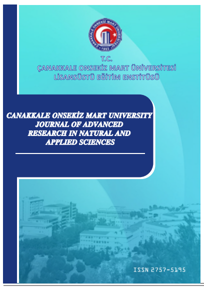
e-ISSN 2757-5195



Çanakkale Onsekiz Mart University
Journal of Advanced Research in Natural and Applied Sciences
Open Access

doi.org/10.28979/jarnas.1037343 2022, Vol 8, Issue 3, Pages: 463-471 dergipark.org.tr/tr/pub/jarnas

Ensemble Based Box-Cox Transformation via Meta Analysis

Muhammed Ali Yılmaz[1], Osman Dag[1,*]

[1] Department of Biostatistics, School of Medicine, Hacettepe University, Ankara, Türkiye.

Article History

Received:	18.12.2021
Accepted:	30.03.2022
Published:	25.09.2022

Research Article

Abstract − Normal distribution has a vital role for the most of statistical methods. Box-Cox power transformation is the most usually applied method when the distribution of data is not normal. In this study, a novel algorithm is proposed assembling different Box-Cox transformation estimates of the well performed six techniques through random effect model in meta analysis. These techniques include the use of goodness-of-fit tests for normality; Anderson–Darling, Lilliefors, Cramer-von Mises, Shapiro–Wilk, Jarque–Bera and Shapiro–Francia tests. For the estimation of Box-Cox parameter, we assemble all possible combinations (63 combinations) of estimates calculated by these six methods. A Monte-Carlo simulation study is implemented to investigate which combination performs better compared to the rest. The simulation study states that the combination of Shapiro–Wilk, Jarque–Bera and Anderson–Darling tests performs well in most of the simulation scenarios constructed under different transformation parameters and sample sizes. In this study, this combination is proposed as ensemble based Box-Cox transformation via meta analysis. The proposed approach is implemented on white blood count data of leukaemia patients which are not normally distributed. Also, the proposed methodology is provided in AID R package with "boxcoxmeta" function for public use.

Keywords − *Data transformation, meta analysis, normality, searching algorithms, statistical software*

1. Introduction

Normal distribution has an essential role for statistical approaches. The reason why the normality assumption has a fundamental role is that it constructs the basis of the approaches, such as t-test and ANOVA. When this assumption is violated, the most widely utilized solution is Box-Cox power transformation (Box and Cox, 1964). Data are not normally distributed in general while analysing data in application. Box-Cox transformation has been carried out in different areas, such as medical studies. Liu et al. (2021) used Box-Cox power transformation to transform the distribution of stay length of type 2 diabetes mellitus patients in public hospital to normal one. Roy et al. (2021) applied Box-Cox transformation for non-normality in brainstem dose-volume histogram points. Singh et al. (2021) applied Box-Cox transformation to normalize absolute insulin dosing. Nelson et al. (2022) used Box-Cox transformation on outcome of cancer-related self-efficacy. In the original article, Box and Cox (1964) used maximum likelihood as an estimation technique of transformation parameter. Dag et al. (2014) proposed a methodology including the usage of an artificial covariate for the estimation of transformation parameter. Rahman and Pearson (2008) and Rahman (1999) used normality tests, Anderson Darling and Shapiro Wilk tests, to estimate the power. Asar et al. (2017) extended the use of normality tests in parameter estimation and included searching algorithm to optimize the parameter. Their study includes well-know seven normality tests; namely, Anderson-Darling, Shapiro-Wilk, Lilliefors (Kolmogorov-Smirnov), Jarque-Bera, Cramer-von Mises, Shapiro-Francia, Pearson chi-square tests. In this study, we propose a novel approach to estimate power transformation parameter. According to the simulation results of the work done by Asar et al. (2017), Pearson chi-square test performed worse compared to the other methods for the parameter estimation. Therefore, we include the other six normality tests in this study. We assemble different Box-Cox transformation estimates through random effect model in meta analysis. Specifically, we assemble all possible combinations (63 combinations) of six methods for the

[1] yilmazmuhammedali@outlook.com.tr
[2] osman.dag@hacettepe.edu.tr
*Corresponding Author

estimation of Box-Cox parameter. We compare all possible combinations via a simulation study. We observe that the combination of Jarque-Bera, Anderson-Darling and Shapiro-Wilk tests performs well in most of the simulation scenarios. Therefore, we propose this combination as ensemble based Box-Cox transformation via meta analysis. The organization of this paper is as follows. Some information regarding Box-Cox power transformation and our proposed methodology are placed in Section 2. The steps of simulation study and its results are placed in Section 3. Moreover, the application of our proposed method is carried out on data of leukaemia patients and the implementation of the proposed approach is placed in Section 3. The article is completed in Section 4.

2. Materials and Methods

Box-Cox power transformation was proposed by Box and Cox (1964). The transformation on y_i (i = 1,2,…,n) is as follows.

$$Z_i = \begin{cases} \frac{(y_i+\lambda_2)^{\lambda}-1}{\lambda}, & \text{if} \quad \lambda \neq 0 \\ \log(y_i+\lambda_2), & \text{if} \quad \lambda = 0 \end{cases} \tag{1}$$

Here, y_i are the data to be transformed, λ is the transformation parameter to be estimated, Z_i are the data on which Box-Cox power transformation is applied and n is the size of sample. λ_2 is the specified fixed value to add each y_i making them positive if $y_i \leq 0$. The power transformation given in Equation (1) is equivalent to

$$Z_i^* = \begin{cases} (y_i+\lambda_2)^{\lambda}, & \text{if} \quad \lambda \neq 0 \\ \log(y_i+\lambda_2), & \text{if} \quad \lambda = 0 \end{cases} \tag{2}$$

since Equation (1) is the linear transformation of Equation (2) (Box and Cox 1964).

The objective of this paper is to propose an ensemble based Box-Cox transformation by assembling the estimates found by Shapiro-Wilk, Jarque–Bera and Anderson–Darling tests via random effect model in meta analysis. Information on estimation process via these tests is available in the work done by Asar et al. (2017). The algorithm of the proposed algorithm can be followed:

i. The possible λ values are specified such as λ = -3, -2.99, -2.98, …, 3.

ii. Any observations in data are not allowed to be negative. If yes, the value is specified to make all observations larger than zero (Box and Cox, 1964).

iii. The estimates via Shapiro-Wilk, Jarque–Bera and Anderson–Darling tests are obtained using the algorithm proposed by Asar et al. (2017).

iv. Standard errors of the estimates are obtained via non-parametric bootstrap.

v. The estimates found in (iii) are assembled using the standard errors obtained in (iv) with random effect model in meta analysis.

The proposed methodology can be reached in **AID** R package with “boxcoxmeta” function. The estimates via Shapiro-Wilk, Jarque–Bera and Anderson–Darling tests are obtained via “boxcoxnc” function in **AID** package. The estimates are assembled via meta package (Balduzzi, 2019). All applications and codes are conducted in R environment (R Development Core Team, 2020). The figures are drawn using ggplot2 R package (Wickham, 2016).

3. Results and Discussion

In this part, we implement a Monte Carlo simulation for the comparison of our proposed method with the other existing methods. The simulation scenarios are provided with their results in this section. Our proposed method is demonstrated on data of leukaemia patients. Brief information is presented together with results. The implementation of the proposed method is provided in the last subsection.

3.1. Monte Carlo Simulation Study

A simulation study is implemented to make a comparison for the estimation performance with our proposed method (OM) with other methods; Shapiro-Francia (SF), Cramer-von Mises (CVM), Shapiro-Wilk (SW), Jarque-Bera (JB), Anderson-Darling (AD), Lilliefors test (LT). The algorithm of simulation scenarios is as follows.

i. Simulate a random data set from N ($\mu = 0$, $\sigma = 5$) with different sample size (n = 20, 30, 50, 100, 500).

ii. If generated variable (Z) includes any non-positive value, add a specified fixed value to make all observations larger than zero.

iii. Make inverse transformation $Z^{(1/\lambda)}$ stated in Equation (2) with Box-Cox parameter (λ= -5, -2, -1, 0, 2, 5).

iv. Estimate λ by all methods.

v. Repeat all steps for 10,000 runs.

When all steps are ended, the performance measures; namely, bias, mean square error (MSE) and standard error (SE) are obtained.

In this section, the performances of methods are investigated through bias, MSE and SE. All results are not reported here for the content integrity, but available at yunus.hacettepe.edu.tr/~osman.dag/supp_materials/ensemble_boxcox.xlsx. We provide the performances in Figure 1 and Table 1.

In general perspective, biases and MSEs get smaller as the magnitude of λ decreases as expected. For example, under the scenario of $n = 50$, absolute biases and MSEs are within a range of 0.005-0.105 and 1.631-2.869 for $\lambda = 5$, respectively; on the other hand, absolute biases and MSEs are found to be in an interval of 0.001-0.043 and 0.266-0.514 for $\lambda = 2$, respectively.

Bias and MSE values become smaller as the number of observations raises. The performances of the methods are similar especially for large sample size. However, as the sample size gets smaller, the differences among the estimation techniques become clear. In most scenarios, our proposed method performs better compared to other estimation techniques. To illustrate, for $n = 30$ and $\lambda = 2$, absolute bias and MSE of our method are 0.001 and 0.449, respectively. Under the same condition, the absolute biases and their MSEs of the other methods are within a range of 0.008-0.054 and 0.412-0.752, respectively.

Table 1

Comparison of our method with six different methods

n	true		λ_{SW}	λ_{AD}	λ_{CVM}	λ_{SF}	λ_{LT}	λ_{JB}	λ_{OM}
	-5	Bias	0.020	-0.103	-0.130	-0.133	-0.061	0.077	-0.001
		MSE	3.287	3.704	4.232	3.391	4.666	3.579	3.440
	-2	Bias	-0.012	-0.073	-0.098	-0.078	-0.087	-0.005	-0.030
		MSE	0.622	0.748	0.914	0.656	1.078	0.748	0.681
	-1	Bias	-0.007	-0.038	-0.051	-0.040	-0.049	-0.006	-0.017
		MSE	0.159	0.194	0.241	0.168	0.287	0.199	0.176
20	0	Bias	0.000	0.000	0.000	0.000	0.001	0.000	0.000
		MSE	0.003	0.003	0.004	0.003	0.005	0.003	0.003
	2	Bias	0.012	0.073	0.098	0.078	0.087	0.005	0.030
		MSE	0.622	0.748	0.914	0.656	1.078	0.748	0.681
	5	Bias	-0.020	0.103	0.130	0.133	0.061	-0.077	0.001
		MSE	3.287	3.704	4.232	3.391	4.666	3.579	3.440

Table 1

Comparison of our method with six different methods (Continued)

n	true		λ_{SW}	λ_{AD}	λ_{CVM}	λ_{SF}	λ_{LT}	λ_{JB}	λ_{OM}
30	-5	Bias	0.041	-0.064	-0.076	-0.097	-0.017	0.105	0.029
		MSE	2.357	2.805	3.339	2.436	3.703	2.569	2.503
	-2	Bias	0.008	-0.041	-0.054	-0.049	-0.038	0.028	-0.001
		MSE	0.412	0.514	0.642	0.432	0.752	0.472	0.449
	-1	Bias	0.003	-0.021	-0.027	-0.025	-0.021	0.013	-0.002
		MSE	0.104	0.130	0.162	0.109	0.190	0.120	0.113
	0	Bias	0.000	0.000	-0.001	-0.001	-0.001	-0.001	-0.001
		MSE	0.002	0.002	0.002	0.002	0.003	0.002	0.002
	2	Bias	-0.008	0.041	0.054	0.049	0.038	-0.028	0.001
		MSE	0.412	0.514	0.642	0.432	0.752	0.472	0.449
	5	Bias	-0.041	0.064	0.076	0.097	0.017	-0.105	-0.029
		MSE	2.357	2.805	3.339	2.436	3.703	2.569	2.503
50	-5	Bias	0.014	-0.074	-0.086	-0.105	-0.026	0.070	0.005
		MSE	1.631	2.070	2.524	1.701	2.869	1.744	1.748
	-2	Bias	0.004	-0.033	-0.042	-0.043	-0.025	0.024	-0.001
		MSE	0.266	0.345	0.435	0.279	0.514	0.290	0.287
	-1	Bias	0.002	-0.017	-0.021	-0.023	-0.014	0.012	-0.001
		MSE	0.067	0.087	0.109	0.070	0.129	0.073	0.072
	0	Bias	-0.001	0.000	0.000	-0.001	0.000	-0.001	-0.001
		MSE	0.001	0.001	0.001	0.001	0.001	0.001	0.001
	2	Bias	-0.004	0.033	0.042	0.043	0.025	-0.024	0.001
		MSE	0.266	0.345	0.435	0.279	0.514	0.290	0.287
	5	Bias	-0.014	0.074	0.086	0.105	0.026	-0.069	-0.005
		MSE	1.631	2.070	2.524	1.701	2.869	1.744	1.748
100	-5	Bias	0.012	-0.034	-0.043	-0.079	-0.021	0.062	0.014
		MSE	0.925	1.265	1.568	0.961	1.856	0.964	0.997
	-2	Bias	0.005	-0.014	-0.019	-0.032	-0.011	0.024	0.005
		MSE	0.149	0.205	0.256	0.155	0.307	0.156	0.161
	-1	Bias	0.002	-0.007	-0.010	-0.016	-0.007	0.010	0.002
		MSE	0.038	0.052	0.065	0.040	0.077	0.039	0.040
	0	Bias	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		MSE	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	2	Bias	-0.005	0.014	0.019	0.032	0.011	-0.024	-0.005
		MSE	0.149	0.205	0.256	0.155	0.307	0.156	0.161
	5	Bias	-0.012	0.035	0.043	0.079	0.021	-0.062	-0.014
		MSE	0.925	1.265	1.568	0.961	1.856	0.964	0.997

Table 1

Comparison of our method with six different methods (Continued)

n	true		λ_{SW}	λ_{AD}	λ_{CVM}	λ_{SF}	λ_{LT}	λ_{JB}	λ_{OM}
500	-5	Bias	-0.012	-0.016	-0.017	-0.052	-0.009	0.008	-0.006
		MSE	0.273	0.400	0.480	0.280	0.571	0.277	0.293
	-2	Bias	-0.005	-0.006	-0.007	-0.021	-0.004	0.003	-0.003
		MSE	0.045	0.065	0.077	0.046	0.091	0.045	0.047
	-1	Bias	-0.003	-0.004	-0.003	-0.011	-0.003	0.000	-0.002
		MSE	0.012	0.017	0.020	0.012	0.023	0.012	0.012
	0	Bias	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		MSE	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	2	Bias	0.005	0.006	0.007	0.021	0.004	-0.003	0.003
		MSE	0.045	0.065	0.077	0.046	0.091	0.045	0.047
	5	Bias	0.012	0.016	0.017	0.052	0.009	-0.008	0.006
		MSE	0.273	0.400	0.480	0.280	0.571	0.277	0.293

When lambda is close to zero, all estimation techniques perform similar. For instance, the absolute bias and MSE values of all tests are lower than 5 × 10^{-3} for all scenarios.

When the data generated under the normal distribution with sample size *n = 20* for *λ = -2, -1, 2*, JB and SW techniques have smaller bias compared to our proposed method.

In brief, our proposed method seems to be more effective than other methods for estimating Box-Cox transformation parameter in most scenarios. However, JB and SW might be preferable when sample size is small and estimated transformation parameter is close to 0.

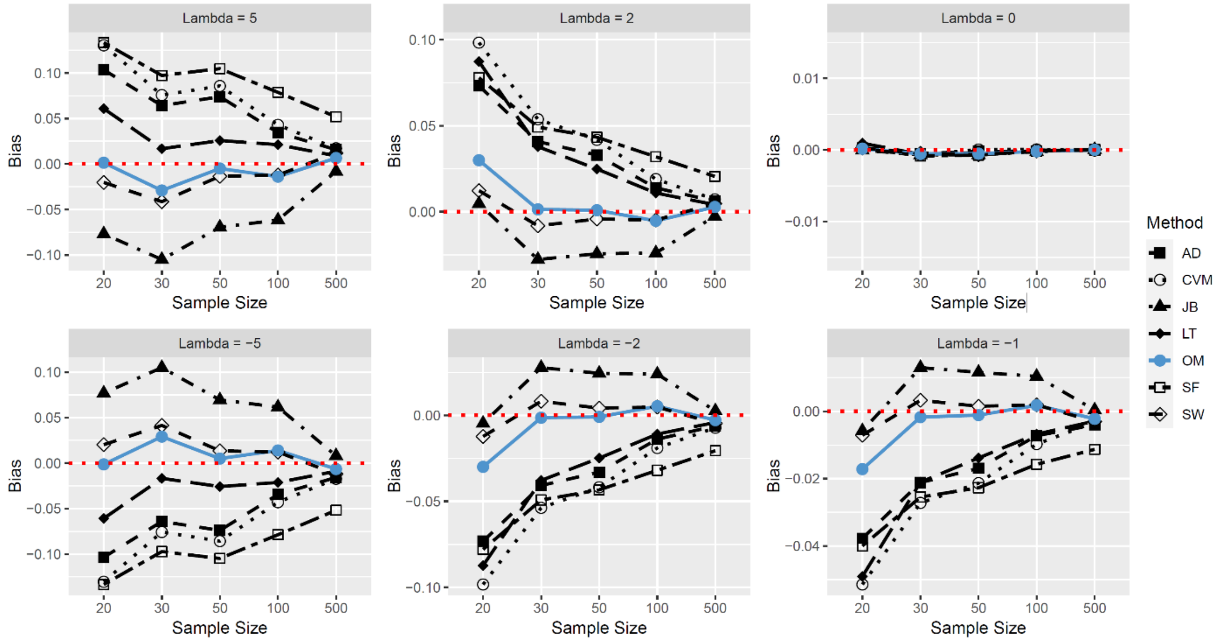


Figure 1. Biases of Methods

3.2. Data of Leukaemia Patients

This dataset involves 33 patients having died from acute myelogenous leukaemia. The data set includes three variables; namely, white blood count, test result and survival time. The data set can be reached in

MASS R package (Venables and Ripley, 2002). For our analysis, the distribution of white blood count (wbc) is examined. This variable shows positively skewed distribution (Figure 2); moreover, the normality tests (Jarque-Bera, Anderson-Darling and Shapiro-Wilk tests) indicate that the distribution of white blood count is non-normal (e.g.. Shapiro-Wilk normality test: p-value $=1.986\times10^{-6}$).

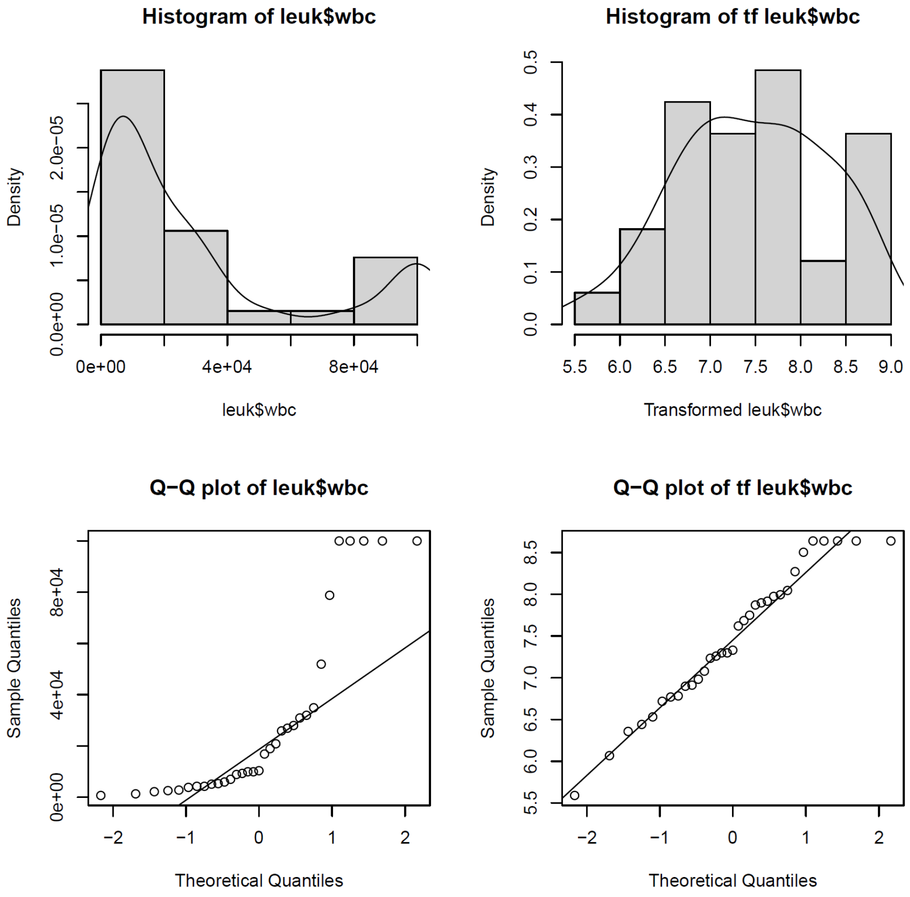


Figure 2. Left: Histogram (upper) and Q-Q (lower) plot of white blood count before Box-Cox transformation; Right: Histogram (upper) and Q-Q plot (lower) of white blood count after Box-Cox transformation.

Our aforementioned method is performed for the estimation of unknown parameter. The estimate is found to be -0.05333092. When we apply Box-Cox transformation with this lambda estimate, we can easily see transformed data are distributed normally (Figure 2). Moreover, the results of normality tests are provided along with p-values in following part.

3.3. Implementation

The proposed methodology is released in "boxcoxmeta" function under **AID** R package. The estimate of λ for white blood count can be acquired with R codes presented below.

```
R> library(AID)
R> library(MASS)

R> set.seed(1)
R> result <- boxcoxmeta(leuk$wbc, lambda = seq(-3,3,0.01), nboot = 100, lambda2 = NULL,
plot = TRUE, alpha = 0.05, verbose = TRUE)

 Box-Cox power transformation via meta analysis
---------------------------------------------------------------------------------------------
```

```
 data : leuk$wbc

 lambda.hat : -0.05251743

 Normality tests for transformed data (alpha = 0.05)
-------------------------------------------------------------------------------------------
        Test                 Statistic       P.Value        Normality
1       Shapiro-Wilk         0.9596797       0.2531485      Not reject
2       Anderson Darling     0.3500850       0.4515463      Not reject
3       Jarque-Bera          0.9444637       0.6236089      Not reject
-------------------------------------------------------------------------------------------
```

Here, lambda.hat is the estimate of λ. After transformation, three normality tests assess the normality of the variable. Since all p-values are larger than alpha, the normality of transformed white blood count is suggested by Shapiro-Wilk, Jarque-Bera and Anderson Darling tests.

In boxcoxmeta function, lambda is the vector of plausible transformation parameter. This vector is fixed to -3:3 (with increment 0.01) as a default. The nboot argument is the number of non-parametric bootstrap samples. Default is set to 100. The lambda2 is the constant to add each value making them positive. The plot option draws a figure given in Figure 2. Default is set to TRUE. The alpha argument is the significance level to check whether normality holds or not after Box-Cox transformation. The alpha argument is set to 0.05 as a default.

Box-Cox transformation is used to transform non-normal variable to a normal one. Therefore, symmetric confidence interval is not appropriate for non-normal data. After transformation, we calculated mean and confidence interval for transformed data. After that, we back transform mean and confidence interval to the original scale. For this reason, asymmetric confidence interval is appropriate for non-normal data. The confInt function is developed and released under **AID** package to obtain asymmetric confidence interval since the original data is asymmetric. This function returns mean, lower confidence limit (LCL) and upper confidence limit (UCL) for original scale of data.

```
R> confInt(result, level = 0.95)

Back transformed data
-------------------------------------------------------------------------
                Mean          2.5%          97.5%
leuk$wbc        13064.64      8148.229      21200.45
-------------------------------------------------------------------------
```

The usage of mean and confidence interval for such a non-normal data set sometimes becomes inappropriate. We reported different usages of mean and confidence interval (CI) in Table 2 to emphasize the difference among them.

In our case, the distribution of white blood count is positively skewed in original scale. Moreover, there exist possible outliers at right tail of the data (Figure 2). Therefore, mean (2.5% - 97.5% CI) of white

Table 2

Difference among the usages of mean and confidence interval

Scale of data	Mean	LCL (2.5%)	UCL (97.5%)
Original data	29165.15	16935.75	41394.56
Transformed data	7.47	7.18	7.76
Back transformed data	13064.64	8148.23	21200.45

blood count moves to right. This does not reflect the central tendency of the data. For transformed data, the distribution of the data is normal and confidence interval is symmetric; however, scale of the data is not interpretable in clinic. For back transformed data, scale of back transformed data is same with original data. Compared to LCL (2.5%), UCL (97.5%) is further to mean since the original data set is positively skewed. Therefore, this usage reflects the central tendency of the data.

4. Conclusion

In this study, a novel technique for the estimation of Box-Cox parameter is proposed. Box-Cox transformation parameter estimates are assembled via random effect model in meta analysis. We assemble the estimates of the well performed six techniques proposed by Asar et al. (2017). These techniques include use of normality tests, Cramer-von Mises, Shapiro–Wilk, Lilliefors, Anderson–Darling, Jarque–Bera and Shapiro–Francia tests. For the estimation of Box-Cox parameter, we assemble not only all estimates by these methods, but also all possible combinations (63 combinations) of estimates calculated by these six methods.

We implement a simulation study. The combination of Shapiro–Wilk, Jarque–Bera and Anderson–Darling tests performs well in most of the simulation scenarios. Therefore, we propose this combination. We call this combination ensemble based Box-Cox transformation via meta analysis. Results show that our proposed method seems to be more effective than other methods for estimating Box-Cox transformation parameter. For an estimate closer to 0, the estimation based on Jarque-Bera and Shapiro-Wilk test might be preferable when sample size is small.

The proposed method is released in **AID** R package under “boxcoxmeta” function. Its implementation is conducted on data of leukaemia patients.

Author’s Contributions

Muhammed Ali Yilmaz: Developing codes, Simulation study, Figures, Table(s), Data analysis, Writing the paper.

Osman Dag: Designing the study, Developing codes, Simulation study, Writing the paper.

Conflicts of Interest

There is no conflict of interest declared by the authors.

References

Asar, O., Ilk, O., & Dag, O. (2017). Estimating Box-Cox power transformation parameter via goodness-of-fit tests. *Communications in Statistics – Simulation and Computation, 46*(1), 91-105. DOI: https://doi.org/10.1080/03610918.2014.957839

Balduzzi, S., Rucker, G., & Schwarzer, G. (2019). How to perform a meta-analysis with R: a practical tutorial. *Evidence-Based Mental Health, 22*, 153-160. DOI: https://doi.org/10.1136/ebmental-2019-300117

Box, G.E.P., & Cox, D.R. (1964). An analysis of transformations (with discussion). *Journal of Royal Statistical Society, Series B (Methodological), 26*(2), 211-243. DOI: https://doi.org/10.1111/j.2517-6161.1964.tb00553.x

Dag, O., Asar, O., & Ilk, O. (2014). A methodology to implement Box-Cox transformation when no covariate is available. *Communications in Statistics – Simulation and Computation, 43*(7), 1740-1759. DOI: https://doi.org/10.1080/03610918.2012.744042

Liu, W., Shi, J., He, S., Luo, X., Zhong, W., & Yang, F. (2021). Understanding variations and influencing factors on length of stay for T2DM patients based on a multilevel model. *Plos One, 16*(3), 1-14. DOI:

https://doi.org/10.1371/journal.pone.0248157

Nelson, D., Law, G.R., McGonagle, I., Turner, P., Jackson, C., & Kane, R. (2022). The Effect of Rural Residence on Cancer-Related Self-Efficacy With UK Cancer Survivors Following Treatment. *The Journal of Rural Health, 38*(1), 28-33. DOI: *https://doi.org/10.1111/jrh.12549*

R Development Core Team (2020). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, Retrieved from http://www.R-project.org.

Rahman, M. (1999). Estimating the Box-Cox transformation via Shapiro-Wilk W statistic. *Communications in Statistics - Simulation and Computation, 28*(1), 223-241. DOI: https://doi.org/10.1080/03610919908813545

Rahman, M., & Pearson, L.M. (2008). Anderson-Darling statistic in estimating the Box-Cox transformation parameter. *Journal of Applied Probability and Statistics, 3*(1), 45-57.

Roy, A., Widjaja, R., Wang, M., Cutright, D., Gopalakrishnan, M., & Mittal, B.B. (2021). Treatment plan quality control using multivariate control charts. *Medical Physics, 48*(5), 2118-2126. DOI: https://doi.org/10.1002/mp.14795

Singh, S.R., Dhanasekara, C.S., Tello, N., Southerland, P., Saleh, A.A., Kesey, J., & Dissanaike, S. (2021). Variations in insulin requirements can be an early indicator of sepsis in burn patients. *Burns, 48*(1), 111-117. DOI: https://doi.org/10.1016/j.burns.2021.02.026

Venables, W.N., & Ripley, B.D. (2002). Modern Applied Statistics with S. Fourth Edition. Springer: New York. ISBN: 978-0-387-21706-2

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag: New York. ISBN: 978-0-387-98141-3