J. Chosun Natural Sci.
Vol. 13, No. 2 (2020) pp. 76 – 82
https://doi.org/10.13160/ricns.2020.13.2.76

Box-Cox Power Transformation Using R

Hoh Yoo Baek[†]

Abstract

If normality of an observed data is not a viable assumption, we can carry out normal-theory analyses by suitable transforming data. Power transformation by Box and Cox, one of the transformation methods, is derived the power which maximized the likelihood function. But it doesn't induces the closed form in mathematical analysis. In this paper, we compose some R the syntax of which is easier than other statistical packages for deriving the power with using numerical methods. Also, by using R, we show the transformed data approximately distributed the normal through Q-Q plot in univariate and bivariate cases with some examples. Finally, we present the value of a goodness-of-fit statistic(AD) and its p-value for normal distribution. In the similar procedure, this method can be extended to more than bivariate case.

Keywords : Normality, Box-Cox Power Transformation, R, Q-Q plot

1. Introduction

Although the assumption of normality of the data is not satisfied, ignoring it and performing data analysis as if it were a normal distribution can lead to incorrect results. Methods to solve this nonnormality are shown in Song et al.(2009) and Yoo et al.(2006). In addition, by properly transforming the original data, nonnormal data can be made closer to normality. In particular, data transformation is essential for data attributes to show normality using the square root of the count, the logarithm of the proportion, and Fisher's Z-transformation of the correlation. With this properly transformed data, we can assume the normality theory and perform the analysis. Depending on the distribution of the data, appropriate transformation methods can be considered (Johnson and Winchern(2007)). It is possible to confirm whether the transformed data was converted to normality through the original data, Q-Q plot, and fitness test. In particular, when the data is positive, a modified method of data transformation by Box and Cox(1964) has been proposed, and since then, some modifications have been made by Andrews and Gnanadesikan(1971) and Hernandez and Johnson(1980) and Yeo and Johnson(2000). They showed the transformation methods from complex multivariate data with more univariate. However, the power required for these transformations cannot derive a mathematically closed form. Only approximate values can be obtained using a computer. In particular, it is easily calculated using a minitab macro. Examples of minitab macro usage are shown in Albert(1996), Berry(1996), Kim et al.(2001), and Lee and Baek(2006) are relatively easy to write syntax in statistical calculations than other statistical packages.

In this paper, it is derived the power of data transformation using R and checked whether the transformed data has normality using Q-Q plots of normal distribution and Anderson-Daring's normality test statistic (AD). In Section 2, the power for Box-Cox transformation from univariate data is derived using R program and shown an example using it. And then, it can be shown that the tranformed data has normality using the Q-Q plot and goodness-of-fit test statistic for a normal distribution. In addition, it is derived the power using a R-program that maximizes the likelihood of the joint bivariate normal distribution with respect to the given bivariate data and an example of its use is presented in Section 3. Also, the normality of the transformed data in the same way as in Section 2 was confirmed by the Q-Q plot including the goodness-of-fit test statistic for a chi-square distribution with 2 degrees of freedom. Finally, conclusions and suggestions are made in Section 4.

Professor, Division of Big Data Financial Statistics, Wonkwang University, #460, Iksandaero Iksan, Jeonbuk 54538, Korea

[†]Corresponding author : hybaek@wku.ac.kr

(Received : June 11, 2020, Revised : June 14, 2020,
Accepted : June 17, 2020)

2. R-program and It's Application for Box-Cox Transformation of Univariate Data

This section shows the Box-Cox transformation form of univariate and positive data and makes R program for this case. In addition, using the derived power, we can transform the raw data to new data and determine whether or not the normality of the transformed data is improved using the Q-Q plot.

Box and Cox(1964) and Hernandez and Johnson(1980) presented the following power transformation.

$$x^{(\lambda)} = \begin{cases} \dfrac{x^{\lambda}-1}{\lambda} & \lambda \neq 0 \\ \ln x & \lambda = 0 \end{cases} \tag{1}$$

where λ is continuous and $x > 0$. Given the observations $x_1, x_2, ..., x_n$, the Box-Cox's solution for the choice of an appropriate power λ is the solution that maximizes the following expression

$$l(\lambda) = -\frac{n}{2}\ln\left[\frac{1}{n}\sum_{j=1}^{n}\left(x_j^{(\lambda)}-\overline{x^{(\lambda)}}\right)^2\right]+(\lambda-1)\sum_{j=1}^{n}\ln x_j \tag{2}$$

where $x_j^{(\lambda)}$ is defined in equation (1) and

$$\overline{x^{(\lambda)}} = \frac{1}{n}\sum_{j=1}^{n}x_j^{(\lambda)} = \frac{1}{n}\sum_{j=1}^{n}\left(\frac{x_j^{\lambda}-1}{\lambda}\right) \tag{3}$$

is the arithmetic mean of the transformed observations. It is difficult to find a mathematically closed form for maximizing equation (2). Therefore, in this paper, we presented an R program file that is easy to use syntax as a numerical analysis method to obtain. In addition, by using this program through examples, the transformed data was derived and it's normality was compared with that of the original data through the Q-Q plots and the test statistic.

2.1. R–program for Univariate Box-Cox Transformation

In this section, it is presented a R-program that is easy to use syntax for univariate Box-Cox transformation. The univariate R program (boxcox.r) is as follows.

```
start=0.1
end=0.4
int=.01
lamda=seq(start, end, int)
lml=length(lamda)
n=length(x)
L_lam=rep(0, lml)
for (i in 1:lml) {
if(abs(lamda[i])<0.0000001) xlam=log(x)
else xlam=(x^lamda[i]-1)/lamda[i]
L_lam[i]=-n/2*log((sum(xlam^2)-n*mean(xlam)^2)/n)
+(lamda[i]-1)*sum(log(x))
}
L_lam
lamda
```

2.2. An Example Using R-Program for Univariate Box-Cox Power Transformation

The example of the execution of the R program in Section 2.1 is as follows. The example is from Johnson and Winchern(2007).

Table 1. Radiation Data

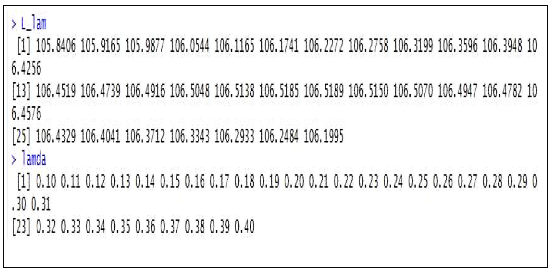
Oven No.	Radiation	Oven No.	Radiation	Oven No.	Radiation	Oven No.	Radiation	Oven No.	Radiation
1	0.15	10	0.10	19	0.01	28	0.09	37	0.20
2	0.09	11	0.07	20	0.40	29	0.08	38	0.30
3	0.18	12	0.02	21	0.10	30	0.18	39	0.30
4	0.10	13	0.01	22	0.05	31	0.10	40	0.40
5	0.05	14	0.10	23	0.03	32	0.20	41	0.30
6	0.12	15	0.10	24	0.05	33	0.11	42	0.05
7	0.08	16	0.10	25	0.15	34	0.30		
8	0.05	17	0.02	26	0.10	35	0.02		
9	0.08	18	0.10	27	0.15	36	0.20		

<Example 2.1> The radiation emissions of 42 microwave oven products from a specific company were investigated. The amount of radiation emitted through the closed doors of these ovens is listed Table 1

Enter 42 observation values of radiation dose into x, input R command.

```
x=c(0.15, 0.09, 0.18, 0.1, 0.05, 0.12, 0.08, 0.05,
0.08, 0.1, 0.07, 0.02, 0.01, 0.1, 0.1, 0.1, 0.02, 0.1,
0.01, 0.4, 0.1, 0.05, 0.03, 0.05, 0.15, 0.1, 0.15, 0.09,
0.08, 0.18, 0.1, 0.2, 0.11, 0.3, 0.02, 0.2, 0.2, 0.3, 0.3,
0.4, 0.3, 0.05)
start=0.1
end=0.4
int=.01
lamda=seq(start, end, int)
lml=length(lamda)
n=length(x)
L_lam=rep(0, lml)
for (i in 1:lml) {
if(abs(lamda[i])<0.0000001) xlam=log(x)
else xlam=(x^lamda[i]-1)/lamda[i]
L_lam[i]=-n/2*log((sum(xlam^2)-n*mean(xlam)^2)/n)
+(lamda[i]-1)*sum(log(x))
}
L_lam
lamda
plot(lamda, L_lam)
xz=(x-mean(x))/sd(x)
xt=x^0.25
xtz=(xt-mean(xt))/sd(xt)
qqnorm(xtz)
qqline(xtz)

##install.packages("nortest")
library(nortest)
qqnorm(x, main="Q-Q Plot of x", xlab="x")
qqline(x)
ad.test(x)
x_1=x^(1/4)
qqnorm(x_1,main="Q-Q Plot of x^lamda",
xlab="x^lamda")
qqline(x_1)
ad.test(x_1)
```



```
> L_lam
 [1] 105.8406 105.9165 105.9877 106.0544 106.1165 106.1741 106.2272 106.2758 106.3199 106.3596 106.3948 10
6.4256
[13] 106.4519 106.4739 106.4916 106.5048 106.5138 106.5185 106.5189 106.5150 106.5070 106.4947 106.4782 10
6.4576
[25] 106.4329 106.4041 106.3712 106.3343 106.2933 106.2484 106.1995
> lamda
 [1] 0.10 0.11 0.12 0.13 0.14 0.15 0.16 0.17 0.18 0.19 0.20 0.21 0.22 0.23 0.24 0.25 0.26 0.27 0.28 0.29 0
.30 0.31
[23] 0.32 0.33 0.34 0.35 0.36 0.37 0.38 0.39 0.40
```

Fig. 1. output of <Example 2.1> using boxcox.r

The results of <Figure 1> above are consistent with Johnson and Winchern(2007). As a result of finding the maximum likelihood value of λ, the maximum likelihood value appears at $\lambda = 1/4$, and the scatter plot of λ is shown in <Figure 2(a)> below. In the above <Figure 2(b)>, it is out of the normality when it is represented by the Q-Q plot including the normality test result (AD=2.101, p-value <0.005) with raw data. Therefore, by the transforming data with approximate $\lambda = 1/4$ we can show the Q-Q plot with the normality test result (AD=0.572 p-value=0.130) in <Figure 2(c)> above. Hence it can be said that the transformed data follow the normal distribution.

3. R-Program and Its Appication for Box-Cox Transformation of Bivariate Data

We can make the approximate the marginal normal data over this bivariate data by deriving the power for each variate as in the previous section. In this case, the problem of obtaining the estimated value of $\lambda' = [\lambda_1, \lambda_2,..., \lambda_n]'$ maximizing the following equation is derived at the same time is the multivariate power transformation method of Box-Cox.

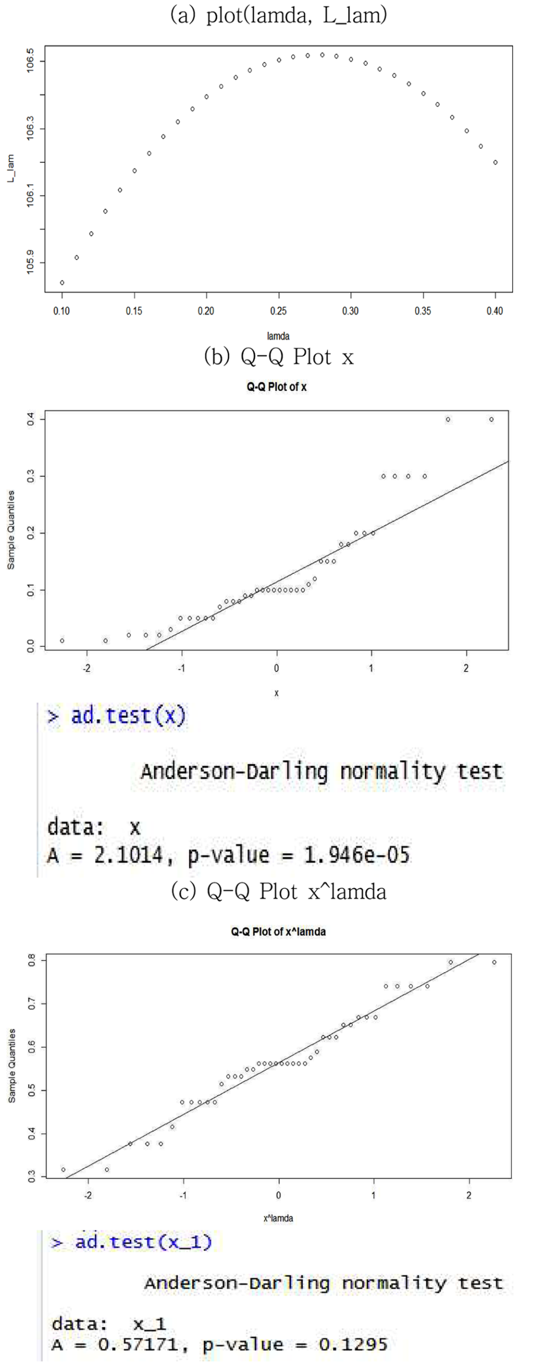
$$l = [\lambda_1, \lambda_2,..., \lambda_p] = -\frac{n}{2}\ln|S(\lambda)| + \sum_{i=1}^{p}(\lambda_i - 1)\sum_{j=1}^{n}\ln x_{ji} \quad (4)$$

where $S(\lambda)$ is the sample covariance matrix computed from the multivariate data below.

$$x_j^{(\lambda)} = l\left[\frac{x_{j1}^{\lambda_1}-1}{\lambda_1}, \frac{x_{j2}^{\lambda_2}-1}{\lambda_2}, \ldots, \frac{x_{jp}^{\lambda_p}-1}{\lambda_p}\right], j = 1, 2,...,n \quad (5)$$

The above formula (4) is presented in Box and Cox(1964) and Andrews and Gnanadesikan(1971) and Hernandez and Johnson(1980).

As in Section 2, it is difficult to find a mathematical closed form of $\lambda' = [\lambda_1, \lambda_2,..., \lambda_n]'$ maximizing formula (4). Therefore, we present a R-program and use this to obtain $\lambda' = [\lambda_1, \lambda_2]'$, in the case of a bivariate data. If the observed bivariate data follows the bivariate normal distribution $d_j = (x_j - \bar{x})'S^{-1}(x_j - \bar{x})$ follows the chi-square distribution with 2 degrees of freedom. Hence we can check the goodness of fit of the distribution through the Q-Q chart including the goodness-of-fit test statistic for the chi-square distribution with 2 degrees of freedom after the data is transformed by using the estimated value of $\lambda' = [\lambda_1, \lambda_2]'$


(a) plot(lamda, L_lam)

(b) Q-Q Plot x

```
> ad.test(x)

        Anderson-Darling normality test

data:  x
A = 2.1014, p-value = 1.946e-05
```

(c) Q-Q Plot x^lamda

```
> ad.test(x_1)

        Anderson-Darling normality test

data:  x_1
A = 0.57171, p-value = 0.1295
```

Fig. 2. likelihood function and Q-Q Plots of <Example 2.1>.

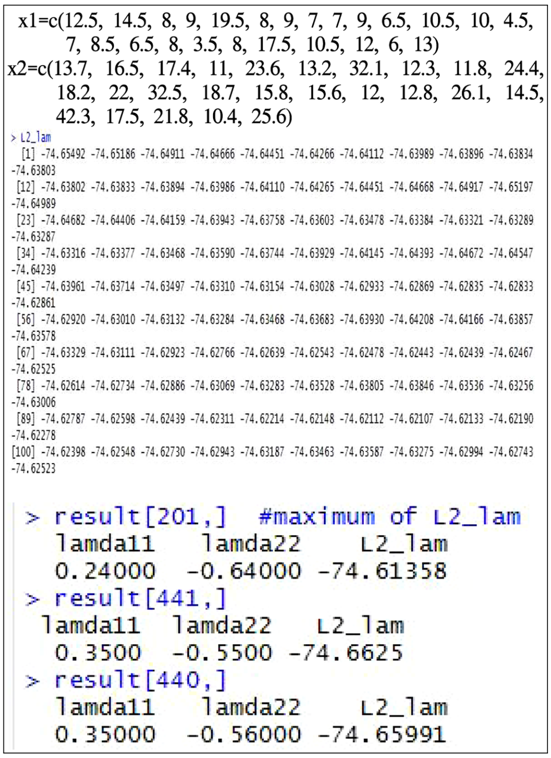
3.1. R-program for Box-Cox Transformation of Bivariate Data

For the bivariate Box-Cox transformation, we present an R program as in Section 2.1. Since the likelihood function for the bivariate case can be obtained, it is easy to derive λ by showing the contour line instead of the scatter plot. The bivariate R program (boxcox2d.r) is as follows.

```
int=0.01
st1=0.15
st2=-0.75
ed1=0.35
ed2=-0.55
lamda1=seq(st1, ed1, int)
m1=length(lamda1)
lamda11=rep(lamda1, each=m1)
lamda2=seq(st2, ed2, int)
m2=length(lamda2)
lamda22=rep(lamda2, m2)
n=length(x1)
L2_lam=rep(0, m1*m2)
for (i in 1:m1) {
if(abs(lamda1[i])<0.0000001) {xlam1=log(x1)}
else {xlam1=(x1^lamda1[i]-1)/lamda1[i]}
for (j in 1:m2) {
if(abs(lamda2[j])<0.0000001) {xlam2=log(x2)}
else {xlam2=(x2^lamda2[j]-1)/lamda2[j]}
s11=(n-1)/n*var(xlam1)
s22=(n-1)/n*var(xlam2)
s12=(n-1)/n*cov(xlam1, xlam2)
dslam=s11*s22-s12^2
L2_lam[(i-1)*m1+j]=-n*log(dslam)/2+(lamda1[i]-1)
*sum(log(x1))+(lamda2[j]-1)*sum(log(x2))
}
}
L2_lam
max(L2_lam)
result=cbind(lamda11, lamda22, L2_lam)
result
result[201,] #maximum of L2_lam
result[441,]
result[440,]
```

Table 2. Data of <Example 3.1>

x_1	x_2	x_1	x_2	x_1	x_2	x_1	x_2	x_1	x_2
12.5	13.7	8.0	13.2	6.5	18.2	8.5	15.6	17.5	42.3
14.5	16.5	9.0	32.1	10.5	22.0	6.5	12.0	10.5	17.5
8.0	17.4	7.0	12.3	10.0	32.5	8.0	12.8	12.0	21.8
9.0	11.0	7.0	11.8	4.5	18.7	3.5	26.1	6.0	10.4
19.5	23.6	9.0	24.4	7.0	15.8	8.0	14.5	13.0	25.6



```
 x1=c(12.5, 14.5, 8, 9, 19.5, 8, 9, 7, 7, 9, 6.5, 10.5, 10, 4.5,
      7, 8.5, 6.5, 8, 3.5, 8, 17.5, 10.5, 12, 6, 13)
x2=c(13.7, 16.5, 17.4, 11, 23.6, 13.2, 32.1, 12.3, 11.8, 24.4,
     18.2, 22, 32.5, 18.7, 15.8, 15.6, 12, 12.8, 26.1, 14.5,
     42.3, 17.5, 21.8, 10.4, 25.6)
> L2_lam
  [1] -74.65492 -74.65186 -74.64911 -74.64666 -74.64451 -74.64266 -74.64112 -74.63989 -74.63896 -74.63834
-74.63803
 [12] -74.63802 -74.63833 -74.63894 -74.63986 -74.64110 -74.64265 -74.64451 -74.64668 -74.64917 -74.65197
-74.64989
 [23] -74.64682 -74.64406 -74.64159 -74.63943 -74.63758 -74.63603 -74.63478 -74.63384 -74.63321 -74.63289
-74.63287
 [34] -74.63316 -74.63377 -74.63468 -74.63590 -74.63744 -74.63929 -74.64145 -74.64393 -74.64672 -74.64547
-74.64239
 [45] -74.63961 -74.63714 -74.63497 -74.63310 -74.63154 -74.63028 -74.62933 -74.62869 -74.62835 -74.62833
-74.62861
 [56] -74.62920 -74.63010 -74.63132 -74.63284 -74.63468 -74.63683 -74.63930 -74.64208 -74.64166 -74.63857
-74.63578
 [67] -74.63329 -74.63111 -74.62923 -74.62766 -74.62639 -74.62543 -74.62478 -74.62443 -74.62439 -74.62467
-74.62525
 [78] -74.62614 -74.62734 -74.62886 -74.63069 -74.63283 -74.63528 -74.63805 -74.63846 -74.63536 -74.63256
-74.63006
 [89] -74.62787 -74.62598 -74.62439 -74.62311 -74.62214 -74.62148 -74.62112 -74.62107 -74.62133 -74.62190
-74.62278
[100] -74.62398 -74.62548 -74.62730 -74.62943 -74.63187 -74.63463 -74.63587 -74.63275 -74.62994 -74.62743
-74.62523

> result[201,]  #maximum of L2_lam
  lamda11   lamda22    L2_lam
  0.24000  -0.64000 -74.61358
> result[441,]
 lamda11  lamda22   L2_lam
  0.3500  -0.5500 -74.6625
> result[440,]
  lamda11   lamda22    L2_lam
  0.35000  -0.56000 -74.65991
```

Fig. 3. output of <Example 3.1> using boxcox2d.r

3.2. An Example Using R-program for Bivariate Box-Cox Power Transformation

It is presented an example of executing the R program boxcox2d.r to perform normality transformation on bivariate data (Johnson and Winchern(2007), P148, P208), and using Q-Q plots of the chi-square distribution the original data and the transfomated data can be checked whether the approach of bivariate normality is approximated.

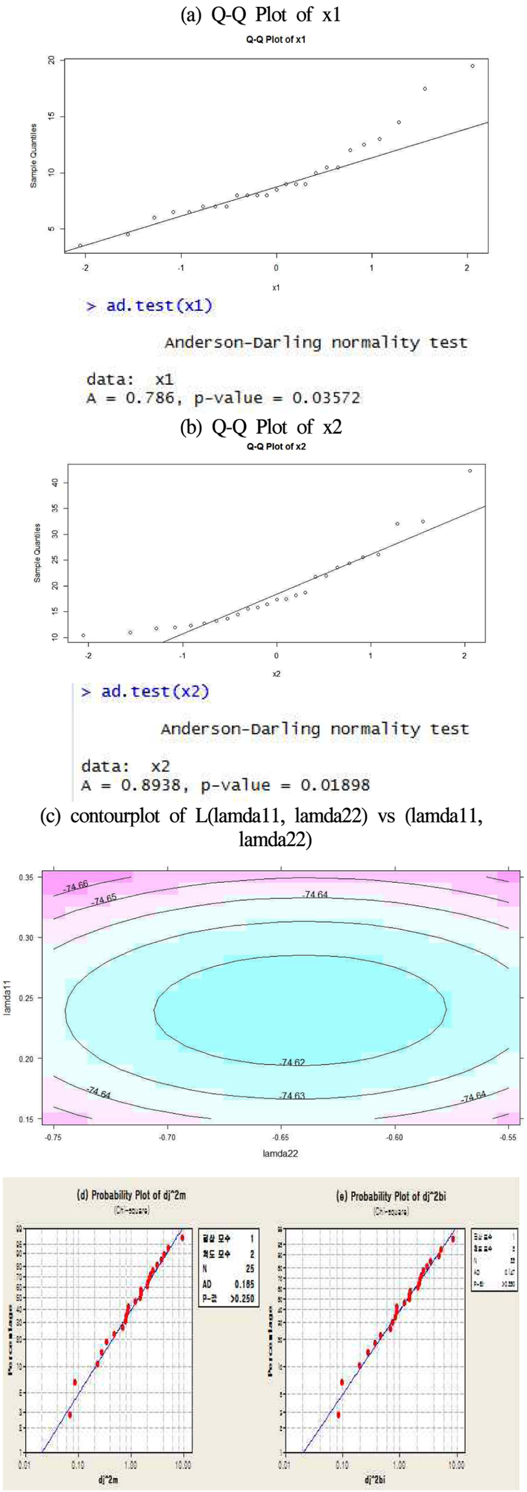
<Example 3.1> In northern climates, roads must be cleared snow quickly following a storm. One measure of storm severity is x_1 = its duration in hours, while the effectiveness of snow removal can be quantified by x_2 = the number of hours crews, men, and machine, spend to clear snow. Here are the results for 25 incidens in Wisconsin.

In <Figure 3>, the maximum likelihood value is shown at $\lambda_1 = 0.24$, $\lambda_2 = -0.64$. As shown in <Figure 4(a), 4(b)> below, the original data is out of the normality with AD=0.786, p-value=0.036, and AD=0.894, P-value=0.019. Therefore, they can be transformed into approximate values λ_1 = 1/4, λ_2 = -2/3, and its normality can be confirmed through the Q-Q plot of the chi-square goodness-of-fit with its degree of freedom for 2 and the p-value of the AD statistic when $d_j = (x_j - \bar{x})'S^{-1}(x_j - \bar{x})$.

In <Figure 4(e)> above, AD=0.147, p-value>0.252, so it can be seen that this result is well approximated for this bivariate normal distribution. Additionally, if two univariate normal distributions are considered as two univariate data, the powers $\lambda_1 = 0.05$, $\lambda_2 = -0.64$ can be obtained by maximizing each likelihood. The above <Figure 4(d)> shows the results of the chi-square plot and the test statistic (AD=0.165, p-value>0.250) when $\lambda_1 = 0$, $\lambda_2 = -2/3$.

4. Conclusions and suggestions

In this paper, the method of transforming to have normality is presented in univariate and bivariate data. In case of biavariate data there is a slight difference in the degree of normality goodness-of-fit between the transformations that maximizes the likelihood functions for each variate (in the form of the marginal normal distribution) and the joint bivariate case itself. In general, it is expected that the normality goodness-of-fit will be stronger in the cases where the correlation of two variates is large and the transformation that maximizes the likelihood is performed. In addition, it can be said that these transformed data follow the normality well when



(a) Q-Q Plot of x1

```
> ad.test(x1)

        Anderson-Darling normality test

data:  x1
A = 0.786, p-value = 0.03572
```

(b) Q-Q Plot of x2

```
> ad.test(x2)

        Anderson-Darling normality test

data:  x2
A = 0.8938, p-value = 0.01898
```

(c) contourplot of L(lamda11, lamda22) vs (lamda11, lamda22)

Fig. 4. likelihood function and Q-Q Plots of <Example 3.1>.

scale transformation according to the attribute of the data is required. The transformation process to observed data in cases of tri and more variates can be performed by modifying as this uni and bivariate case in the paper. In addition, if negative observation values are included (Yeo and Johnson(2000)), the macro presented in the text can be modified and applied as appropriate. Other computer language programs or statistical packages will also be able to write programs for Box-Cox transformation based on the algorithm of this paper.

Acknowledgements

This study was supported by Wonkwang University Research Fund 2019.

References

[1] J. H. Albert, Bayesian Computation Using Minitab, Duxbury Press, Belmont, CA, 1996.

[2] D. F. Andrews, R. Gnanadesikan, and J. L. Warner, Transformation of Multivariate Data, Biometrics, Vol.27, No. 4, pp. 825-840, 1971.

[3] D. A. Berry, Statistics : A Bayesian Perspective, Duzbury Press, Belmont, CA., 1996.

[4] G. E. Box and D. R. Cox, An Analysis of Transformations, Journal of the Royal Statistical Society, Ser. B, Vol. 26, pp. 211-252, 1964.

[5] F. Hernandez and R. A. Johnson, The Large-Sample Behavior of Transformations to Normality, Journal of the American Statistical Association, Vol. 75, No. 352, pp. 855-861, 1980.

[6] B. H. Kim, H. Y. Baek, T. R. Park, H. S. Oh, and I. H. Jang, Bayesian statistical calculation, Free academy.(in Korean), 2001.

[7] H. J. Kim, C. Park, H. Y. Woon, and Y. G. Moon, A Comparative Study on the Parameter Estimation of Bivariate Regular Population Using Variation Coefficients, Journal of Korean Data Analysis Society, Vol. 3, No. 3, pp. 255-265.(in Korean), 2001.

[8] J. M. Lee and H. Y. Baek, Minitab macros for application of Bayes' law, Journal of the Korean Data Analysis Society, Vol. 8, No. 4, pp. 1585-1599.(in Korean), 2006.

[9] Minitab Inc, MINITAB, User's Guide Release 14 for Windows, 2003.

[10] A. Richard, D. Johnson, and W. Winchern, Applied Multivariate Statistical Analysis, 6th edition, Peason Education, Inc, 2007.

[11] J. W. Song, Application of multiple substitution method using latent variable for nonnormal variable, Journal of the Korean Data Analysis Society, Vol. 11, No. 3B, 1377-1387.(in Korean), 2009.

[12] I. K. Yeo and R. A. Johnson, A New Family of Power Transformations to Improve Normality or Symmetry, Biometrika, Vol. 87, No. 4, pp. 954-959, 2000.

[13] S. M. Yoo, G. H. Kim, and D. H. Kim, Stock price normalization process, Journal of the Korean Data Analysis Society, Vol. 8, No. 2, pp. 615-624.(in Korean), 2006.