Journal of Mathematics and Statistics

Original Research Paper

Adjusted Transformation Methods for Reproduction Quality Control

Emmanouil-Nektarios Kalligeris, Alex Karagrigoriou, Kostantinos Ladopoulos and Christina Parpoula

Department of Statistics and Actuarial-Financial Mathematics, Lab of Statistics and Data Analysis, Samos, Greece

Article history
Received: 26-07-2019
Revised: 01-09-2019
Accepted: 12-09-2019

Corresponding author:
Emmanouil-Nektarios Kalligeris
Department of Statistics and Actuarial-Financial Mathematics, Lab of Statistics and Data Analysis, Samos, Greece
Email: ekalligeris@aegean.gr

Abstract: In this paper we introduce adjusted transformation methods for off-line quality control. The proposed methods result in proper performance measures for the determination of the controllable factors that affect the mean and variability of the response variable of interest. The details of the proposed adjusted transformations are provided and compared with the well-known Box and Coxtransformation and the safeguard method of Logothetis. The performance abilities of the proposed methodology are demonstrated on quality control data, considering both a real dataset and a simulated one.

Keywords: Quality Control, Adjusted Transformation Methods, Control Factors, Performance Measures, ANOVA

Introduction

Data transformation has been long recognized to play a key role in statistical quality control with particular interest lying in the determination of control factors that affect the mean and the variability of a response variable of interest. The well-known Box and Cox (1964) transformation has been found to interact effectively with the off-line quality control analysis (Taguchi and Konishi, 1987) serving adequately the above purposes. Transformations play also an important role in regression analysis (Cook and Weisberg, 1999). Such transformations often succeed in resolving issues related to the variability of the error term and as a result achieving to a quite satisfactory degree of homoskedasticity. Homoskedasticity is directly connected to the independence between the mean and the variance which should be reflected into the sample equivalen characteristics.

Logothetis (1990) attempted to provide an improved technique of safeguarding against the possibility of violating the homoskedasticity, or rather, the independence between mean and variance.

Another issue of great importance in data transformation is the case where negative values are involved. Several attempts to define transformation classes that include negative values have been suggested. Logothetis (1990), Cook and Weisberg (1999) and Yeo and Johnson (2000) have addressed the problem and proposed a number of such important transformations.

Based on the aforementioned, the need of a proper choice for data transformation is considered of high importance in statistical quality control. One of the drawbacks of the Box-Cox transformation is the risk that oversimplification via a model-linearly-oriented transformation could induce a mean bias in the variability. Such a risk jeopardies the homoskedasticity assumption or rather the independence between the mean and the variance. The safeguard method of Logothetis although effective is not fully satisfactory since the proposed model may be weak in terms of the prediction error and the model accuracy. In addition, for the case of negative values an alternative model is required making the methodology less attractive.

In this paper we propose a general model selection approach for the proper determination of an appropriate transformation that safeguards against the violation of homoskedasticity and at the same time provides a model of the highest possible accuracy. In addition, the proposed approach ensures the applicability even in the case where negative values are involved without any additional adjustments or corrections. The rest of the paper is organized as follows. In Section 2 standard transformation methods are briefly discussed. The proposed methodology is developed in Section 3 while Section 4 is devoted to applications based on real case and simulation studies for evaluating the performance of the new approach. Finally the conclusions are discussed.

Methodology

The family of transformations that is most often used in practice is the one introduced by Box and Cox (1964) which is defined as follows:

$$y_{bc} = \begin{cases} \dfrac{y^{\lambda}-1}{\lambda}, & \lambda \neq 0 \\ \log(y), & \lambda = 0 \end{cases},$$

where, y the original observations, y_{bc} the transformed observations and λ a properly selected value that defines the transformation applied to the original data.

Note that although the above class is applicable for strictly positive values of y a proper adjustment was also proposed in the same work, useful for $y > \lambda_1$, for λ_1 wisely selected by the experimenter. This adjustment could be viewed as a shifted location transformation given by:

$$y_{bc}^{(\lambda_1)} = \begin{cases} \dfrac{\left(y+\lambda_1\right)^{\lambda}-1}{\lambda_1}, & \lambda \neq 0 \\ \log\left(y+\lambda_1\right), & \lambda = 0 \end{cases}.$$

An important component of statistical quality control is the accurate decision regarding the improvement of the product or the process for which there is a strong need in identifying properly the role of each of the controllable factors involved in the process. More specifically there is a need in identifying those factors affecting the variability from those affecting the mean. It is frequently observed an ambiguity or an overlap in the classification of the factors. This fact could be resolved by a proper transformation leading to the independence between the variance and the mean. Logothetis (1990) suggested the general functional form $\sigma = f(\mu)$ for describing the relationship between the standard deviation σ and the mean μ. For the particular case where:

$$f(\mu) = a\mu^{k} \tag{1}$$

the power k could be estimated through the simple linear regression model between the $\log(\sigma)$ and the $\log(\mu)$ if a sample of n observations is available.

In statistical quality control, the Noise Performance Measure (NPM) is the measure which is used to identify the variability controllable factors of the process under investigation. In case there is no functional relation established then the noise performance measure is given by:

$$NPM = -\log_{10}\left(s_T^2\right),$$

where, s_T^2 is the sample variance of the transformed data. Otherwise, the measure should be defined in such a way so that the relationship established is removed. In such a case, the measure is given by:

$$NPM = 10\log_{10}\left(\frac{f(\bar{x})}{\sigma}\right)^{2} \tag{2}$$

where, $\bar{X}$ the sample mean of the data.

It should be noted that irrespectively of the noise measure chosen, the standard sample mean is always considered as the mean measure for the identification of the factors affecting the mean of the process.

The case of negative values could be handled in the above case in a similar fashion as in the Box-Cox transformation, with the implementation of an extra wisely chosen parameter k_1 entering into the function form, namely $\sigma = a(\mu + k_1)^k$. It should be pointed out that various approaches for handling negative values have been proposed over the years. Yeo and Johnson (2000) have proposed a general class of transformations applicable without restrictions which resembles the Box-Cox methodology. This class of transformations is given by:

$$y^{(\lambda,YJ)} = \begin{cases} \dfrac{(y+1)^{\lambda}-1}{\lambda}, & if\ \lambda \neq 0, y \geq 0 \\ \log(y+1), & if\ \lambda = 0, y \geq 0 \\ \dfrac{-\left[(-y+1)^{2-\lambda}-1\right]}{2-\lambda}, & if\ \lambda \neq 2, y < 0 \\ -\log(-y+1), & if\ \lambda = 2, y < 0 \end{cases}.$$

Observe that Yeo and Johnson transformation is equivalent to the two-parameter Box-Cox with $\lambda = 1$ for $y > 0$. On the other hand, for $y < -1$ the transformation is equivalent to Box-Cox transformation of the variable $-y+1$ with power $2-\lambda$ instead of λ.

Alternative classes of transformations for negative values like the folded power family (see Cook and Weisberg (1999), p. 330) have been proposed, but rarely used due to limited properties of the resulting transformations. For other such approaches see Hawkins and Weisberg (2017).

The Proposed Transformations

The Simple Polynomial Approach

For describing the relationship between the standard deviation and the mean, we propose:

$$\sigma = a\mu^{k} \tag{3}$$

and then identify, via an exhaustive search, the best estimates a and k for which the mean squared error is

minimized. Note that the approach is equivalent to the model selection criteria technique since all competing models have the same penalty term. Indeed, for Akaike Information Criterion (AIC) given by:

$$AIC = -2\log Lik + pn,$$

where, *Lik* the likelihood, *p* is fixed and equal to 2 representing the number of parameters involved for all competing models.

It should be pointed out that the approach based on the simple linear regression used by Logothetis is not a proper one since for its implementation a further "internal" log-transformation is required for the implementation of the regression analysis technique. As a result, the analysis attempts to model not the intended standard deviation but the logarithm of it with all the unavoidable consequences of reduced variability. This defect of the Logothetis method is resolved through the proposed procedure which attempts the modeling of the standard deviation taking into consideration the actual variability of the process.

The Full Polynomial Approach

Instead of the first approach as previously described, we further recommend the polynomial regression:

$$\sigma = a_0 + a_1\mu + a_2\mu^2 + ... + a_k\mu^k, \quad (4)$$

where, via an exhaustive search the best polynomial can be identified. In fact we choose the polynomial of degree *k* for which AIC is minimized.

Note that the proposed model is applicable not only for positive but also for negative values. Indeed, observe that the models proposed by Box-Cox and Logothetis can be considered as special cases of the proposed polynomial regression. The advantages of the full polynomial regression is the selection of the best possible model which possesses the highest possible accuracy in terms of the mean squared error and the coefficient of determination.

Based on the above methods the noise performance measures introduced in the previous section are reformulated accordingly.

Additionally, the control charts associated with the mean and variability of the process ($\bar{X} - R$ chart) could be appropriately adjusted to incorporate the proposed methodology. Indeed, such charts are useful in verifying whether a process is in control or whether changes have affected the process or product resulting in an out-of-control procedure. The implications of the proposed procedure will be presented through both real and simulated data in the following section.

Remark 1

In the KKLP method for selecting the optimal value of the parameter *k*, we first preassigned a sufficiently large range of candidate values and then for each value of k, through ordinary least squares (OLS), the estimate $\hat{a}$ of a is being estimated. The optimal k is chosen to be the one for which the minimum MSE is attained. Thus, the optimal estimate of a is the OLS estimator $\hat{a}$ obtained when k takes its optimal value. As a result, $\hat{a}$ has all standard properties of OLS i.e., consistency, unbiasedness and asymptotic normality. For the Ladopoly method, the aforementioned procedure was also implemented with the exception that instead of using MSE as a criterion for the determination of k, we made use of AIC, due to the fact that it assigns a considerable penalty for too many terms in the polynomial regression equation. Based on the above, all the estimates of the coefficients a involved in KKLP and Ladopoly methods, have the standard properties of OLS. It remains as an open problem to examine the asymptotic theory associated with the estimate of the power k which is left for future work.

Applications

Real Data - Performance Measures

The dataset consists of three (3) measurements for each combination of six (6) factors with three (3) levels each, according to the design of $OA_{18}3^6$ (see Taguchi and Konishi (1987)). The full dataset is given in Table 0 in the Appendix.

By applying the:

(a) Standard Taguchi performance measure
(b) Logothetis measure
(c) Box-Cox transformation
(d) The simple polynomial approach referred to as KKLP (Kalligeris-Karagrigoriou-Ladopoulos-Parpoula)
(e) The full polynomial approach referred to as Ladopoly (Ladopoulos polynomial)

to both the mean and variability, the results presented in the Appendix are obtained.

The purpose of this analysis is to identify the factors affecting the mean and those affecting the variability with the least possible overlapping. The mean and variance have been evaluated for each of the 18 experiments using the 3 available measurements and the ANOVA results are presented in Tables 1-5 (see Appendix). Table 1 refers to the mean and Tables 2-5 refer to the variability of the real data.

The Taguchi and Box-Cox transformations (Table 2 and Table 4) recognize both A and B as factors affecting the variability with almost identical Pvalues (i.e., Pvalues<0.05). Observe though, that the same

factors are recognized as factors affecting the mean (i.e., Pvalues = 0.00) resulting in a complete confusion (Table 1). Logothetis method (Table 3) resolves partly the problem by identifying only the factor A (i.e., Pvalue<0.05) while the proposed method of KKLP manages to fully resolve the issue by recognizing neither A nor B as factors affecting the variability (i.e., Pvalues>0.05).

In conclusion, the proposed methodology removes the dependence between mean and variance which results in clear discrimination between the factors affecting the main characteristics of the procedure.

Simulations

The simulations that took place, mimic the real dataset of Table 0 and performed with the use of R (Fox and Weisberg, 2019). The formulation of the simulated dataset is fully based on the one of the real dataset presented in the previous section. Finally, the proposed technique is applied to $\bar{X}-S$ control chart (Shewhart (1931)) which is used to evaluate the two basic characteristics of a procedure.

For this purpose we apply the:

(a) Standard $\bar{X}-S$ control chart;
(b) Logothetis transformation given in (1), $\sigma = f(\mu) \equiv \sigma_1$ and $\mu = f^{1}(\sigma s) \equiv \mu_1$
(c) Box-Cox transformation, $y_{bc} = \frac{y^{\lambda}-1}{\lambda-1}$
(d) KKLP given in (3), $\sigma = f_{KKLP}(\mu) \equiv \sigma_2$ and $\mu = f^{-1}_{KKLP}(\sigma)$ $f^{-1}_{KKLP}(\sigma) \equiv \mu_2 + (\sigma) \equiv \mu_3$
(e) Ladopoly given in (4), $\sigma = f_{Ladopoly}(\sigma)\_ \sigma_3$ and $\mu = f_{Ladopoly}$.

Figure 1 presents the charts for $\bar{X}$, μ_1, μ_2 and μ_3 while Fig. 2 the charts for S; σ_1, σ_2 and σ_3. We observe that the multiple $\bar{X}$-Graph reveals the out-of-control point via Logothetis and KKLP approaches. All others fail to reveal the out-of-control single point (observed at time point t_{14}). Ladopoly approach appears to behave similarly to the original $\bar{X}$-Graph at least in reference to the out-of-control point t_{14}. Note that the Ladopoly approach is the one that tends to describe as accurately as possible the observed $\bar{X}$-values regressed on the S-values. Therefore as expected the similarities between the original $\bar{X}$ and Ladopoly are observed. KKLP is superior to Logothetis due to the fact that in general, this approach results in a more accurate modeling of the underline characteristic (σ). Ladopoly, as before, attempts to describe as accurately as possible the original S-values and as a result the out-of-control point is easily revealed. Finally, Logothetis and KKLP are comparable in terms of identifying the out-of-control point but it should be stressed out that the modeling for Logothetis method is based on the logarithm transformation of the original (S, $\bar{X}$) data as opposed to KKLP for which the modeling is based on the raw data.

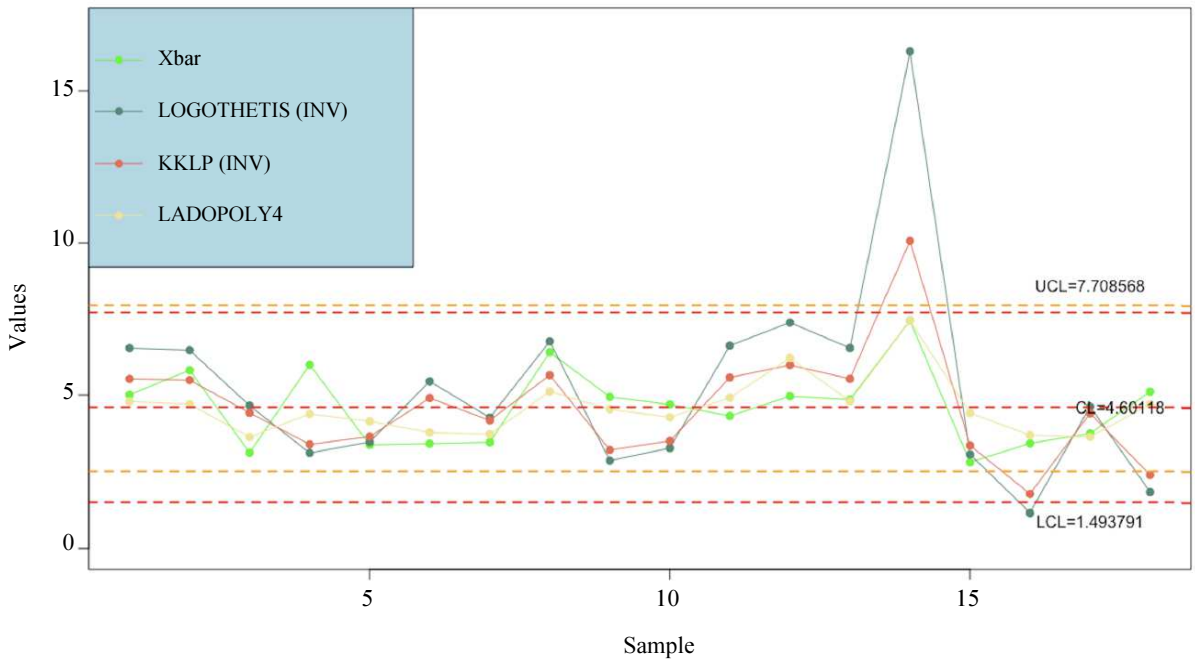

Fig. 1: $\bar{X}$-Graph

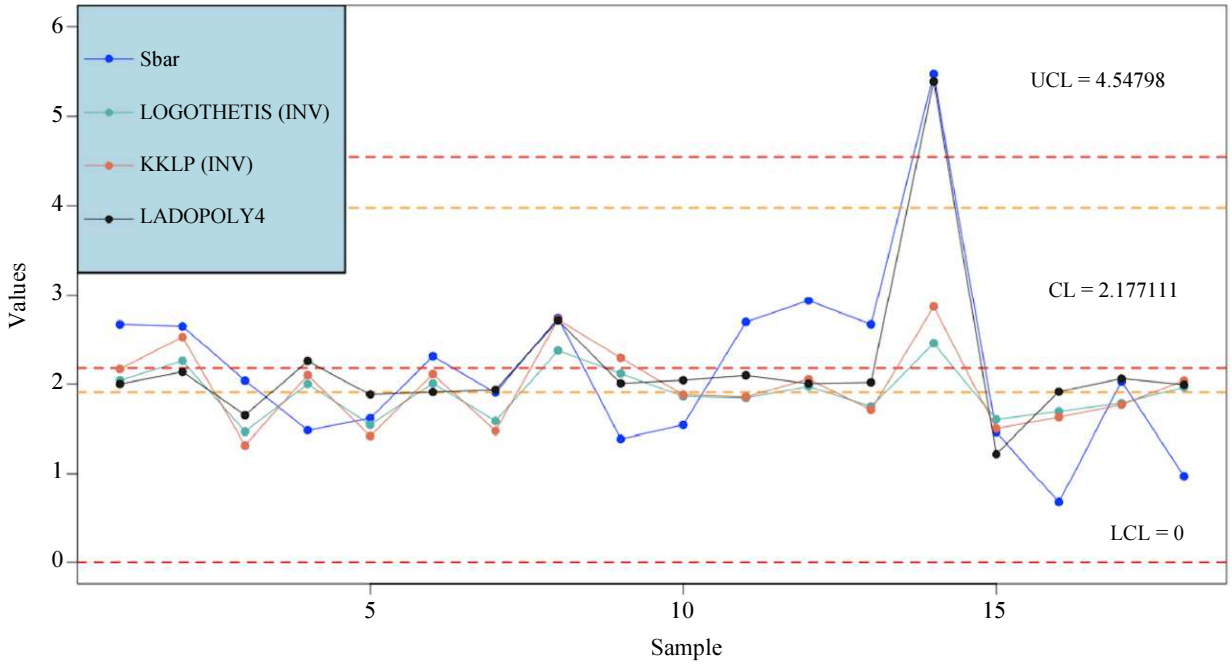

Fig. 2: S-Graph

Note that "Ladopoly4" and "Ladopoly3" in Fig 1 and 2, represent the full polynomial in (4) with k = 4 and 3, respectively.

Conclusion

A proper data transformation is considered to be of high importance in statistical quality control for achieving a satisfactory degree of homeskedasticity and at the same time ensuring high accuracy and great applicability. In this paper by relying on a general model selection approach we propose appropriate transformations that safeguard against the violation of homoskedasticity and at the same time provide the highest possible accuracy. In addition, the proposed transformations ensure the applicability even in the cases where negative values are involved.

The KKLP method through the model $\sigma = f_{KKLP}(\mu) = a\mu^k$ succeeds in recognizing discrepancies in the noise behavior which are depicted in the $\bar{X}$ -Graph through the transformation $\mu = (\sigma/k)^{1/a}$. Hence the KKLP methodology provides the double $\bar{X}$ -control chart which suffices to reveal the behavior of both noise and mean of the data.

Concludingly, we proposed adjusted transformations for off-line quality control. The proposed methods result in proper performance measures for the determination of the controllable factors that affect the mean and variability of the response variable of interest.

Acknowledgement

The authors wish to express their appreciation to the Editor and two anonymous Referees for taking the time to evaluate our work. Their comments and suggestions improved the quality as well as the presentation of the manuscript. This work was completed as part of the research activities of the Laboratory of Statistics and Data Analysis of the University of the Aegean.

Author's Contributions

E.N. Kalligeris: Conceptualization, data curation, formal analysis, methodology, software, supervision, writing – original draft.

A. Karagrigoriou: Conceptualization, formal analysis, methodology, supervision, writing – review and editing.

K. Ladopoulos: Data curation, Formal analysis, software.

C. Parpoula: Methodology, supervision, writing – review and editing.

References

Box, G.E.P. and D.R. Cox, 1964. An analysis of transformations. J. Royal Stat. Society, 26: 211-252. DOI: 10.1111/j.2517-6161.1964.tb00553.x

Cook, R.D. and S. Weisberg, 1999. Applied Regression Including Computing and Graphics. 1st Edn., Wiley, New York, ISBN-10: 047131711X, pp: 623.

Fox, J. and S. Weisberg, 2018. An R Companion to Applied Regression. 3rd Edn., SAGE Publications, LOS ANGELES, ISBN-10: 1544336470, pp: 608.

Hawkins, D. and S. Weisberg, 2017. Combining the Box-Cox power and generalized log transformations to accommodate nonpositive responses in linear and mixed effects linear models. South African Stat. J., 51; 317-328.

Logothetis, N., 1990. Box-Cox transformation and the Taguchi method. J. Royal Stat. Society, 39: 31-48. DOI: 10.2307/2347809

Shewhart, W.A., 1931. Economic Control of Quality of Manufactured Product. 1st Edn., Martino Fine Books, Eastford, pp: 501.

Taguchi, G. and S. Konishi, 1987. Taguchi methods orthogonal arrays and linear graphs: Tools for quality engineering. American Supplier Institute, Egypt.

Yeo, I.K. and R. Johnson, 2000. New family of power transformations to improve normality or symmetry. Biometrika, 87: 954-959. DOI: 10.1093/biomet/87.4.954

Appendix

Real Data Results

Table 0: Data used for the real case study

	Internal Order-Control Factors						Data		
	1	2	3	4	5	6			
Trial	A	B	C	D	E	F	Y_1	Y_2	Y_3
1	1	1	1	1	1	1	10.4	10.6	10.8
2	2	2	2	2	2	2	9.8	9.9	9.7
3	3	3	3	3	3	3	9.1	9.1	9.2
4	1	1	2	2	3	3	10.2	10.3	10.5
5	2	2	3	3	1	1	9.5	9.6	9.7
6	3	3	1	1	2	2	9.1	9.0	8.9
7	1	2	1	3	2	3	9.9	9.6	9.5
8	2	3	2	1	3	1	9.2	9.3	9.1
9	3	1	3	2	1	2	9.3	9.4	9.5
10	1	3	3	2	2	1	9.4	9.5	9.0
11	2	1	1	3	3	2	10.0	10.3	9.9
12	3	2	2	1	1	3	9.0	9.2	9.1
13	1	2	3	1	3	2	9.8	9.6	9.9
14	2	3	1	2	1	3	9.2	9.1	9.5
15	3	1	2	3	2	1	9.3	9.2	9.3
16	1	3	2	3	1	2	9.2	9.1	9.4
17	2	1	3	1	2	3	10.5	10.4	10.7
18	3	2	1	2	3	1	9.5	9.4	9.6

Table1: General linear model average versus A; B; C; D; E; F

Source	DF	Adj SS	Adj MS	F-Value	P-Value
A	2	9.645	4.823	1813.840	**0.000**
B	2	0.612	0.306	115.030	**0.000**
C	2	0.032	0.016	5.960	0.013
D	2	0.517	0.026	9.720	0.002
E	2	1.004	0.502	188.790	0.000
F	2	0.011	0.005	1.990	0.174
Error	14	0.037	0.003		
Total	26	11.392			

Table 2: General linear model NPM (Taguchi) versus A; B; C; D; E; F

Source	DF	Adj SS	Adj MS	F-Value	P-Value
A	2	0.979	0.490	6.86	**0.037**
B	2	1.698	0.849	11.88	**0.013**
C	2	0.095	0.047	0.66	0.555
D	2	0.088	0.044	0.62	0.576
E	2	0.046	0.023	0.32	0.740
F	2	0.037	0.019	0.26	0.780
Error	5	0.357	0.071		
Total	17	3.300			

Table 3: General linear model NPM (logothetis) versus A; B; C; D; E; F

Source	DF	Adj SS	Adj MS	F-Value	P-Value
A	2	146.039	73.019	10.24	**0.017**
B	2	2.189	1.094	0.15	0.862
C	2	43.208	21.604	3.03	0.137
D	2	0.316	51.658	0.79	0.502
E	2	4.379	2.190	0.31	0.749
F	2	2.800	1.400	0.2	0.828
Error	5	35.658	7.132		
Total	17	245.588			

Table 4: General linear model NPM (Box-Cox) versus A; B; C; D; E; F

Source	DF	Adj SS	Adj MS	F-Value	P-Value
A	2	0.059	0.030	6.95	**0.036**
B	2	0.103	0.051	12.06	**0.012**
C	2	0.006	0.003	0.66	0.555
D	2	0.005	0.003	0.59	0.587
E	2	0.003	0.001	0.33	0.731
F	2	0.002	0.001	0.25	0.787
Error	5	0.021	0.004		
Total	17	0.199			

Table 5: General linear model: NPM (KKLP) versus A; B; C; D; E; F

Source	DF	Adj SS	Adj MS	F-Value	P-Value
A	2	88.317	44.158	4.56	**0.075**
B	2	12.942	6.471	0.67	**0.553**
C	2	33.252	16.626	1.72	0.271
D	2	8.570	4.285	0.44	0.665
E	2	7.279	3.639	0.38	0.705
F	2	2.119	1.059	0.11	0.899
Error	5	48.435	9.687		
Total	17	200.915			