

BOX-COX TRANSFORMATION ANALYSIS MACRO

Steve Orlich and Mike Delozier
Minitab Inc.
3081 Enterprise Drive
State College, PA 16801

March 30, 2001

ABSTRACT

This macro determines the likelihood estimate of the Box-Cox power transformation parameter in regression and response surface modeling applications. Here, we consider modeling $(y^\lambda - 1)/\lambda$ ($\log y$ if $\lambda = 0$) where $\lambda \in [-2, 2]$ in terms of one or more predictor variables and estimate the parameter λ from the data. Note that if a nonzero value of λ is chosen to transform the response, y^λ may be used in the analysis since the difference is simply an origin shift and a scale change. A plot of the log-likelihood function over a range of λ values is displayed showing the likelihood estimate of λ and an approximate 95% confidence interval for λ . Also displayed is a plot of the values of the PRESS statistic transformed back to the original response scale over the 95% confidence interval for λ . Optionally, the user may choose to specify the range for λ in the plot of PRESS, display an index plot due to Cook and Wang (1983) showing the influence of individual cases on the likelihood estimate, and store all computed results.

MACRO COMMAND AND SUBCOMMANDS

%BCtrans response C and predictors C...C

This is the command that executes the macro. Enter the response variable column first followed by the predictor variable columns. If %BCtrans C1-C5 is entered, the macro considers C1 to be the response and C2-C5 to be the predictors.

range K K

Use this subcommand to specify the range for λ in the PRESS plot. Specifying a range for λ overrides the default range that covers the 95% confidence interval.

influence

Use this subcommand to display an index plot showing the influence of individual cases on the likelihood estimate of λ .

bcstore C C

Use this subcommand to store the log-likelihood and corresponding λ values used in creating the log-likelihood plot. Two columns must be specified. The first column will contain the log-likelihood values and the second column will contain the λ values.

infstore C C

Use this subcommand to store the approximate likelihood distances and case numbers used in creating the index plot. Two columns must be specified. The first column will contain the approximate likelihood distance values and the second column will contain the case numbers.

presstore C C

Use this subcommand to store the PRESS statistic and corresponding λ values used in creating the PRESS plot. Two columns must be specified. The first column will contain the PRESS statistic values and the second column will contain the λ values.

IMPORTANT FEATURES OF THE MACRO**Missing data**

Rows in the data set containing missing data are removed from the analysis.

Model specification

If terms other than linear (such as interactions or squared terms) are to be included in the model, the user must create the appropriate columns in the worksheet and enter these columns in the command line as predictors.

Example data sets

Two example data sets (see below) are included that may be used for verification of output.

The first data set included is the MINITAB trees data and may be used to verify the likelihood estimate of λ and the index plot. The estimate reported by the macro in the sample run matches that given in Cook and Weisberg (1982) for the first-order model using volume as the response and height and diameter as the predictors. Further, one can verify that the estimates reported by the macro match those given in Cook and Weisberg for the model using height and diameter² as the predictors, and for the model using height, diameter, and diameter² as the predictors. The index plot for the first-order model matches that in Cook and Wang (1983). Note that a range of 0 to 0.5 for λ is specified for the PRESS plot shown in the sample run. This overrides the default range of 0.1 to 0.5 covering the 95% confidence interval.

The second data set included is the surgical services data from Myers (1990) and may be used to verify the PRESS calculations. Using man-hours per month as the response, 1/surgical cases as the predictor, a specified range of -1 0, and storage of the PRESS statistic values, one can verify that the PRESS values stored in the worksheet for $\lambda = -1$ and 0 match those given in the reference.

EXAMPLES

Results for: MINITAB Trees Data

```
MTB > %BCtrans c1 c2 c3;  
SUBC> range 0 .5;  
SUBC> influence.  
Executing from file: BCtrans.MAC
```

Macro is running ... please wait

Box-Cox Power Transformation Analysis

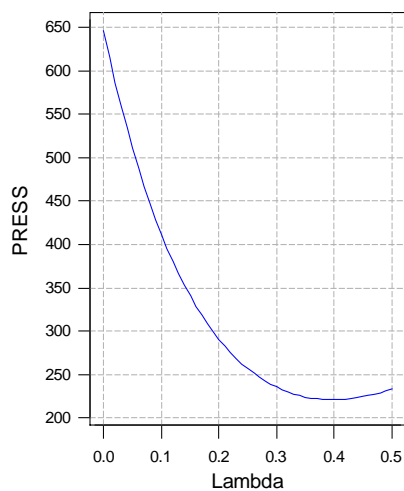
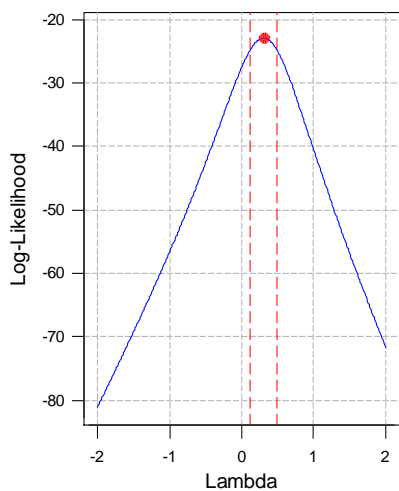
Model Information

```
-----  
Response:      Volume  
Predictor(s):  Diameter , Height  
-----
```

Estimated Lambda: 0.31

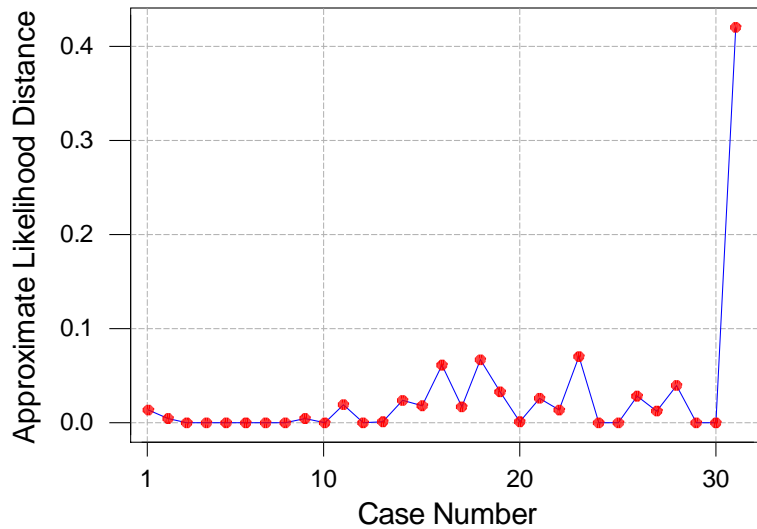
Approximate 95% CI for Lambda: (0.12 , 0.49)

Box-Cox Power Transformation Analysis



Estimated Lambda: 0.31
Approximate 95% CI for Lambda: (0.12 , 0.49)

Cook-Wang Index Plot



Results for: Surgical Services Data

```
MTB > %BCtrans c1 c3;  
SUBC> range -1 0;  
SUBC> presstore c5 c6.  
Executing from file: BCtrans.MAC
```

Macro is running ... please wait

Box-Cox Power Transformation Analysis

Model Information

Response: Man-Hours pe

Predictor(s): 1/Surgical C

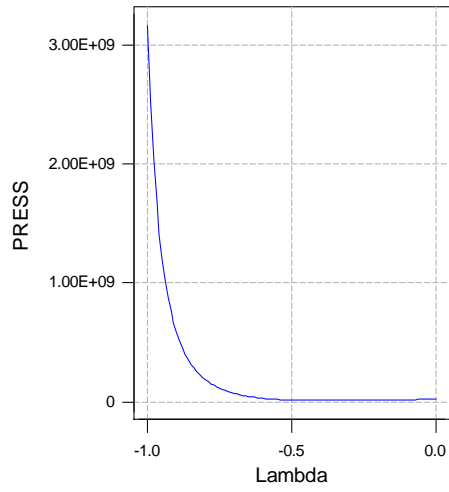
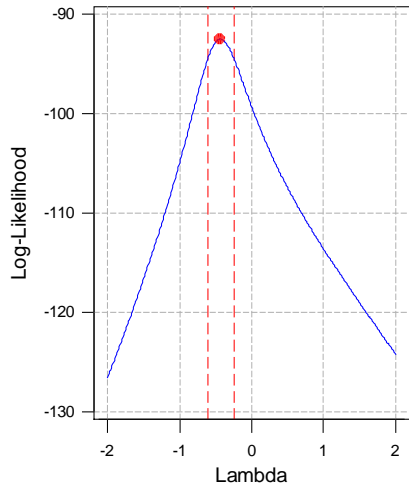
Estimated Lambda: -0.44

Approximate 95% CI for Lambda: (-0.6 , -0.25)

```
MTB > print c8 c9
```

Row	PRESS	Lambda
1	3166829641	-1
2	20684140	0

Box-Cox Power Transformation Analysis



Estimated Lambda: -0.44
Approximate 95% CI for Lambda: (-0.6 , -0.25)

DATA SETS

MINITAB Trees Data

Case	Volume	Diameter	Height
1	10.3	8.3	70
2	10.3	8.6	65
3	10.2	8.8	63
4	16.4	10.5	72
5	18.8	10.7	81
6	19.7	10.8	83
7	15.6	11.0	66
8	18.2	11.0	75
9	22.6	11.1	80
10	19.9	11.2	75
11	24.2	11.3	79
12	21.0	11.4	76
13	21.4	11.4	76
14	21.3	11.7	69
15	19.1	12.0	75
16	22.2	12.9	74
17	33.8	12.9	85
18	27.4	13.3	86
19	25.7	13.7	71
20	24.9	13.8	64
21	34.5	14.0	78
22	31.7	14.2	80
23	36.3	14.5	74
24	38.3	16.0	72
25	42.6	16.3	77
26	55.4	17.3	81
27	55.7	17.5	82
28	58.3	17.9	80
29	51.5	18.0	80
30	51.0	18.0	80
31	77.0	20.6	87

Surgical Services Data

Case	Man-Hours per Month	Surgical Cases
1	1275	230
2	1350	235
3	1650	250
4	2000	277
5	3750	522
6	4222	545
7	5018	625
8	6125	713
9	6200	735
10	8150	820
11	9975	992
12	12200	1322
13	12750	1900
14	13014	2022
15	13275	2155

REFERENCES

1. Cook, R. D. and Wang, P. C. (1983), "Transformations and Influential Cases in Regression," *Technometrics*, **25**, 337-343.
2. Cook, R. D. and Weisberg, S. (1982), *Residuals and Influence in Regression*, London: Chapman and Hall, 66-74.
3. Myers, R. H. (1990), *Classical and Modern Regression with Applications, Second Edition*, Duxbury Press (PWS-KENT Publishing Company), 299-305.