0^{\infty} \int_{-\infty}^{+\infty} p(\mathbf{y}, \beta, \sigma, \lambda) d\beta d\sigma$$

$$= p_0(\lambda) \int_0^\infty (2\pi\sigma^2)^{-v/2} \sigma^{-1} |\mathbf{X}'\mathbf{X}|^{-1/2} \exp\{-v s_\lambda^2 / 2\sigma^2\} m(\mathbf{y})^{v(\lambda-1)} d\sigma$$

This is obtained by integrating with regard to β . Next, it is

$$= 2^{1/2} \pi^{-v/2} |\mathbf{X}'\mathbf{X}|^{-1/2} m(\mathbf{y})^{v(\lambda-1)} p_0(\lambda) \int_0^\infty (2\sigma^2)^{-v/2-1} \exp\{-v s_\lambda^2 / 2\sigma^2\} d\sigma.$$

Now make change of variable: $\omega = 1/(2\sigma^2)$

$$\rightarrow \sigma = (2\omega)^{-1/2}$$

$$\rightarrow d\sigma = -(2\omega)^{-3/2} d\omega$$

$$\begin{aligned} \rightarrow p(\mathbf{y}, \lambda) &= 2^{-1/2} \pi^{-v/2} |\mathbf{X}'\mathbf{X}|^{-1/2} m(\mathbf{y})^{v(\lambda-1)} p_0(\lambda) \int_0^\infty \exp\{-\omega(v s_\lambda^2)\} \omega^{v/2-1} d\omega \\ &= 2^{-1/2} \pi^{-v/2} |\mathbf{X}'\mathbf{X}|^{-1/2} m(\mathbf{y})^{v(\lambda-1)} p_0(\lambda) (v s_\lambda^2)^{-v/2} \Gamma(v/2) \end{aligned}$$

Therefore the posterior conditional distribution of λ is

$$p(\lambda | \mathbf{y}) = p(\mathbf{y}, \lambda) / p(\mathbf{y}) = K m(\mathbf{y})^{v(\lambda-1)} p_0(\lambda) (s_\lambda^2)^{-v/2},$$

where K is a constant.

Let $\mathbf{z}^{(\lambda)} = \mathbf{y}^{(\lambda)} / m(\mathbf{y})^{(\lambda-1)}$, Then $s_\lambda^2 = s^2(\mathbf{y}; \lambda) = m(\mathbf{y})^{2(\lambda-1)} s^2(\mathbf{z}; \lambda)$ and

$$p(\lambda | \mathbf{y}) = K p_0(\lambda) [s^2(\mathbf{z}; \lambda)]^{-v/2}.$$

The constant, K , must be chosen so that

$$\int_{-\infty}^{+\infty} K p_0(\lambda) [s^2(\mathbf{z}; \lambda)]^{-v/2} d\lambda = 1.$$

For example, take $p_0(\lambda)$ to be uniform, then,

$$p(\lambda | \mathbf{y}) = K / [s^2(\mathbf{z}; \lambda)]^{v/2}.$$

Now use numerical integration to find K , plot the posterior density function, $p(\lambda | \mathbf{y})$ to λ . The λ value corresponding to the peak is what we want.

Find L and U such that

$$\int_L^U p(\lambda | \mathbf{y}) d\lambda = 0.95.$$

Then (L, U) is a 95% Bayesian Confidence interval.

5 Example: A Biological Experiment using a 3×4 Factorial Design with Replication.

Box and Cox (1964) included the following example and analyzed it. This experiment is a 3×4. The factors are (a) three poisons and (b) 4 treatments. Each cell is randomly assigned 4 animals. Table 1 shows their survival times.

Table 1: Survival times(unit, 10 hr) of animals

<i>Poison</i>	<i>Treatment</i>			
	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>
1	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
2	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
3	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

The residual is obtained by reducing the row and column effect from the original data. No interactions should be included. Then the model is: $\mu_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij}$, where $i=1,2,3,4$; $j=1,2,3$. The

maximum likelihood and posterior distribution are functions of the residual sum of squares for $z^{(\lambda)}$ which is denoted by $s^2(z; \lambda)$, where $z^{(\lambda)}$ is the standardized Box-Cox transformation.

Table 2 contains $s^2(z; \lambda)$, $D(\lambda_0)$ and $p(\lambda | y)$. The appendix contains the SAS code to calculate these three values. K in $Kp_0(\lambda)$ $[s^2(z; \lambda)]^{-v/2} = p(\lambda | y)$ is a constant, which equals the reciprocal of the area under the curve $Y = [s^2(z; \lambda)]^{-v/2}$, because it is reasonable to select $p_0(\lambda)$ as a Uniform distribution.

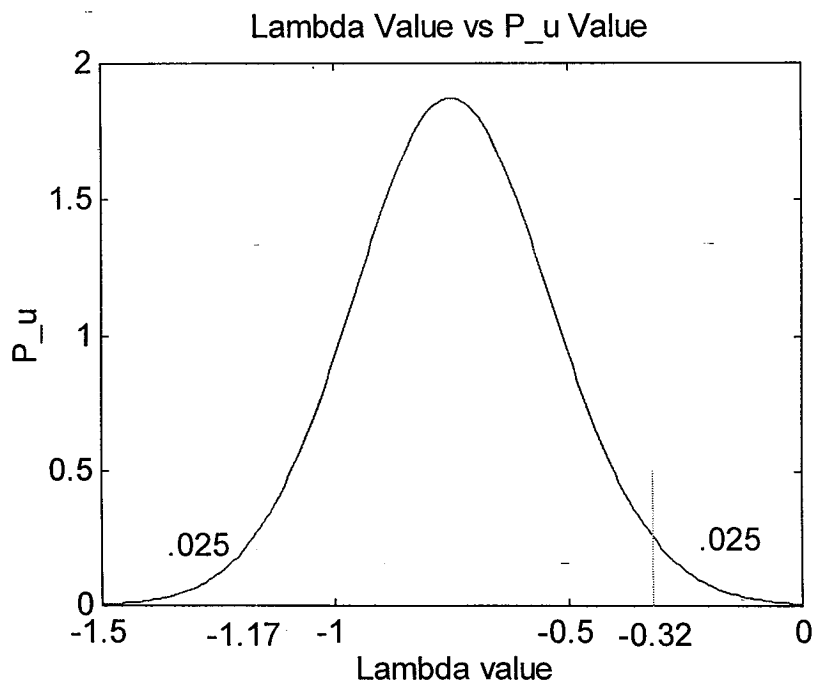
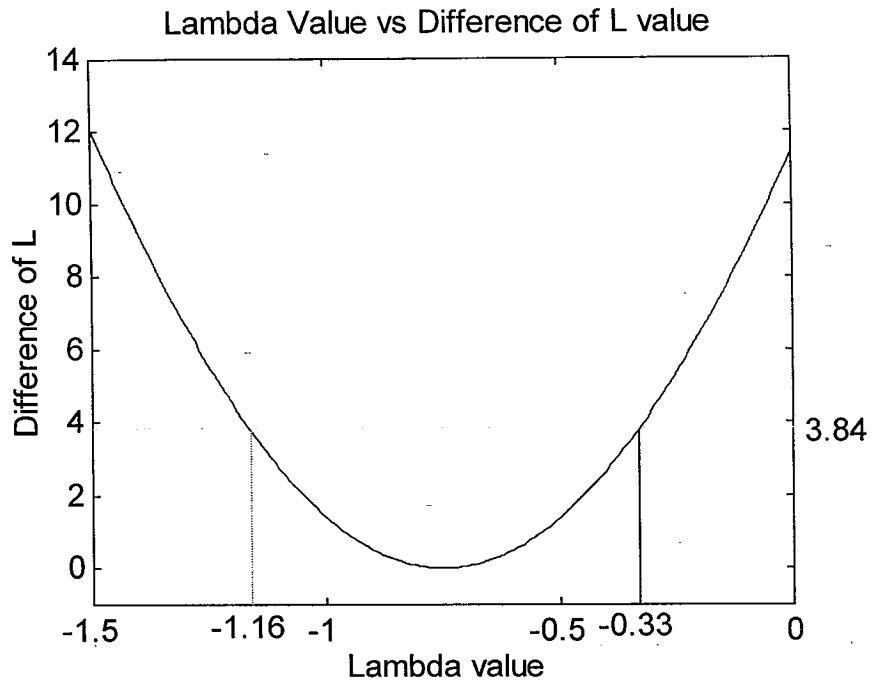
The enclosed Matlab code will calculate K value, the maximizer $\hat{\lambda}$, and the exact lower and upper confidence limits. Here $\hat{\lambda} = -0.7502$, and the CI is $(-1.1675, -0.3220)$. I got $8.7136E-11$ for K , which is a little different with the value Box and Cox got: $8.66E-11$. I chose the former one to do my other calculations and I think the Box-Cox result is not as accurate as mine.

By plotting $D(\lambda_0)$ and $p(\lambda | y)$ vs λ , it yields that in maximized likelihood method, the optimal value is about -0.75 . According to $D(\lambda_0) = 2L_{\max}(\hat{\lambda}) - 2L_{\max}(\lambda) < 3.84$, the approximate 95% confidence interval is $(-1.16, -0.33)$. In Bayes's method, the posterior distribution $p(\lambda | y)$ is approximately normal: $N(-0.75, 0.22)$. The 95% confidence interval is $(-1.17, -0.32)$.

Table 2

λ	$s^2(z; \lambda)$	$D(\lambda_0)$	λ	$s^2(z; \lambda)$	$D(\lambda_0)$
1.0	1.0509	56.76	-1.0	0.3331	1.61
0.5	0.6345	32.51	-1.2	0.3586	5.14
0.0	0.4239	13.08	-1.4	0.4007	10.49
-0.2	0.3752	7.30	-1.6	0.4625	17.39
-0.4	0.3431	3.05	-2.0	0.6639	34.71
-0.6	0.3258	0.57	-2.5	1.1331	60.52
-0.8	0.3225	0.06	-3.0	2.0489	88.82

λ	$p(\lambda \mathbf{y})$	λ	$p(\lambda \mathbf{y})$
0.0	0.01	-0.8	1.82
-0.1	0.02	-0.9	1.42
-0.2	0.08	-1.0	0.92
-0.3	0.26	-1.1	0.47
-0.4	0.49	-1.2	0.19
-0.5	0.94	-1.3	0.07
-0.6	1.46	-1.5	0.01
-0.7	1.82		



6 References

1. Anscombe, F. J. (1948). The transformation of Poisson, binomial and negative binomial data. *Biometrika*, **35**, 246-54.
2. Bartlett, M. S. (1936). The square-root transformation in analysis of variance. *J. Roy. Stat. Soc. Suppl.* **3**, 68-78.
3. Beall, G. (1942). The transformation of data from entomological field experiments so that the analysis of variance becomes applicable. *Biometrika*, **29**, 243-62.
4. Bickel, P. J. and Doksum, K. A. (1981). An analysis of transformations revisited. *Journal of the American Statistical Association*, **76**, 296-311.
5. Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations. *J. Roy. Stat. Soc. B*, **26**, 211-52.
6. Cox, D. R. (1964). Some applications of exponential ordered scores. *J. Roy. Stat. Soc. B*, **26**, 103-10.
7. Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Phil. Trans. Roy. Soc. London*, **222**, 309-68.
8. Fisher, R. A. (1925). *Statistical Methods for Research Workers* (1st edition). Griffin, London.
9. Fisher R. A. (1930). The distribution of gene ration for rare mutations. *Proc. Roy. Soc. Edinburgh*, **50**, 205-20.
10. Fisher R. A. (1954). The analysis of variance with various binomial transformations. *Biometrics*, **10**, 130-51.

11. Fisher R. A. and Yates, F. (1957). *Statistical Tables for Biological Agricultural and Medical Research* (5th edition). Oliver and Boys Edinburgh.
12. Freeman, M. F. and Tukey, J. W. (1950). Transformation related to the angular and the square root. *Ann. Math. Statist*, **21**, 607-11.
13. Hoyle, M.H. (1973). Transformations -- An introduction and a bibliography. *International Statistical Review*, **41**, 203-223.
14. John, J. A. and Draper, N. R. (1980). An alternative family of transformations. *Applied Statistics*, **29**, 190-197.
15. Kihlberg, J. K., Herson, J. H. and Schotz, D. E. (1967). Square root transformations revisited. Presented to American Statistical Association Annual Meeting.
16. Manly, B. F. (1976). Exponential data transformation. *The Statistician*, **25**, 37-42.
17. Sakia, R. M. (1992). The Box-Cox transformation technique: A review. *The Statistician*, **41**, 169-178.
18. Tukey, J. W. (1957). The comparative anatomy of transformations. *Annals of Mathematical Statistics*, **28**, 602-632.
19. Wilson, E. B. and Hilferty, M. M. (1931). The distribution of chi-square. *Proc. Nat. Acad. Sci.* **17**, 684-8.

7 Appendix

***** SAS code *****

```
title1 "Likelihood Evaluation of SSM";
title2 "Using different lamda values: -3.0 to 1.0";
options ls=76 ps=66;
```

```
data one;
  do poison='aaa', 'bbb', 'ccc';
    do treat='a', 'b', 'c', 'd';
      do repeat=1 to 4;
        input survtime @@;
        output;
      end;
    end;
  end;
```

```
cards;
0.31 0.45 0.46 0.43 0.82 1.10 0.88 0.72
0.43 0.45 0.63 0.76 0.45 0.71 0.66 0.62
0.36 0.29 0.40 0.23 0.92 0.61 0.49 1.24
0.44 0.35 0.31 0.40 0.56 1.02 0.71 0.38
0.22 0.21 0.18 0.23 0.30 0.37 0.38 0.29
0.23 0.25 0.24 0.22 0.30 0.36 0.31 0.33
;
run;
```

```
proc iml;
  /* Input X matrix. */
  X = { 1 1 0 0 1 0 0 0,
        1 1 0 0 1 0 0 0,
        1 1 0 0 1 0 0 0,
        1 1 0 0 1 0 0 0,
        1 1 0 0 0 1 0 0,
        1 1 0 0 0 1 0 0,
        1 1 0 0 0 1 0 0,
        1 1 0 0 0 1 0 0,
        1 1 0 0 0 1 0 0,
        1 1 0 0 0 0 1 0,
        1 1 0 0 0 0 1 0,
        1 1 0 0 0 0 1 0,
        1 1 0 0 0 0 1 0,
        1 1 0 0 0 0 0 1,
```

```

1 1 0 0 0 0 0 1,
1 1 0 0 0 0 0 1,
1 1 0 0 0 0 0 1,
1 0 1 0 1 0 0 0,
1 0 1 0 1 0 0 0,
1 0 1 0 1 0 0 0,
1 0 1 0 1 0 0 0,
1 0 1 0 0 1 0 0,
1 0 1 0 0 1 0 0,
1 0 1 0 0 1 0 0,
1 0 1 0 0 1 0 0,
1 0 1 0 0 0 1 0,
1 0 1 0 0 0 1 0,
1 0 1 0 0 0 1 0,
1 0 1 0 0 0 1 0,
1 0 1 0 0 0 0 1,
1 0 1 0 0 0 0 1,
1 0 1 0 0 0 0 1,
1 0 1 0 0 0 0 1,
1 0 0 1 1 0 0 0,
1 0 0 1 1 0 0 0,
1 0 0 1 1 0 0 0,
1 0 0 1 1 0 0 0,
1 0 0 1 0 1 0 0,
1 0 0 1 0 1 0 0,
1 0 0 1 0 1 0 0,
1 0 0 1 0 1 0 0,
1 0 0 1 0 0 1 0,
1 0 0 1 0 0 1 0,
1 0 0 1 0 0 1 0,
1 0 0 1 0 0 1 0,
1 0 0 1 0 0 0 1,
1 0 0 1 0 0 0 1,
1 0 0 1 0 0 0 1,
1 0 0 1 0 0 0 1 };

```

```

G = ginv(t(X)*X); /* Calculate generalized inverse*/
GX = G*t(X); /* To save calculation. */
invK = 0;

```

```

RESET AUTONAME;
USE one;
READ ALL INTO DATA;
survtime=DATA(1,2);

```

```

/* Calculate the Jacobian. */
logy = log(survtime);

```

```

jacobian = exp(sum(logy)/48);
/* Following calculate L value under maximizer -0.75 */
z0 = (survtime##(-0.75)-1)/(-0.75*jacobian##(-1.75));
beta0 = GX * z0;
SSE0 = SSQ(z0 - X*beta0);
Lmax0 = -24 * log(SSE0)+92.91; /* n=48 */

do i = 1 to 41;
    lamda = (i-31)/10;
    if lamda=0 then z_i=logy*jacobian;
    else z_i = (survtime##lamda-1)/(lamda*jacobian##(lamda-1));
    /* Calculate SSE here. */
    beta_i = GX * z_i;
    SSE_i = SSQ(z_i - X*beta_i);
    Lmax_i = -24 * log(SSE_i)+92.91; /* n=48 */
    /* number of independent component is 42 */
    D_i = 2*( Lmax0 - Lmax_i); /* calculate the difference */
    pu_i = 0.871*10**(-10)*(SSE_i)**(-21); /* It is p(lamda given y). */
    /* I used K value from Matlab programming here. */
    print, "Value of SSE ", lamda SSE_i D_i pu_i ;
end;
quit;

run;

```

***** Matlab code *****

Drawing the two plots:

Drawing 'Lambda Value vs Difference of L value' plot:

```

const=0;
lamhat=fmin('pdf_bayes',-3,3,options,l_H,y,G,nu,const)
Lmax=log(-pdf_bayes(lamhat,l_H,y,G,nu,const));
lamb=[-1.5:.01:0]';
unit=lamb.^0;
D=[];
for j=1:length(lamb)
    lam=lamb(j);
    d=2*(Lmax-log(-pdf_bayes(lamb(j),l_H,y,G,nu,const)));
    D=[D;d];
end
c = [-1.5, 0, -1, 14];
yrange = [-1:.1:3.84]';

```

```

unit2 = yrange.^0;
plot(lamb,unit*3.84,'-',lamb,D,'-',unit2*(-1.16),yrange,'-',unit2*(-0.33),yrange,'-')
axis(c)
xlabel('Lambda value')
ylabel('Difference of L')
gtext('3.84')
gtext('-1.16')
gtext('-0.33')
title('Lambda Value vs Difference of L value')

```

To construct the pdf plot:

```

const=new_const;
lamb=[-1.5:.01:0]';
D=[];
for j=1:length(lamb)
    lam=lamb(j);
    d=-pdf_bayes(lamb(j),l_H,y,G,nu,const);
    D=[D;d];
end
yrange33 = [0:.1:.5]';
unit33 = yrange33.^0;
plot(unit33.*(-1.17),yrange33,'-',lamb,D,'-',unit33.*(-0.32),yrange33,'-')
xlabel('Lambda value')
ylabel('P_u')
gtext('-1.17')
gtext('-0.32')
title('Lambda Value vs P_u Value')
gtext('.025')
gtext('.025')

```

To calculate K value, lambda_hat, and confidence interval:

program find_k.m

```

y = [0.31 0.45 0.46 0.43 0.82 1.10 0.88 0.72 0.43 0.45 0.63 0.76 ...
0.45 0.71 0.66 0.62 0.36 0.29 0.40 0.23 0.92 0.61 0.49 1.24 0.44 ...
0.35 0.31 0.40 0.56 1.02 0.71 0.38 0.22 0.21 0.18 0.23 0.30 0.37 ...
0.38 0.29 0.23 0.25 0.24 0.22 0.30 0.36 0.31 0.33 ]';
G=exp(sum(log(y))/48);
b=kron(ones(3,1),kron([1 2 3 4]',ones(4,1)));
a=kron([1 2 3]',ones(16,1));
n=length(y);

```

```

X=[ones(n,1) dummyvar(a) dummyvar(b)];

% For a model with interaction, use
% X=dummyvar((a-1)*4+b);
%

H = X*pinv(X'*X)*X';
I_H = speye(n)-sparse(H);
nu=n-rank(full(X));
options=foptions;
const=0;
lamhat=fmin('pdf_baye',-3,3,options,I_H,y,G,nu,const)

## lamhat =

## -0.7502

const=pdf_baye(lamhat,I_H,y,G,nu,const);
const=-log(-const);
Tol=[1.e-10 0]';
for a=5:5
    K = -1./quad8('pdf_baye',lamhat-a,lamhat+a,Tol,[],I_H,y,G,nu,const);
    new_const=const+log(K);
    disp(['Limits of Intergration are'])
    disp([lamhat-a lamhat+a])
    disp(['Value of K is'])
    disp(exp(new_const))
end;

## Limits of Intergration are
## -5.7502 4.2498

## Value of K is
## 8.7136e-011

const=new_const;
Hi=fmin('CI_bayes',lamhat,lamhat+1,options,I_H,y,G,nu,const,lamhat,.975)

## -0.3682 0.9603

```

```
## -0.1321 0.9974
## -0.5141 0.8622
## -0.2528 0.9883
## -0.2967 0.9809
## -0.3069 0.9787
## -0.3303 0.9728
## -0.3222 0.9750
## -0.3215 0.9751
## -0.3220 0.9750
## -0.3221 0.9750
## -0.3220 0.9750
## -0.3220 0.9750
```

```
## Hi =
```

```
## -0.3220
```

```
Low=fmin('Cl_bayes',lamhat-1,lamhat,options,l_H,y,G,nu,const,lamhat,.025)
```

```
## -1.3682 0.0021
## -1.1321 0.0362
## -0.9862 0.1318
## -1.2460 0.0102
## -1.1306 0.0367
## -1.1839 0.0209
```



```

## -1.2076 0.0160
## -1.1745 0.0232
## -1.1583 0.0276
## -1.1680 0.0249
## -1.1678 0.0249
## -1.1674 0.0250
## -1.1675 0.0250
## -1.1675 0.0250
## -1.1675 0.0250
## Low =
## -1.1675

```

```

## Program pdf_baye.m (Only 8 char is permitted in name).

```

```

function pdf=pdf_baye(lam,l_H,y,G,nu,const)
n=length(y);
pdf=[];
m=length(lam);
for j=1:m
if lam(j) == 0
z = log(y) * G;
else z = ((y.^lam(j)-1)/(lam(j)*G^(lam(j)-1)));
end
s2= (z^l_H*z);
lnpdf=-nu/2*log(s2) + const;
pdf=[pdf; -exp(lnpdf)];
end

```

```

## Program CI_bayes.m

```

```

function P = CI_bayes(upper_limit,l_H,y,G,nu,const,lamhat,p)
Tol=[1.e-10 0]';
options=foptions;

```

```
t=-quad8('pdf_bayes',lamhat-5,upper_limit,Tol,[],l_H,y,G,nu,const);  
disp([upper_limit t])  
P=(t-p)^2;
```