Statistika, Vol. 6 No. 2, 1 – 6
Nopember 2006

An Approached of Box-Cox Data Transformation to Biostatistics Experiment

WAN MUHAMAD AMIR W AHMAD[1], NYI NYI NAING[2], NORHAYATI ROSLI[3]

[1] Jabatan Matematik, Fakulti Sains dan Teknologi, Malaysia, Kolej Universiti Sains dan Teknologi Malaysia (KUSTEM), 21030 Kuala Terengganu, Terengganu Malaysia.

[2]Unit Biostatistik, Pusat Pengajian Sains Perubatan, Universiti Sains Malaysia, USM, Kampus Kesihatan, 16150 Kubang Kerian, Kelantan, Malaysia.

[3] Fakulti Kejuruteraan Kimia dan Sumber Asli, Kolej Universiti Kejuruteraan dan Teknologi Malaysia, 25000 Kuantan Pahang, Malaysia.

ABSTRACT

The Box-Cox family of transformation is a well-known approach to make data behave accordingly to assumption of linear regression and ANOVA. The regression coefficients, as well as the parameter λ defining the transformation are generally estimated by maximum likelihood, assuming homoscedastic normal error. In application of ANOVA for hypothesis testing in biostatistics science experiments, the assumption of homogeneity of errors often is 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. We will illustrate the use of the Box-Cox method by using MINITAB software.

Keywords: Box-Cox transformation and parameter λ

1. 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 according 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. 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 often are accompanied by changes in error variances (a scale effect). Transformation of the data is usually the 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 are analyzed so that the variances are not heterogeneous. However, the transformation that is most effective in reducing the heterogeneity of variance often is not obvious and often must be 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 untransformed data and the results from the transform data are discussed.

2. 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 and 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).

3. Case Study

This study is done based on the data that has been used by Box and Cox. 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 1.1 shows the dataset of lifetime animal in 3×4 factorial design of experiment with 3 level of poison factor and 4 level of treatment factor.

Table 1.1 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

3.1 Diagnostic Checking

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. If, there are a sign that show the data is deflecting from the assumption of normality, then the transformation is needed. Figure 1.1 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 1.1 displays the deflections in point. Thus, there is enough evidence to say that the normality assumption is contravened in this case. To support this interpretation let us look at the Figure 1.2. 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 shape or nearly bell a shape. 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 plot Box-Cox in Figure 1.3 shows that, the value 1 do not include in the 95% of confidence interval. Since these limits do not include the value 1, we conclude that the transformation is needed.

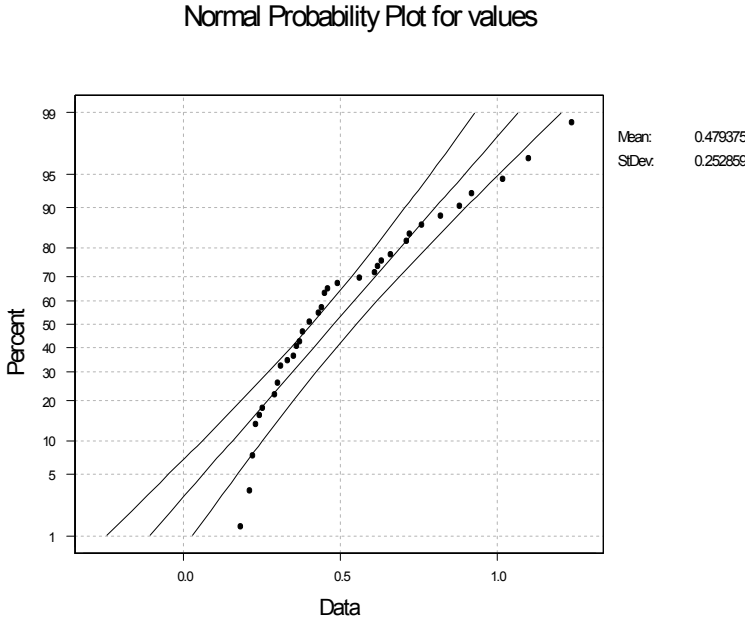

Figure 1.1 Normal probability plot of the data

3.2 Remedial Measurement

From the diagnostics checking techniques, 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 1.3 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 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 1.4 given the Normal probability plot of the data after transformation.

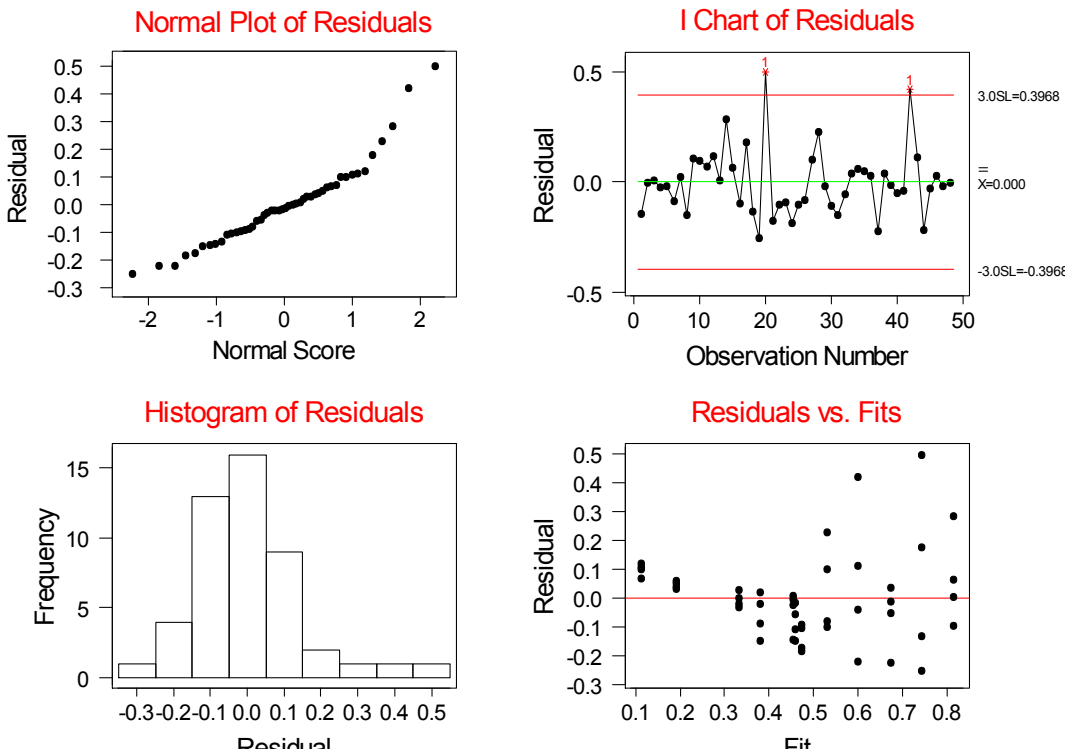

Figure 1.2 Residual Model Diagnostics for Data Before Transformation

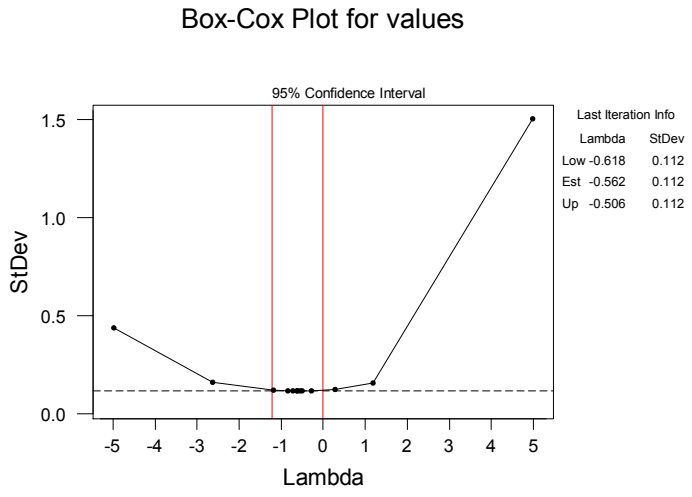

Figure 1.3 Box-Cox Plot for Values

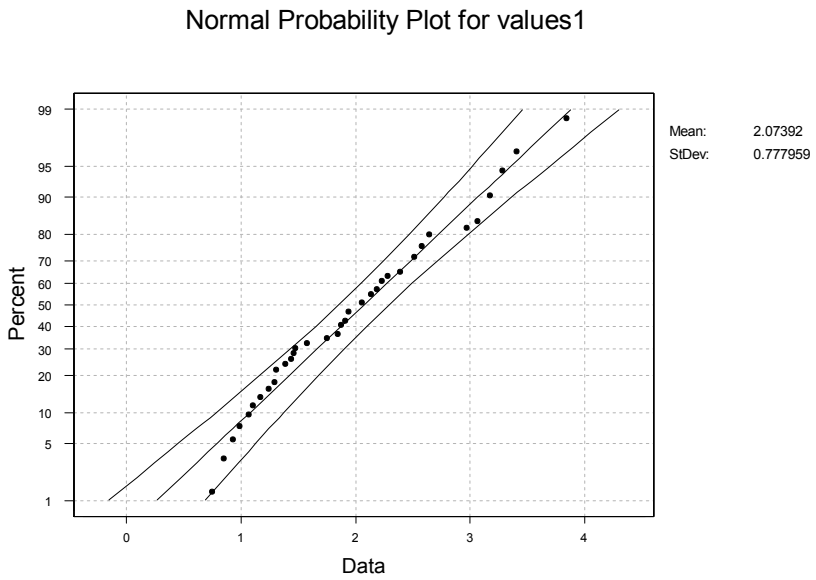

Figure 1.4 Normal Probability Plot of the Data After Transformation

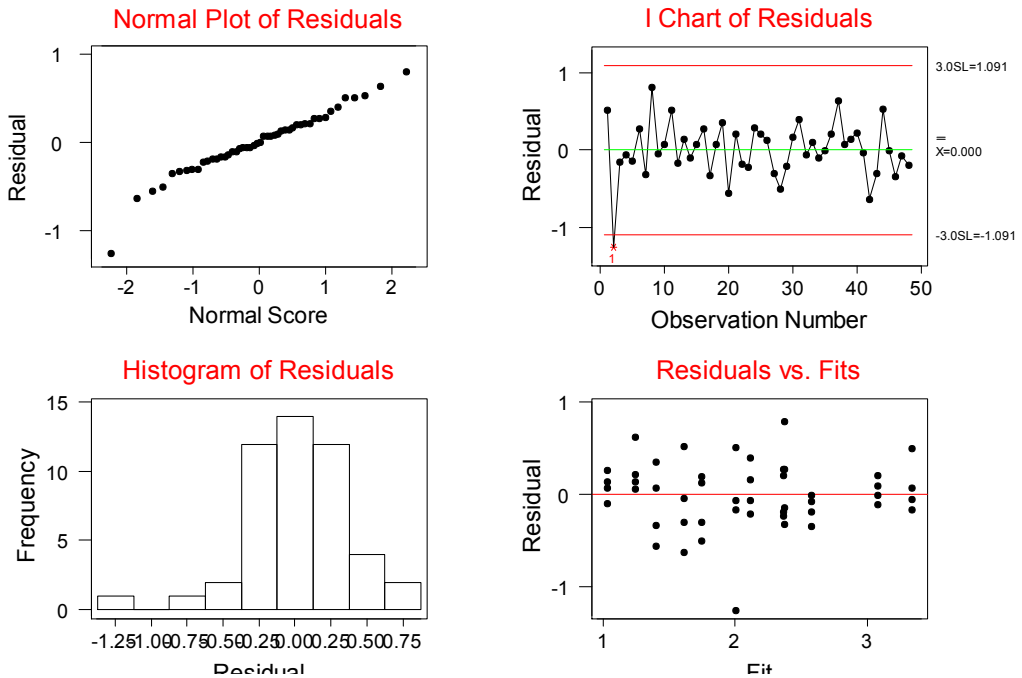

Figure 1.5 Residual Model Diagnostics for Data After Transformation

The Normal Plot of Residual is shown in Figure 1.5. From Figure 1.5 we can see that this plot is almost linear and this trend explained that it is normally distributed. Plot of Residual

Histogram illustrate almost bell a shape. This mentioned that the residual is approximating normally distributed as well. Lastly, the plot of Residual versus Fit 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 do a Box-Cox plot and it is shown in Figure 1.6. The Box-Cox plot in Figure 1.6 demonstrate that there is a value 1 in 95% of confidence interval. Thus, the Box-Cox transformation is accomplishment. Once a Box-Cox transformation has been done, the transformation data can be used in order to employ a parameter analysis.

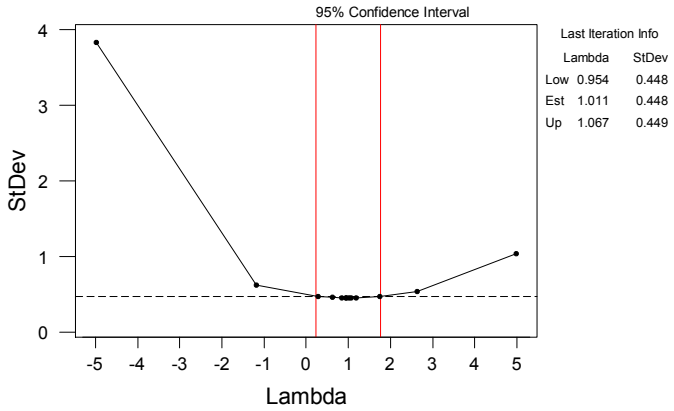

Figure 1.6 Box-Cox Plot for Values After Transformation

4. Discussion

The Box-Cox algorithm provides a simple method to determine the best way to transform the data for reducing heterogeneity of errors. This transformation also is well adapted to bringing heavily skewed data sets to near normality (Draper and 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.

5. Reference

[1] Andrew, D.F. (1971). A Note on the Selection of Data Transformation. Biometrical 58, 249-254.
[2] Anscombe, F.J. (1961). Examination of Residual. Proc. Fourth Berkeley Symp. On Mathematical Statistics and Probability 1, 1-36. University of California Press, Berkeley.
[3] Atkinson, A.C. (1973). Testing Transformation to Normality. Journal of the Royal Stat. Society Ser. B, 35, 473-479.
[4] Atkinson, A.C. (1986a). Diagnostics Test for Transformations, Technometrics 28, 29-38
[5] Bartlett, M.S. (1947) The Use of Transformation. Biometrical 3 39-52.
[6] Box, G.E.P., and D.R. Cox (1964). An Analysis of Transformation. J.R. Stat Soc. B 26:211-252.
[7] Damon, R. A., Jr., and W. R. Harvey. 1987. Experimental Design, ANOVA and Regression. Harper and Row, New York.
[8] Draper, N. R., and D.R. Cox (1969). On Distribution And Their Transformation to Normality. .R. Stat Soc. B 31:472-476.
[9] Montgomery, D. C. 1991. Design and Analysis of Experiments (3rd Ed.). John Wiley & Sons, New York.
[10] Peltier, M. R. 1996. Effect of melatonin implantation at the summer solstice and ovarian status on the annual reproductive rhythm in pony mares. M.S. Thesis. University of Florida, Gainesville.