USING SAS FOR BOX-COX TRANSFORMATION IN REGRESSION ANALYSIS

Ollie Frazier and Deepak A. Keshani, Duke Power Company

In the regression analysis of data, it is assumed that observations $y_1, y_2, \ldots, y_n$ are independently normally distributed with constant variance and with expectations specified by a model linear in a set of parameter θ. In most applications this assumption is not satisfied. Many times, one uses a log model or a square root model in order to obtain a normally distributed error term. This manual process of trial and error can be automated with a proper procedure.

The linear regression model is appropriate after some suitable transformation suggested by Box and Cox (1964). It involves using the maximum log likelihood function technique to estimate the parameter values. However, the authors have set up a SAS procedure to obtain these parameter values for an univariate model. This procedure is applied to load research data at Duke Power Co. to estimate a missing value of electric energy use of a customer.

Many utilities (applicable to other industries) are now or will be billing their customers either on the highest demand for electricity for a certain time period or for each hour of every day. Under these complex billing arrangements a continuous hourly (half-hourly) reading of the electric usage is required. This information will be gathered by either an electronic or an electro-mechanical recording device. With all due respect to the manufacturers of these devices, a failure will occur at some point in history. During this time period when data is required and the recording device fails, there exists a need for a procedure to estimate any of the missing points. Our main interest at Duke Power Co. is being able to estimate our customers demands at the time when the aggregate of all these customers causes peak generation (commonly called system peak). At the present time, this variable is expensive to measure and is not known for the entire population.

Because of the expense of measuring demand at the time of system peak, sampling techniques are usually employed. In the utility industry, it is a common practice to use a stratified sample design. Problems may arise, however, when data points are lost due to malfunctions in the equipment used to obtain the data. The final sample used in the analysis may not be a proper subset of the population and the single most widely used statistical tool - regression analysis-usually is employed.

Generally, a straight line regression equation is fitted to the available points and in some cases the assumptions may have been overlooked. The residual plot may indicate heteroskedasticity, or it may indicate non-normality. Another major problem seen in the study is sort of an "outlier" type, a point remote in X-space controlling the regression line.

According to the paper by Box and Cox (1964), a transformation form $Z_i = h(Y_i, \lambda)$ of original data $\{Y_i\}$ satisfies a linear model. The transformation considered here is

$$TY = \frac{Y^{\lambda}-1}{\lambda}$$

$$TX = \frac{X^{\lambda}-1}{\lambda}$$

where λ is a real variable ($\lambda \neq 0$). This transformation holds for $X, Y > 0$. λ is obtained thru an iterative process using a maximum likelihood approach. In this transformation, if

$$TY = \log Y \quad , \text{ and}$$

$$TX = \log X$$

are considered at $\lambda = 0$ then it is continuous at $\lambda = 0$.

Using this transformation, a SAS procedure was set up at Duke Power Co. To use this procedure, a user may specify a range, and an increment value for λ. A data set is then created containing values of λ, α, β and a CHECK variable. The CHECK variable is a flag to indicate whether the likelihood function was maximized or not. The predicted value of Y can then be obtained with

$$\hat{Y} = (1 + \alpha\lambda + \beta(X^{\lambda}-1))^{1/\lambda}$$

A copy of users literature is provided in the appendix.

To illustrate the use of the Box-Cox technique, a set of data was generated and the results are presented in the figures shown on the following page. Figure 1 contains scatter diagrams with two regression lines. One was obtained using Box-Cox technique, while the other was obtained without any transformation. Without using transformation, a regression line was fitted and the residual plot is shown in Figure 2. This graph displays problems of heteroskedasticity and non-normality. The residual points start from the negative region and then spread out in the positive and negative regions as the predicted value increases. To look at the effect of the transformation on the regression analysis, a regression line was fitted on the transformed variables (TX, TY). The residual plot is shown in Figure 3. This graph does not indicate problems with heteroskedas-

WITH AND WITHOUT USING BOXCOX TRANSFORMATION
IN ORIGINAL FORM
GROUP=1

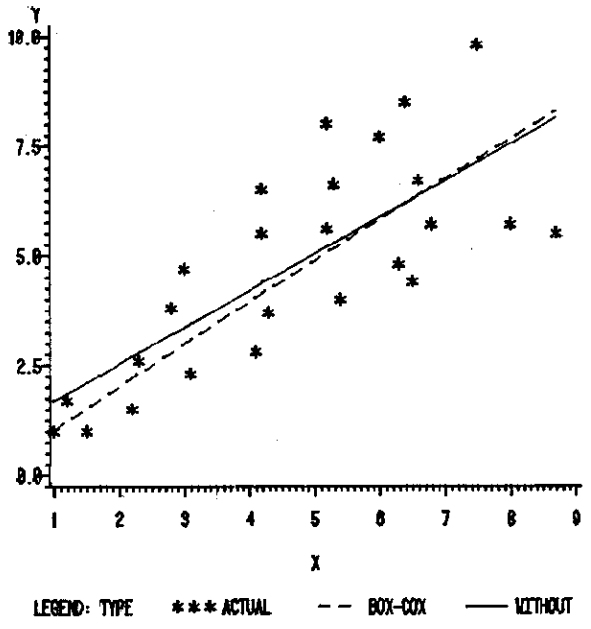

FIGURE 1

RESIDUAL PLOT USING BOXCOX TECHNIQUE
IN TRANSFORMED FORM
GROUP=1

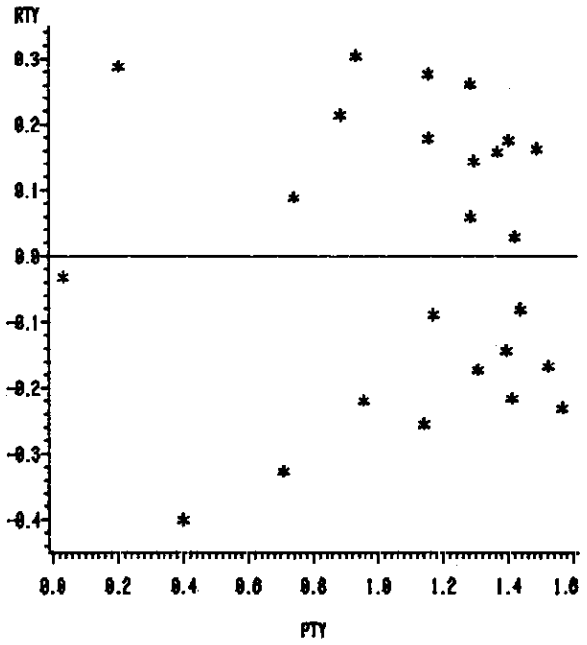

FIGURE 3

RESIDUAL PLOT WITHOUT USING TRANSFORMATION
GROUP=1

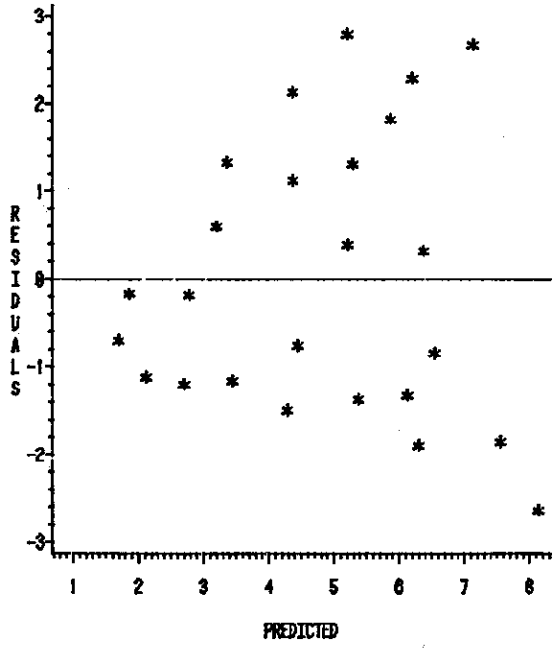

FIGURE 2

RESIDUAL PLOT USING BOXCOX TECHNIQUE
IN ORIGINAL FORM
GROUP=1

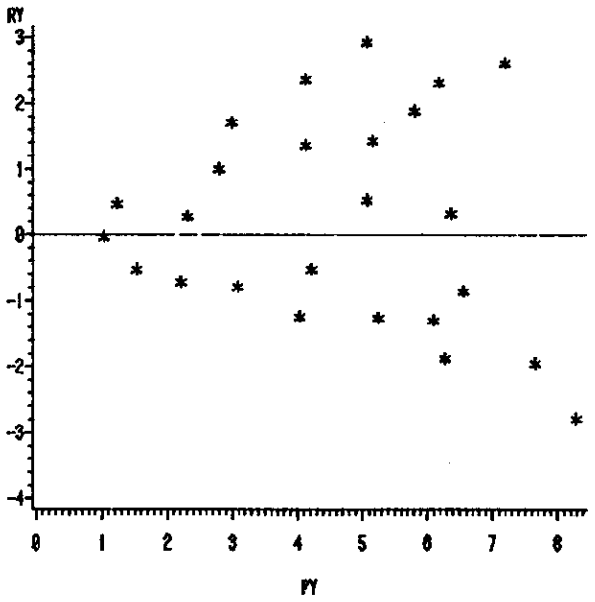

FIGURE 4

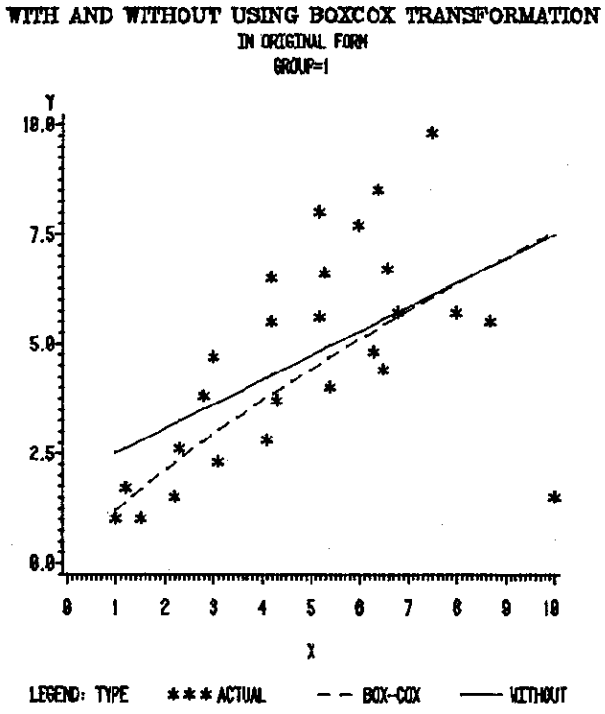

FIGURE 5

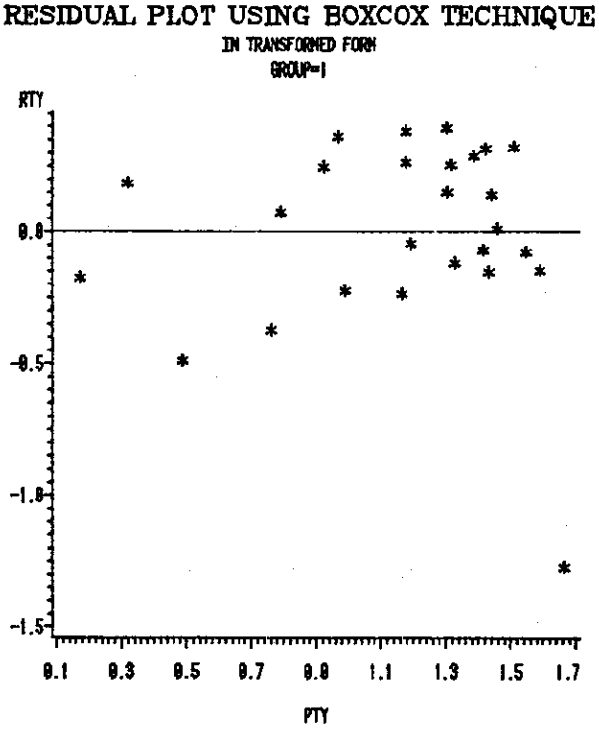

FIGURE 7

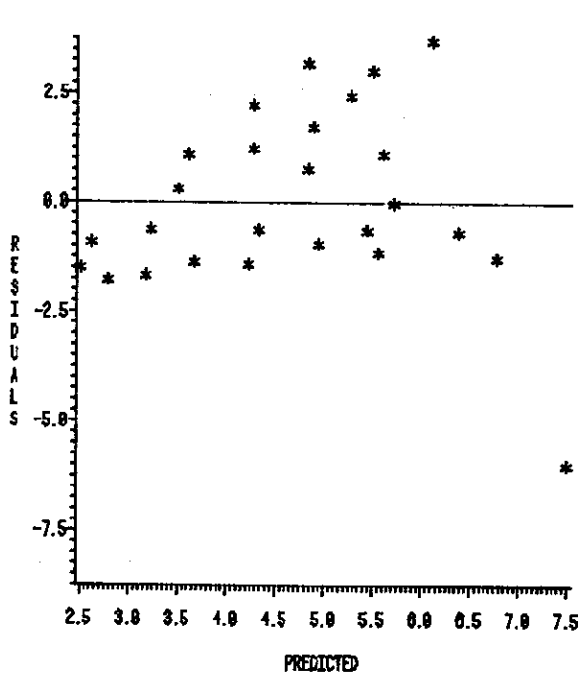

FIGURE 6

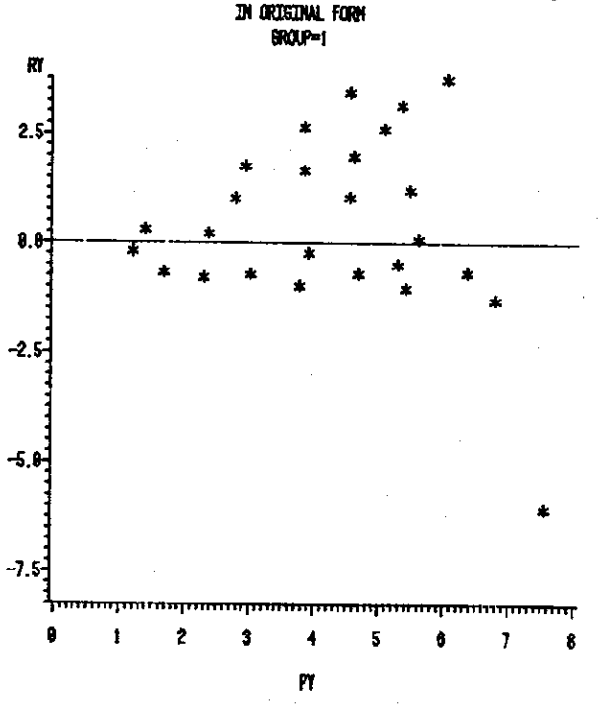

FIGURE 8

ticity. Also the residual points started by spreading out in neither the negative nor the positive region. The residual varies from positive side to negative side as the magnitude of predicted value increases.

In order to see the effect of the transformation on the original form of the variables (X and Y), the predicted value of TY, that is PTY, was untransformed to PY. The residual plot is shown in Figure 4. This plot exhibits heteroskedasticity because the variables are in the original form. The points, however, spread out in neither the negative nor the positive region, but they spread out almost evenly in both the regions.

Thus, Figure 2 and Figure 4 can be compared to determine the better fit of the regression line and they can help to determine which regression line will provide better estimation for Y over the entire range. For example, Figure 2 indicates that in this case, for smaller values of Y, the predicted values will have negative error terms, while Figure 4 indicates that in the same case for smaller values of Y, the predicted values may have positive or negative error terms.

In order to see the effect of an outlier in the data, an additional point was added to the set of data used in the previous example.

Similar to Figures 1 thru 4, the Figures 5 thru 8 are shown. Figure 5 indicates a difference in the intercept value of the lines. This difference may look small but it may have a great effect when it is applied to the entire population points in the smaller region of X. The residual plot of the regression line obtained without using a transformation is shown in the Figure 6. It does indicate constant variance and normality problems. After the transformation, the regression line was obtained on transformed variables and the residual plot is shown in Figure 7. It does indicate minor problems with heteroskedasticity and non-normality, but compared to Figure 6 these problems appear to be minor. Figure 8 shows the residual plot for the fitted regression line on the transformed variables, but the residual is obtained from the untransformed predicted values.

Thus, Box-Cox transformation technique provides a proper transformation to obtain a better fit for a regression line. In this technique, when $\lambda = 0.5$, it is a square root transformation. If $\lambda = 0$ is considered in the model, it is a log transformation. Since it covers most of the general types of transformation and determines the best one to use, user may find it easy to use this technique. However, precautions must be taken and diagnostic checks should be made before using a particular transformation. For further information, users may refer to Box, G.E.P., and Cox, D.R. (1964), Bickel, P.J., and Doksum, K.A. (1981).

In this type of utility study, it is very hard, nearly impossible, to replace the sample and to try using another sample. Also, the level of accuracy is very important. Many times, linear regression analysis is used to approximate the desired estimates. The Box-Cox transformation technique can resolve many of the problems in meeting the assumptions, allowing a linear regression analysis technique to be used and to obtain better estimates of demands.

APPENDIX

THE PROC BOXCOX STATEMENT

PROC BOXCOX Options:

The options below may appear in the PROC BOXCOX statement.

DATA = data set name
DATA = gives the name of the data set to be used by PROC BOXCOX. If it is omitted, the last data set created will be used.

OUT = data set name
OUT = gives the name of the data set to be built by the PROC BOXCOX. If it is omitted, the data set name DATA1 will be used.

The variables in the BY statement as well as the computed parameter variables (ALPHA, BETA, LAMDA), and condition check variable (CHECK: value is 1 if the MLF is maximized, otherwise 0) will be included in the new data set.

INDEP = 2
INDEP = gives the position of independent variable in the VARIABLE statement. If it is omitted, the position value 2 will be used.

DEP = 1
DEP = gives the position of dependent variable in the VARIABLE statement. If it is omitted, the position value 1 will be used.

BEGIN = 2
BEGIN = gives the initial value for the parameter λ. If it is omitted, the initial value -2 will be used.

STEP = 0.1
STEP = gives the increment value for λ. If it is omitted, the increment value 0.1 will be used.

FINAL = 8
FINAL = gives the final value of λ, unless the likelihood function is maximized before this final value. If it is omitted, the final value is computed using BEGIN value and STEP value as:

FINAL = BEGIN + 100 * STEP

PRINT

PRINT tells BOXCOX to print a table containing values of α, β, λ, and $\sigma^2$and the likelihood for every value of λ.

STATEMENTS USED WITH BOXCOX

BY variable name:

A BY statement may be used with PROC BOXCOX if the data set has been sorted by the variables in the BY statement.

VAR variable names:

A VAR (variable) statement is used with PROC BOXCOX to provide the dependent and independent variable names.

EXAMPLE

This example shows how to use the Box-Cox procedure. In this same example the SYSREG procedure is invoked to compare the residuals and predicted values.

```
00010 DATA A ;
00020    INFILE IN ;
00030    INPUT X Y ;
00040    GROUP = 1 ;
00050
00060 PROC SORT ;
00070    BY GROUP ;
00080
00090 PROC PRINT ;
00100    BY GROUP ;
00110 TITLE1 ORIGINAL DATA LISTING ;
00120
00130 PROC BOXCOX DATA=A OUT=B P ;
00140    BY GROUP ;
00150    VAR Y X ;
00160 TITLE1 LISTING OF PRINT OPTION IN
       PROC BOXCOX ;
00170
00180 PROC PRINT ;
00190    BY GROUP ;
00200 TITLE1 OUTPUT DATA SET FROM PROC
       BOXCOX ;
00210
00220 PROC SYSREG DATA=A OUT=C ;
00230    BY GROUP ;
00240    MODEL Y = X ;
00250    OUTPUT P = P_Y
00260           R = R_Y ;
00270 TITLE1    ;
00280
00290 PROC SORT ;
00300    BY GROUP X Y ;
00310
00320 DATA E ;
00330    MERGE A B ;
00340      BY GROUP ;
00350    PY = (1+ALPHA*LAMDA+BETA*(X**
              LAMDA-1))**(1/LAMDA) ;
00360    RY = Y - PY ;
00370    TX = ( X**LAMDA - 1)/LAMDA ;
00380    TY = ( Y**LAMDA - 1)/LAMDA ;
00390   PTY = ALPHA + BETA*TX ;
00400   RTY = TY - PTY ;
00410   KEEP GROUP X Y PY RY TX TY PTY
         RTY ;
00420
00430 PROC SORT ;
00440    BY GROUP X Y ;
00450
00460 DATA PRT ;
00470   MERGE C E ;
00480     BY GROUP X Y ;
00490
00500 PROC PRINT ;
00510   BY GROUP ;
00520 TITLE1 FINAL RESULT LISTING ;
00530
```

The printout of the SAS code:

ORIGINAL DATA LISTING

GROUP=1

OBS	X	Y			
1	1.0	1.0	13	5.4	4.0
2	1.5	1.0	14	5.2	5.6
3	1.2	1.7	15	5.3	6.6
4	2.2	1.5	16	5.2	8.0
5	2.3	2.6	17	6.0	7.7
6	3.1	2.3	18	6.3	4.8
7	2.8	3.8	19	6.4	8.5
8	3.0	4.7	20	6.5	4.4
9	4.1	2.8	21	6.6	6.7
10	4.3	3.7	22	6.8	5.7
11	4.2	5.5	23	8.0	5.7
12	4.2	6.5	24	8.7	5.5
			25	7.5	9.8

LISTING OF PRINT OPTION IN PROC BOXCOX

LIK.FUN.	LAMDA	ALPHA	BETA	SS
-43.096	-2.00	-0.003	0.986	0.005
-39.576	-1.90	-0.003	0.988	0.005
-36.056	-1.80	-0.004	0.989	0.006
-32.536	-1.70	-0.005	0.991	0.006
-41.516	-1.60	-0.005	0.992	0.007
-37.997	-1.50	-0.006	0.993	0.007
-34.477	-1.40	-0.006	0.993	0.008
-30.957	-1.30	-0.006	0.993	0.009
-27.437	-1.20	-0.006	0.993	0.010
-23.917	-1.10	-0.005	0.992	0.011
-20.397	-1.00	-0.004	0.991	0.013
-16.877	-0.90	-0.002	0.989	0.015
-13.357	-0.80	-0.000	0.987	0.017
-22.338	-0.70	0.003	0.983	0.020
-18.818	-0.60	0.007	0.979	0.025
-15.298	-0.50	0.014	0.975	0.030
-11.778	-0.40	0.021	0.970	0.037
- 8.258	-0.30	0.031	0.964	0.047
-17.238	-0.20	0.044	0.957	0.060
-13.718	-0.10	0.059	0.949	0.078
-19.178	0.10	0.102	0.933	0.136
-15.659	0.20	0.130	0.923	0.183
-12.139	0.30	0.164	0.914	0.248

OUTPUT DATA SET FROM PROC BOXCOX

GROUP = 1

OBS	LAMDA	ALPHA	BETA	CHECK
1	-0.3	0.0319455	0.96404	1

GROUP = 1

MODEL: MODEL01 SSE 59.040332 F RATIO 30.69
DFE 23 PROB>F 0.0001
DEP VAR: Y MSE 2.566971 R-SQUARE 0.5716

VARIABLE	DF	PARAMETER ESTIMATE	STANDARD ERROR	T RATIO	PROB>^T^
INTERCEPT	1	0.865486	0.779864	1.1098	0.2786
X	1	0.835848	0.150890	5.5395	0.0001

FINAL RESULT LISTING

GROUP = 1

OBS	X	Y	P_Y	R_Y	PY	RY	TX	TY	PTY	RTY
1	1.0	1.0	1.70133	-0.7013	1.03262	-0.0326	0.00000	0.00000	0.03195	-0.03195
2	1.2	1.7	1.86850	-0.1685	1.23298	0.4670	0.17743	0.49055	0.20299	0.28756
3	1.5	1.0	2.11926	-1.1193	1.53128	-0.5313	0.38178	0.00000	0.39999	-0.39999
4	2.2	1.5	2.70435	-1.2044	2.21877	-0.7188	0.70214	0.38178	0.70884	-0.32707
5	2.3	2.6	2.78794	-0.1879	2.31615	0.2839	0.73700	0.83076	0.74244	0.08832
6	2.8	3.8	3.20586	0.5941	2.80035	0.9996	0.88578	1.10005	0.88588	0.21417
7	3.0	4.7	3.37303	1.3270	2.99288	1.7071	0.93592	1.23802	0.93421	0.30381
8	3.1	2.3	3.45661	-1.1566	3.08891	-0.7889	0.95939	0.73700	0.95684	-0.21984
9	4.1	2.8	4.29246	-1.4925	4.04156	-1.2416	1.15038	0.88578	1.14096	-0.25518
10	4.2	5.5	4.37605	1.1240	4.13612	1.3639	1.16611	1.33453	1.15612	0.17841
11	4.2	6.5	4.37605	2.1240	4.13612	2.3639	1.16611	1.43223	1.15612	0.27611
12	4.3	3.7	4.45963	-0.7596	4.23056	-0.5306	1.18135	1.08211	1.17082	-0.08871
13	5.2	5.6	5.21189	0.3881	5.07556	0.5244	1.30061	1.34531	1.28579	0.05952
14	5.2	8.0	5.21189	2.7881	5.07556	2.9244	1.30061	1.54704	1.28579	0.26126
15	5.3	6.6	5.29548	1.3045	5.16893	1.4311	1.31220	1.44092	1.29695	0.14397
16	5.4	4.0	5.37906	-1.3791	5.26219	-1.2622	1.32350	1.13415	1.30785	-0.17370
17	6.0	7.7	5.88057	1.8194	5.81980	1.8802	1.38603	1.52644	1.36813	0.15831
18	6.3	4.8	6.13133	-1.3313	6.09737	-1.2974	1.41433	1.25121	1.39541	-0.14420
19	6.4	8.5	6.21491	2.2851	6.18972	2.3103	1.42337	1.57924	1.40413	0.17511
20	6.5	4.4	6.29850	-1.8985	6.28198	-1.8820	1.43223	1.19614	1.41268	-0.21653
21	6.6	6.7	6.38208	0.3179	6.37416	0.3258	1.44092	1.44944	1.42105	0.02839
22	6.8	5.7	6.54925	-0.8492	6.55827	-0.8583	1.45779	1.35583	1.43732	-0.08148
23	7.5	9.8	7.13434	2.6657	7.20014	2.5999	1.51212	1.65255	1.48969	0.16286
24	8.0	5.7	7.55227	-1.8523	7.65633	-1.9563	1.54704	1.35583	1.52336	-0.16752
25	8.7	5.5	8.13736	-2.6374	8.29199	-2.7920	1.59143	1.33453	1.56615	-0.23162

REFERENCES

Bickel, P.J. and Doksum, K.A. (1981), "An Analysis of Transformations Revisited", Journal of the American Statistical Association, Vol. 76, 296-311.

Box, G.E.P. and Cox, D.R. (1964), "An Analysis of Transformations", Journal of Royal Statistical Society, Ser. B, 26, 211-252.