Journal of Bioscience, 19(1), 137–145, 2008

SHORT COMMUNICATION

AN APPLICATION OF BOX-COX TRANSFORMATION TO BIOSTATISTICS EXPERIMENT DATA

Wan Muhamad Amir Wan Ahmad[*1], Nyi Nyi Naing and Nurfadhlina Abd Halim

Jabatan Matematik, Fakulti Sains dan Teknologi, Malaysia, Universiti Malaysia Terengganu (UMT), 21030 Kuala Terengganu, Terengganu, Malaysia.

Abstrak: Teknik penapisan data adalah merupakan satu kaedah yang penting untuk menyemak sifat semula jadi data. Salah satu kaedah untuk tujuan ini adalah Transformasi Box-Cox. Kaedah transformasi Box-Cox diketahui sebagai kaedah yang dapat mentransformasikan data supaya memenuhi andaian regresi linear dan ANOVA. Setelah data ditapis, kaedah-kaedah berparameter boleh diaplikasikan terhadap data-data tersebut. Seperti pekali regresi juga, parameter λ amnya ditentukan dengan menggunakan kaedah kebolehjadian maksimum dengan beranggapan ralat normal tertabur secara homogen (Bartlett 1947). Dalam aplikasi ANOVA bagi tujuan hipotesis dalam eksperimen sains biostatistik, andaian bagi kehomogenan ralat selalunya terabai disebabkan oleh kewujudan kesan skala dan andaian dalam ukuran. Kami menjalankan ujian dengan kaedah transformasi data supaya andaian bagi ANOVA dapat dipenuhi (atau gangguan dapat dikurangkan) dan dapat diguna pakai dalam analisa data eksperimen biostatistik. Dalam penyelidikan ini, kami akan menunjukkan kaedah Box-Cox dengan menggunakan perisian MINITAB.

Kata kunci: Transformasi Box-cox, Parameter λ , ANOVA, Regresi

Abstract: Data screening is the most important technique to check the nature of the data. One of the methods to screen the data is the Box-Cox Transformation. The Box-Cox family of transformation is a well-known approach to make data behave accordingly to assumption of linear regression and ANOVA. After screening the data method, the parametric method can be applied. The regression coefficients, as well as the parameter λ defining the transformation, are generally estimated by maximum likelihood, assuming homoscedastic normal error (Bartlett 1947). In application of ANOVA for hypothesis testing in biostatistics science experiments, the assumption of homogeneity of errors is often violating because of scale effects and the nature of the measurements. We demonstrate a method of transformation data so that the assumptions of ANOVA are met (or violated to a lesser degree) and apply it in analysis of data from biostatistics experiments. In this paper, we will illustrate the use of the Box-Cox method by using MINITAB software.

Keywords: Box-Cox Transformation, Parameter λ , ANOVA, Regression

[*]Corresponding author: wmamir@umt.edu.my

INTRODUCTION

Since the seminal by Box and Cox (1964), the Box-Cox types of power transformation have generated a great deal of interests, both in theoretical work and in practical applications. The Box-Cox family of transformation has become a widely used tool to make data behave accordingly to a linear regression model. Sakia (1992) has given an excellent review of the work relating to this transformation. The response variable, transformed according to the Box-Cox procedure, is usually assumed to be linearly related to its covariates and the errors normally distributed with constant variance.

The data that required an ANOVA application, errors need to be independent, have a normal distribution with zero mean, and be similar between treatments (Andrew 1971). The assumption of homogenous variances often is violated in biological experiments because treatments that result in a change in the mean of a given response variable are often accompanied by changes in error variances (a scale effect). Transformation of the data is usually a useful method of alleviating heterogeneity because it is applicable to all experimental designs analyzed with ANOVA, and conversion of data from interval or ratio values to ranks, as in nonparametric procedures, results in a loss of information. This loss in information is usually reflected by a loss of power of the statistical test. The goal of data transformation is to change the scale by which the data is analyzed so that the variances are not heterogenous. However, the transformation that is most effective in reducing the heterogeneity of variance is often not obvious and found by trial and error.

Box and Cox (1964) presented a family of transformations and a computational technique to select a transformation that will best resolve the problems of non-normality and heterogeneity of error. Despite its power and desirable properties, the Box-Cox transformation apparently is rarely used in the statistical analysis of biostatistics data. In this report, we provide an overview of the Box-Cox data transformation and provide an illustrative example for its application in the analysis of data from a biostatistics experiment. The results from the transformed and untransformed data are discussed.

MATERIALS AND METHODS

The method presented by Box and Cox (1964) is based on the observation that the mean (μ) is often proportional to the standard deviation (σ) of a population such that

$$\sigma \propto \mu^{a}$$

(Damon & Harvey 1987; Montgomery 1991). The purpose of transformation is to raise the data to power λ such that the correlation between the mean and the standard deviation is reduced or eliminated. Box and Cox (1964) provided an

algorithm by which the optimum value for the transformation parameter λ is selected by the method of maximum likelihood. This technique involves performing a series of analyses of variance, for various values of λ, transformed as

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

where,

$$\dot{y} = \ln^{-1}\left(\frac{\sum \ln y}{n} \right)$$

the geometric mean of the observation (y). The analysis of variance for λ that yields the lowest error sums of squares then is used for hypothesis testing (Peltier 1998).

CASE STUDY

This study is done based on the data that has been used by Box and Cox (1964). We used the Box-Cox data after considering the results of their study which is very useful in general statistics and the used of transformation especially. Although the data that has been studied by Box and Cox is presented in this study case, but we will illustrated it by the different way. Table 1 and Table 2 show the full data set of lifetime animal in 3 $\times$ 4 factorial design of experiment with 3 level of poison factor and 4 level of treatment factor.

Table 1: Descriptions of the data

Poison	I - Poison Type I
	II - Poison Type II
	III - Poison Type III
Treatment	A - Treatment Type I
	B - Treatment Type II
	C - Treatment Type III
	D - Treatment Type IV

Table 2: Full dataset of lifetime animal in 3×4 factorial design of 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

(Source : Box and Cox 1964)

Method Applied and Diagnostic Checking of Original Data Sets

The first step that we should do is plotting the normal probability plot to the response variable. Normal probability plot is done to verify whether the data that we study is fulfilling the assumption of normality or not. If there are a sign that show the data is deflecting from the assumption of normality, then the transformation is needed. Figure 1 illustrates the normal probability plot of the data. From the normal probability plot, we can see that the normality assumption is not fulfilled by the response variables. The structure of the plot in Figure 1 displays the deflections in point. Thus, there is enough evidence to say that the normality assumption is contravened in this case.

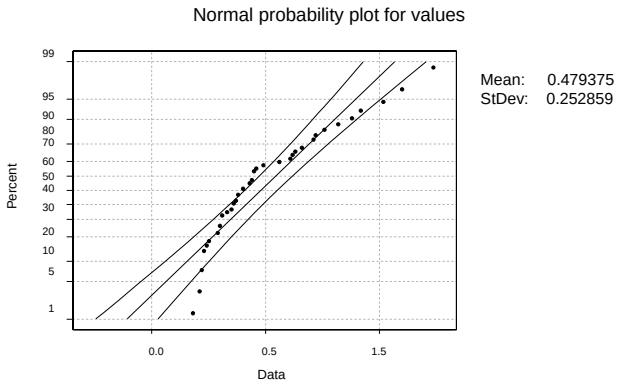

Figure 1: Normal probability plot of the original data

To support this interpretation, let us look at the Figure 2. From the figure, we can see that the Normal Plot of Residual swerves. These results indicate that the residual is not normally distributed. Residual histogram also shows that the residual is not normally distributed. The right skews in histogram give us the sign that the data is not normally distributed. When the data is normally distributed, the residual should be in a bell-shaped or nearly a bell-shaped. Finally, the plot of Residual versus Fit indicates that the variance of error is not homogenous. We can say that the data is not fulfilled the assumption of normal distribution. So, the transformation of the data is required. The Box-Cox plot in Figure 3 shows that, the value 1 does not include in the 95% of confidence interval. Since these limits do not include the value 1, we conclude that the transformation is needed.

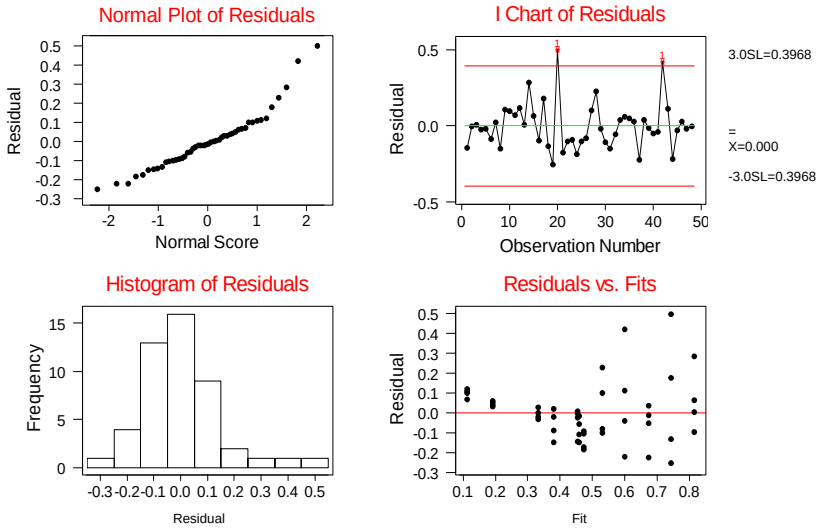

Figure 2: Residual model diagnostics for data before transformation

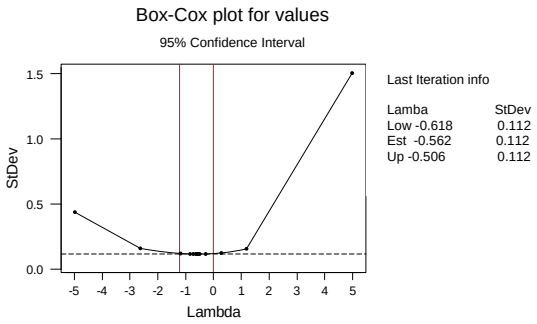

Figure 3: Box-Cox plot for value

Therefore we can conclude that there is a need of transformation. According to the Box-Cox, we have to determine the suitable values of parameter λ. According to the Figure 3 the best value for λ is –1. Difference values of λ in 95% of given confidence interval also can be used for calculation but it is easy to select the value of $\lambda = -1$. The transformation process can be employ after we select the value of parameter λ. Box-Cox (1964) mentioned that the value of -1 for parameter λ is representing the inverse of transformation. Once the transformation of the data has been done, the normality of the data transformation need to be verified. Figure 4 given the normal probability plot of the data after transformation.

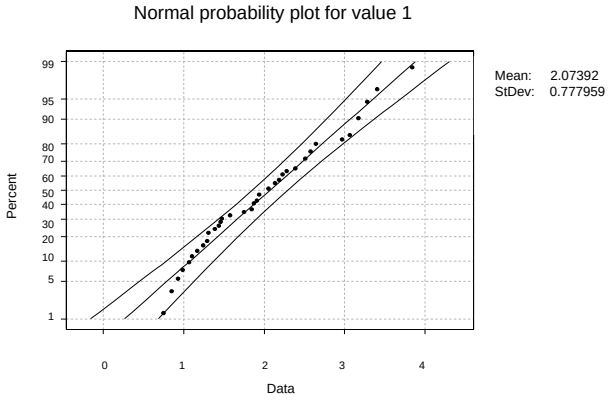

Figure 4: Normal probability plot of the data after transformation

The Normal Plot of Residuals is shown in Figure 5. From Figure 5 we can see that this plot is almost linear and this trend explained that it is normally distributed. Plot of Residuals Histogram illustrates almost a bell-shaped. This mentioned that the residual is approximating normally distributed as well. Lastly, the plot of Residuals versus Fits indicates that the variance of error is homogenous. So, we can conclude that the data transformation is fulfilled the assumption of normal distribution. The adequacy of data transformation, once again will be check by a Box-Cox plot and it is shown in Figure 6. The Box-Cox plot in Figure 6 demonstrate that there is a value 1 in 95% of confidence interval. Thus, the Box-Cox transformation is accomplished. Once a Box-Cox transformation has been done, the transformation data can be used in order to perform a parameter analysis.

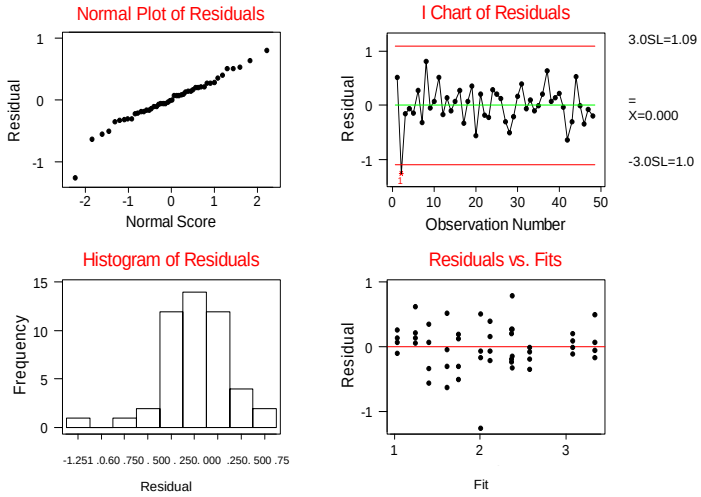

Figure 5: Residual model diagnostics for data after transformation

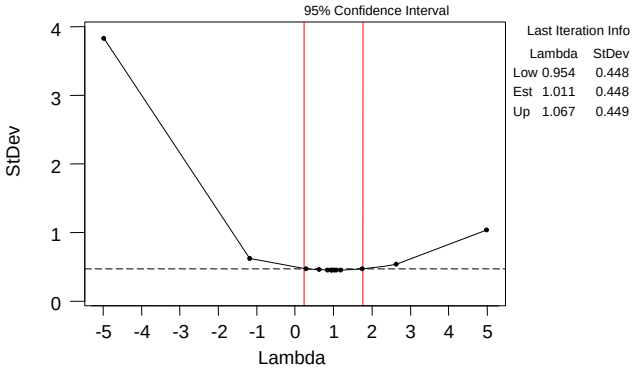

Figure 6: Box-Cox plot for values after transformation

DISCUSSION AND CONCLUSION

Most of the researchers use the method of parametric statistics without checking the nature of the data. When we applied this kind of method to the analysis, the results of this analysis is not accurate until we clean the data. The process of cleaning data is called data transformation. There are various methods to clean the data, but one of the famous methods is through the Box-Cox transformation. The Box-Cox algorithm provides a simple method to determine the best way to transform the data for reducing heterogeneity of errors (Atkinson 1986). This transformation also is well-adapted to bring heavily skewed data sets to near

normality (Draper & Cox 1969). The Box-Cox data transformation is a simple method that enable analysis of heteroscedastic and non-normal data sets so that the assumption of the analysis of variance might be satisfied especially when other transformation procedure fail. From the Figure 1 until Figure 6, we can clearly see how transformation leading the role of normalization. After the data gone through the transformation and fulfilled the normality assumption, we can applied the method of parametric successfully. Table 3 shows the transformation value of the data set after using the Box-Cox method.

Table 3: Data set of lifetime animal in 3×4 factorial design of experiment after transformation

Poison	Treatment			
	A	**B**	**C**	**D**
I	3.22581	1.21951	2.32558	2.22222
	0.68966	0.90909	2.22222	1.40845
	2.17391	1.13636	1.58730	1.51515
	2.32558	1.38889	1.31579	1.61290
II	2.77778	1.08696	2.27273	1.78571
	3.44828	1.63934	2.85714	0.98039
	2.50000	2.04082	3.22581	1.40845
	4.34783	0.80645	2.50000	2.63158
III	4.54545	3.33333	4.34783	3.33333
	4.76190	2.70270	4.00000	2.77778
	5.55556	2.63158	4.16667	3.22581
	4.34783	3.44828	4.54545	3.03030

REFERENCE

Andrew D F. (1971). A note on the selection of data transformation. *Biometrical* 58: 249–254.

Anscombe F J. (1961). Examination of Residual. *Proceedings of the 4th Berkeley Symposium on of the Mathematical Statistics and Probability,* 1–36.

Atkinson A C. (1986). Diagnostics test for transformations. *Technometrics* 28: 29–38.

Bartlett M S. (1947). The use of transformation. *Biometrical* 3: 39–52.

Box G E P and D R Cox. (1964). An analysis of transformation. *J. R. Stat Soc. B* 26: 211–252.

Damon R A Jr and W R Harvey. (1987). *Experimental design, ANOVA and regression*. New York: Harper and Row.

Draper N R and D R Cox. (1969). On distribution and their transformation to normality. *R. Stat Soc. B* 31: 472–476.

Montgomery D C. (1991). *Design and analysis of experiments*. 3rd ed. New York: John Wiley & Sons.

Peltier M R. (1996). *Effect of melatonin implantation at the summer solstice and ovarian status on the annual reproductive rhythm in pony mares*. MS thesis, University of Florida, Gainesville.

Sakia R M. (1992). The Box-Cox transformation technique: A review. *The Statistician* 41: 169–178.