Journal of Econometrics 10 (1979) 57–69. © North-Holland Publishing Company

EXPERIENCE WITH USING THE BOX–COX TRANSFORMATION WHEN FORECASTING ECONOMIC TIME SERIES

Harold L. NELSON, Jr. and C.W.J. GRANGER

University of California at San Diego, La Jolla, CA 92093, USA

Received March 1978, final version received September 1978

Experience using twenty-one actual economic series suggests that using the Box–Cox transform does not consistently produce superior forecasts. The procedure used was to consider transformations $x^{(\lambda)}=(x^{\lambda}-1)/\lambda$, where λ is chosen by maximum likelihood, a linear ARIMA model fitted to $x^{(\lambda)}$ and forecasts produced, and finally forecasts constructed for the original series. A main problem found was that no value of λ appeared to produce normally distributed data and so the maximum likelihood procedure was inappropriate.

1. Introduction

The logarithmic transformation is frequently used by econometricians, either because the change in logarithm of variables approximates percentage changes, or rate of return, or because it is observed that the variability of a series appears to be related to the level, so that using logarithms may produce relationships with more homogeneous residuals. There is no strong reason why the logarithmic transformation should be the best available for this second objective and recently there has been considerable interest in using a broader class of power transformations introduced by Box and Cox (1964). These are given by

$$x^{(\lambda)}=(x^{\lambda}-1)/\lambda, \tag{1.1}$$

which include the logarithmic transformation by taking $\lambda\to 0$. As x has to be positive for use of the transformation, it is sometimes necessary to use $x+\mu$ instead of x in (1.1) where μ is chosen so that $\text{Prob}(x+\mu<0)$ is negligible. However, the majority of levels of economic activity are naturally positive and for the series used in this study it was not necessary to add this extra parameter. An example of the use of these transformations, with seasonal models, is the work by Ansley et al. (1977).

Our objective was to investigate whether or not x_{n+h} could be better forecast from the information set I_n: x_{n-j}, $j\geqq 0$, by the use of the Box–Cox transformations. There has been some controversy on this point in the literature, with Wilson (1973) and Box and Jenkins (1973) coming down in

Table 1

Basic characteristics of the data.[a]

Series identification	Series title	First observation	Last observation	Observation frequency	Total number of observations
A	Average workweek of production workers, ordnance and accessories	1/47	11/74	Monthly	335
B	Steel prices	1/48	12/74	Monthly	324
C	Employees on non-agricultural payrolls, ordnance and accessories	1/45	11/74	Monthly	359
D	Industrial production, primary metals	1/47	11/74	Monthly	335
E	Sales of grocery stores	1/48	10/74	Monthly	322
F	Wholesale prices, lumber and wood products	1/45	11/74	Monthly	359
G	Average weekly initial claims for unemployment insurance, state programs	1/45	11/74	Monthly	359
H	Ratio, price to unit labor cost, manufacturing	1/47	11/74	Monthly	335

I	Index of stock prices, 500 common stocks	1/45	12/74	Monthly	360
J	Index of industrial materials prices	7/46	12/74	Monthly	342
K	Net cash flow, corporate, in current dollars	3/47	9/74	Quarterly	111
L	Unemployment rate, total	1/48	12/74	Monthly	324
M	Sales of retail stores	1/46	11/74	Monthly	335
N	Index of labor cost per unit of output, manufacturing	1/47	11/74	Monthly	335
O	Federal funds rate, unadjusted	2/54	12/74	Monthly	245
P	Gross national product, in current dollars	3/46	9/74	Quarterly	115
Q	Per capita gross national product, in 1958 dollars	3/47	9/74	Quarterly	111
R	Personal income, in current dollars	3/46	9/74	Quarterly	115
S	Gross private domestic investment, total	3/46	9/74	Quarterly	115
T	Net interest	3/46	9/74	Quarterly	115
U	General imports, total	1/48	11/74	Monthly	323

[a]The data was taken from the *Business Conditions Digest* Tape, 1975 edition, supplied by the Bureau of Economic Analysis, U.S. Department of Commerce.

favour of the use of the transformations but with Chatfield and Prothero (1973) finding the extra complication adding insufficient precision. However, all of this published discussion used just a single series for purposes of illustration. We here use twenty-one economic series so that an opinion can be formulated on a much wider sample. The list of series used and their observed time periods are listed in table 1. The members of the sample were chosen in a fairly haphazard fashion. In each case, models were fitted to the first part of the sample and the last twenty terms were reserved for the purpose of evaluating forecasts.

2. Estimation and normality

The estimation procedure used with time series is to form $x_t^{(\lambda)}$ for some λ, to then identify and estimate an ARIMA model for $x_t^{(\lambda)}$, to form the likelihood of the residual $\varepsilon_t^{(\lambda)}$ to this model under the assumption that $\varepsilon_t^{(\lambda)}$ is normally distributed, and finally to search over the λ-space to find the maximum likelihood value, denoted by λ_0. There are a number of difficulties with this procedure, the most obvious of which is that there may exist no λ_0 such that the residuals $\varepsilon_t^{(\lambda)}$ are normally distributed, and even if such a λ_0 does exist, then the residuals will almost certainly not be normally distributed for $\lambda \neq \lambda_0$. As will be discussed later, it is unclear if maximum likelihood is the correct criterion to be applying when selecting the best transformation, as one is not strictly comparing similar quantities. A further problem is that the identified ARIMA model changes as λ changes. This was shown to occur using theoretical arguments by Granger and Newbold (1976). As an example of this occurrence, table 2 shows the estimated autocor-

Table 2

Autocorrelations of $x_t^{(\lambda)}$.

Lag k	$\lambda=0$	$\lambda=0.2$	$\lambda=0.4$	$\lambda=0.6$	$\lambda=0.8$	$\lambda=1$
1	0.03	0.13	0.30	0.50	0.66	0.75
2	0.02	0.09	0.24	0.42	0.57	0.66
3	−0.11	−0.03	0.13	0.32	0.48	0.58
4	0.11	0.15	0.25	0.39	0.51	0.58
5	−0.19	−0.09	0.09	0.31	0.48	0.58
6	−0.04	0.01	0.13	0.31	0.46	0.56
7	−0.10	−0.02	0.13	0.32	0.47	0.56
8	0.13	0.16	0.24	0.35	0.46	0.52
9	0.11	0.14	0.34	0.35	0.45	0.51
10	−0.03	0.02	0.14	0.29	0.42	0.50
12	0.18	0.23	0.33	0.43	0.50	0.54
15	0.06	0.11	0.20	0.29	0.35	0.38
18	0.08	0.10	0.15	0.22	0.28	0.31
21	−0.01	−0.01	0.04	0.11	0.17	0.20
24	−0.06	−0.01	0.07	0.15	0.20	0.21

relations of the changes in series T (net interest) for various values of λ, where $\rho_k^{(\lambda)}$ = correlation $(\Delta x_t^{(\lambda)}, \Delta x_{t-k}^{(\lambda)})$.

The autocorrelations are based on 95 observations, so the approximate 95% confidence interval is ± 0.205. It is seen from table 2 that the changes in the logarithm of the series, corresponding to $\lambda=0$, might well be identified as white noise, but the autocorrelations increase in size as λ increases and for λ larger than about 0.5 a further differencing might be thought appropriate. For example, using also the partial autocorrelations, which are not shown, with $\lambda=0.8$ or 1.0 an ARIMA (1, 2, 0) model is identified compared to ARIMA (0, 1, 0) for $\lambda=0$ or 0.2. For this particular series, the maximum likelihood method produces $\lambda_0=0.71$.

λ_0 is estimated by calculating values of the log-likelihood for various values of λ. Each of these calculations is quite costly since it involves a non-linear estimation of the remaining parameters. To economize we employed a Fibonacci search procedure after the initial diagnostic estimations, which are necessary to check model adequacy as λ varies and obtain upper and lower bounds for λ_0. Walsh (1975) gives a comprehensive development of this search procedure and related numerical optimization techniques.

Another technique that has been suggested by Box and Cox (1964) and Zarembka (1974) is data-scaling which removes the component of the log-likelihood function arising from the Jacobian. This makes the optimization amenable to a least-squares algorithm, but the resulting increase in the number of exponentiations performed seems to negate the potential cost reduction.

It was previously mentioned that the 'maximum likelihood' λ_0 may not correspond to normally distributed residuals. That this is a real problem is illustrated in table 3, which shows coefficients of skewness and kurtosis, β_s and β_k, for the residuals of ARIMA models using $x_t^{(\lambda)}$ with $\lambda=0$, 1 and λ_0. These coefficients are defined by

$$\beta_s = m_3/m_2^{3/2},$$

and

$$\beta_k = (m_4/m_2^2) - 3,$$

where

$$m_j = \frac{1}{n}\sum_{i=1}^{n} (x_t - \bar{x})^j,$$

for a sample of size n, $\bar{x}$ being the sample mean. Table 3 also shows the values of λ_0 found by the maximum likelihood procedure.

The majority of the β_k values are significantly positive, being greater than four times the estimated standard deviation of β_k. The only exceptions are

Table 3
λ_0, coefficient of skewness and kurtosis.

Series		β_s			β_k		
	λ_0	λ_0	$\lambda=0$	$\lambda=1$	λ_0	$\lambda=0$	$\lambda=1$
A	6.62	−0.4	−1.2	−1.0	4.5	12.9	10.6
B	0.09	−0.1	−0.1	0.0	0.7	0.7	1.5
C	−0.23	−11.9	−13.0	−10.9	188.8	211.9	174.0
D	2.81	−0.6	−1.1	−0.7	9.0	22.8	17.7
E	0.17	0.8	0.9	1.4	3.1	4.0	5.9
F	0.18	1.4	1.7	0.6	13.6	15.7	15.8
G	1.14	2.8	4.9	3.1	31.1	61.3	34.7
H	−1.52	0.2	0.2	0.2	1.3	1.2	1.2
I	0.10	−0.7	−0.8	−0.6	2.1	2.5	3.8
J	0.14	1.8	1.9	0.9	17.2	18.7	10.4
K	0.88	−0.1	−0.2	−0.0	0.1	0.4	0.2
L	0.10	0.1	0.1	−0.1	3.2	3.0	6.8
M	0.32	0.3	0.3	0.6	1.8	2.6	2.2
N	0.99	0.0	0.1	0.0	1.4	1.4	1.4
O	1.02	0.1	1.0	0.1	3.3	11.0	3.3
P	0.63	−0.2	0.2	−0.2	−0.1	1.2	0.1
Q	2.46	−0.0	0.1	−0.0	−0.1	1.3	0.4
R	1.57	0.3	0.2	0.2	0.5	6.4	1.7
S	0.75	−0.5	0.2	−0.6	−0.1	1.4	0.6
T	0.71	−0.1	0.8	−0.2	1.4	4.2	0.9
U	0.06	−0.9	−0.8	−1.0	7.3	6.8	15.1

series K, P, Q, S (all three λ values), $B(\lambda_0$ and $\lambda=0)$, $R(\lambda_0$ and $\lambda=1)$ and $T(\lambda_0$ and $\lambda=1)$. Similarly, many β_s values are greater than twice the standard deviation of β_s in magnitude. For λ_0, series A, C, D, E, F, G, H, I and U have β_s values clearly significantly non-zero. Lowest β_s values occur 8 times with $\lambda=\lambda_0$, 10 times with $\lambda=1$ and 3 times with $\lambda=0$. Lowest β_k values occur 12 times with $\lambda=\lambda_0$, 5 times with $\lambda=1$ and 4 times with $\lambda=0$. It thus seems that using the Box–Cox transformation takes the residuals nearer to normality but rarely achieves it. The only series that appear to have normal residuals with $\lambda=\lambda_0$ are K, P, Q, R and S. The results for β_s do not agree with the theory of Draper and Cox (1969) who indicate that even when no value of λ achieves normality, then the distribution of errors corresponding to λ_0 will be more symmetric.

Turning to the λ_0 values, 7 are seen to be near zero (in the range ± 0.2), 4 are near one (range 0.8 to 1.2), 2 are less than -0.2 and 4 are larger than 1.2. The median value is 0.63 and the range is -1.52 to 6.62. Given that the series are generally non-normal, it is difficult to justify the use of a likelihood ratios test to see if λ_0 is significantly different from 0 or 1. However, if such a test is used, then the λ_0 values of 6.62 (series A), 2.81 (D), 2.46 (Q) and 1.57 (R) are 'significantly' greater than 1 and -0.23 (C) is 'significantly' less than

0, but the value -1.52 (H) is not significantly different from 0. In any case, it is clear that the search for λ_0 has to extend outside the range 0 to 1.

3. Forecasting performance

The main reason for using the Box–Cox transformation, according to Box and Jenkins (1973), is to produce improved forecasts. In this section, the usefulness of the transformation in this respect is evaluated for the 21 series in our sample. Using the post-sample data reserved for the purpose, 20 pieces for each series, as many h-step forecasts as could be evaluated were produced. Thus, 20 one-step forecasts, 19 two-step forecasts, 18 three-step and so forth, up to 11 forecasts with a horizon of ten steps were made, the resulting forecast errors were recorded and sums of squared residuals formed. Predictions with $\lambda=\lambda_0$, $\lambda=1$ and $\lambda=0$ were produced and compared. However an inverse transformation is required to produce equivalent forecasts as, if a model is produced for $x_t^{(\lambda)}$ with $\lambda \neq 1$, then this model will only produce directly forecasts $f_{n,h}^{(\lambda)}$ of $x_{n+h}^{(\lambda)}$ rather than of the actual series x_{n+h} as required. If a government department requests a forecast of unemployment, they would not be interested in being told that $\lambda_0 = 1/2$ and then being given a forecast of the square root of unemployment. A naive procedure is to note that

$$\begin{aligned} x_t &= (\lambda x_t^{(\lambda)} + 1)^{1/\lambda}, && \lambda \neq 0, \\ &= \exp(x_t^{(0)}), && \lambda = 0, \end{aligned}$$

and then to use as a forecast of x_{n+h} made at time n,

$$g_{n,h}^{(\lambda)} = (\lambda f_{n,h}^{(\lambda)} + 1)^{1/\lambda}, \qquad \lambda \neq 0, \tag{3.1}$$

$$= \exp(f_{n,h}^{(0)}), \qquad \lambda = 0. \tag{3.2}$$

However, it is easy to show that these forecasts are frequently biased or sub-optimal, producing forecast errors with non-zero means. If $x_t^{(\lambda)}$ is normally distributed, Granger and Newbold (1976) have determined the extent of these biases and have found the optimal, unbiased forecasts $f_{n,h}$, in terms of $f_{n,h}^{(\lambda)}$ and $\sigma_{n,\lambda}^2$ the variance of the h-step forecast errors $x_{n+h}^{(\lambda)} - f_{n,h}^{(\lambda)}$. For example, if $\log x_t = x_t^{(0)}$ is normally distributed, it was found that the optimal h-step forecast of x_t is

$$f_{n,h,0} = \exp(f_{n,h}^{(0)} + \sigma_{h,0}^2/2). \tag{3.3}$$

Unfortunately, there is no closed form for $f_{n,h}$ for general λ and it has to be obtained from the integral

$$f_{n,h,\lambda}=\frac{1}{\sigma_{h,\lambda}\sqrt{2\pi}}\int_{-\infty}^{\infty}\exp\left[-1/2\left(\frac{y-f_{n,h}^{(\lambda)}}{\sigma_{h,\lambda}}\right)^{2}\right](\lambda y-1)^{1/\lambda}\mathrm{d}y. \tag{3.4}$$

These formulae are based on an assumption of normality for $x_t^{(\lambda)}$ and may be quite inappropriate under non-normality. Nevertheless, five different forecasts made at time n of x_{n+h} were considered:

A Naive using $\lambda=\lambda_0$, from (3.1).
B Non-biasing using $\lambda=\lambda_0$, from (3.4).
C Naive using $\lambda=0$, the logarithmic case, from (3.2).
D Non-biasing using $\lambda=0$, from (3.3).
E Linear case, $\lambda=1$, $f_{n,h}^{(1)}$ derived directly from ARIMA model for x_t.

With twenty-one series, five forecasting procedures and up to ten forecasting horizons there are a large number of results to report. Only summaries will be presented here, full details may be found in Nelson (1977). For each individual series, comparisons of forecast methods were made using the root mean squared error

$$\mathrm{RMSE}(i,h,r)=\left[\frac{1}{20-h+1}\sum_{t=r+h}^{n+20}(x_{i,t}-\hat{f}_{t-h,h,i,r})^2\right]^{\frac{1}{2}},$$

for series i, horizon h and forecast method r. To compare alternative forecasting methods, the geometric means of ratios of root mean squares are used,

$$\mathrm{GM}(h,r,s)=\left[\prod_{h=1}^{21}\frac{\mathrm{RMSE}(i,h,r)}{\mathrm{RMSE}(i,h,s)}\right]^{\frac{1}{21}},$$

for horizon h and methods r and s.

Table 4 shows which individual methods had the lowest root mean squares for each of the twenty-one series in the sample. For each of the three forecast horizons shown, the most successful method is D which uses logarithms of the data and then the non-biasing inversion formula (3.3). The methods using the 'optimum' transformation with λ_0, A and B, do not perform particularly well.

Table 5 shows the proportions of times that one forecasting method has a lower root mean squared error than another for the twenty-one series and for forecasting horizons $h=1$, 5 and 10. The interesting comparisons are between the transformed series against the untransformed ones. The use of the unbiasing formulae do generally seem to be worthwhile, although not

Table 4
The best forecasting methods for $h=1$, 5 and 10 for all twenty-one series.

Series number	λ_0	$h=1$	$h=5$	$h=10$
A	6.625	C	C	B
B	0.088	C	D	D
C	−0.231	C	C	E
D	2.811	C	B	B
E	0.168	D	D	D
F	0.180	A	E	E
G	1.142	D	D	D
H	−1.515	B	B	B
I	0.103	E	E	E
J	0.139	E	E	D
K	0.877	D	D	D
L	0.103	D	D	D
M	0.321	E	D	D
N	0.986	D	D	D
O	1.028	A	A	A
P	0.630	D	D	D
Q	2.465	A	D	D
R	1.573	E	E	D
S	0.747	D	D	D
T	0.713	D	E	E
U	0.063	D	D	D
Count	A	3	1	1
	B	1	2	3
	C	4	2	0
	D	9	11	13
	E	4	5	4

dramatically so, as the proportion of times that method B is superior to A is about 0.6 for each h value and similarly method D is superior to C at about the same proportion. There is also some evidence that biasing forecasts on transformed data is better than staying with the untransformed data. The proportions in the column *E* are under 0.5 in every case but one and method E does particularly poorly compared to C and D. If the five methods are ranked on RMSE for each of the 21 series with $h=1$, the average ranks found are:

Method	Average rank
A	3.095
B	2.904
C	2.809
D	2.809
E	3.381

Table 5
Entries are proportions of cases in which the column outperforms the row methods.

	A	B	C	D	E
(i) $h=1$					
A	–	0.571	0.476	0.524	0.476
B	0.429	–	0.571	0.476	0.476
C	0.429	0.524	–	0.571	0.286
D	0.524	0.476	0.429	–	0.381
E	0.524	0.714	0.714	0.619	–
(ii) $h=5$					
A	–	0.667	0.524	0.667	0.476
B	0.333	–	0.619	0.667	0.524
C	0.381	0.476	–	0.619	0.333
D	0.333	0.333	0.381	–	0.429
E	0.476	0.524	0.667	0.571	–
(iii) $h=10$					
A	–	0.619	0.476	0.667	0.333
B	0.381	–	0.667	0.667	0.381
C	0.333	0.524	–	0.667	0.296
D	0.333	0.333	0.333	–	0.333
E	0.619	0.667	0.714	0.667	–

Thus, again the methods C and D are somewhat superior to B and E is generally the poorest.

These rankings and proportions provide interesting descriptive statistics but do not indicate the significance or extent of any differences. Table 6 shows the geometric means of ratios of RMSE, denoted by $\mathrm{GM}(h,r,s)$ for $h=1$, 5 and 10. The results suggest that, in practice, there is very little to be gained by using transformed data rather than untransformed, a gain of about 2% at most in reduction of root mean squared error on the average when $h=1$, although possibly more with $h=5$ or 10.

In summation, little improvement in forecast performance is obtained by using the Box–Cox optimum transformation over using untransformed data, using the logarithmic transformation is somewhat better on average than the forecasts from untransformed data and the unbiasing formula is somewhat worthwhile. None of the differences observed are consistently convincing and forecasts based just on untransformed data are likely to be very little, if at all, inferior to forecasts resulting from more costly transformation techniques.

It should be pointed out that the method selecting the optimal transform parameter λ_0 used here, which is similar to that used by most other workers, is not necessarily the most appropriate. It could be suggested that λ_0 should

Table 6

Geometric means of ratios of root mean squares.

$s=$	A	B	C	D	E
(i) *values of* GM$(1,r,s)$					
$r=$A	1.0000	0.9998	1.0132	1.0101	0.9972
B	1.0002	1.0000	1.0130	1.0099	0.9970
C	0.9871	0.9869	1.0000	0.9969	0.9841
D	0.9902	0.9900	1.0031	1.0000	0.9872
E	1.0031	1.0028	1.0161	1.0129	1.0000
(ii) *values of* GM$(5,r,s)$					
$r=$A	1.0000	0.9898	1.0566	1.0377	1.0039
B	1.0103	1.0000	1.0459	1.0271	0.9937
C	0.9561	0.9464	1.0000	0.9821	0.9501
D	0.9836	0.9637	1.0182	1.0000	0.9674
E	1.0064	0.9461	1.0525	1.0357	1.0000
(iii) *values of* GM$(10,r,s)$					
$r=$A	1.0000	0.9597	1.1222	1.0590	1.0399
B	1.0420	1.0000	1.0770	1.0164	0.9980
C	0.9285	0.8911	1.0000	0.9437	0.9266
D	0.9839	0.9443	1.0597	1.0000	0.9819
E	1.0020	0.9617	1.0792	1.0184	1.0000

be chosen so that, after applying the unbiasing formula, the best forecasts of x_t are obtained. The forecasting results suggest that λ_0 chosen by this criterion, which would be an expensive one to apply, would be quite different from the values for λ_0 obtained by the maximum likelihood criterion. It should be noted that in a non-forecasting context, Cleveland, Douglas and Terpenning (1978) have used a 'best-model' criterion for choosing λ_0.

4. Simulation experiments

To throw some more light on the problems found with real data, similar calculations were performed with synthetic, computer generated data. The data were all generated from the model

$$\nabla x_t^{(\lambda)} = 0.6 \nabla x_{t-1}^{(\lambda)} + m + e_t,$$

where e_t is a zero-mean, white noise with variance σ_e^2, and m, σ_e and the starting value for the series are all taken from series 0 (Federal funds rate). To form e_t uniform random numbers were obtained from the CDC FTN4 intrinsic and for most experiments were then converted to normally distributed random numbers using the polar conversion method. Each experiment was performed on a sample of 100 generated series. Each generated series contained 450 observations, but observations 1 to 300 were discarded to avoid starting-up, transient behaviour, observations 301 to 400 were used for estimation and diagnostic checking and the final 50 terms used to evaluate forecasts.

Once again, the quantity of empirical results are too great to be reported fully, and so only the more interesting or important results will be presented. More details may be found in Nelson (1977).

The first experiment has $\lambda=1$ and Gaussian disturbances with $\sigma_e=4$ and $m=8$. The maximum likelihood estimates of λ_0 had mean 0.947, standard deviation of 0.565, a minimum of -0.464 and a maximum of 2.464. The coefficients of skewness and kurtosis were both small and the distribution of λ_0 appears to be Gaussian. The size of the standard deviation of λ_0, and the consequent range, is disturbing. The parameter of the AR process using $\lambda=\lambda_0$ had a mean of 0.565, compared to a true value of 0.6, and standard deviation 0.084. As might be expected with $\lambda=1$, there was virtually no difference between forecasting methods A, B and E, but methods C and D were inferior.

The second experiment was similar to the first, but the disturbances were chi-squared with three degrees of freedom. The results were very similar to those in the first experiment, with average λ_0 0.991, but now the standard deviation of estimated λ_0 is 0.88 and taking $\lambda=\lambda_0$ does not move the residuals nearer to normality.

The third experiment considers the logarithmic form with $\lambda_0=0$, $m=0.01$ and $\sigma_e=0.01$ and e_t Gaussian. The mean estimated λ_0 is -0.002 with standard deviation of 0.085, a minimum of -0.200 and a maximum of 0.218. Forecast methods C and D were equally superior to A and B, with E the worst for all three forecast horizons as expected.

The fourth and final experiment used $\lambda=0.5$, $\sigma_e=0.2$ and $m=0.4$. Maximum likelihood estimates of λ_0 had mean 0.472, standard deviation 0.298, minimum of -0.275, maximum of 1.267 and low coefficients of skewness and kurtosis. Of the five forecasting methods used previously, methods A and B are the best using λ_0 but methods using the true $\lambda=0.5$ were better yet.

In general, the results contain few surprises. It seems that when the necessary underlying assumptions are true, the Box–Cox transformation works well and does produce superior forecasts when a transformation is really justified. It was found that the unbiasing procedure had little effect in improving forecasts for the cases considered in the simulations. The one case using non-normal disturbances did not behave differently. It should be noted however that the real data used in previous sections was generally much less near normal than the data used in the second experiment. One disturbing feature of the simulation results is the observation that the standard deviation of the maximum likelihood estimate of λ_0 increases as the true value of λ increases. When estimating λ_0 with 100 pieces of data and with true $\lambda = 1$, the approximate 95% confidence interval for λ_0 is -0.183 to 2.077 with normal disturbances and -0.769 to 2.751 with chi-squared disturbances. These ranges are too wide for interpretations of the maximum likelihood estimate of λ_0 to be made with any comfort.

5. Conclusion

Although in some circumstances there might be some advantage in using the Box–Cox transformation, the evidence when using actual data is that the extra inconvenience, effort and cost is usually such as to make the use of these transformations not worthwhile. The main problem seems to be the extreme non-normality of actual economic data, and the use of the transformation does not dramatically reduce this problem.

References

Ansley, C.F., W.A. Spivey and W.J. Wrobleski, 1977, A class of transformations for Box–Jenkins seasonal models, Applied Statistics 26.

Box, G.E.P. and D.R. Cox, 1964, An analysis of transformations, Journal of the Royal Statistical Society B 26, 211–252.

Box, G.E.P. and G.M. Jenkins, 1970, Time series analysis (Holden-Day, San Francisco, CA).

Box, G.E.P. and G.M. Jenkins, 1973, Some comments on a paper by Chatfield and Prothero and on a review by Kendall, Journal of the Royal Statistical Society A 136, 337–345.

Chatfield, C. and D.L. Prothero, 1973, Box–Jenkins' seasonal forecasting: Problems in a case study, Journal of the Royal Statistical Society A 136, 295–314.

Cleveland, W.S., D.M. Dunn and I.J. Terpenning, 1978, SABL – A robust seasonal adjustment procedure with graphical methods for interpretation and diagnosis, in: Proceedings of NBER-Bureau of Census Conference on Seasonal Adjustment, forthcoming.

Draper, N.R. and D.R. Cox, 1969, On distributions and their transformation to normality, Journal of the Royal Statistical Society B 31, 472–476.

Granger, C.W.J. and P. Newbold, 1976, Forecasting transformed variables, Journal of the Royal Statistical Society B 38, 189–203.

Nelson, H.L., Jr., 1977, The use of Box–Cox transformations in economic time series analysis: An empirical study, Ph.D. thesis (Economics Department, University of California, San Diego, CA).

Walsh, G.R., 1975, Methods of optimization (Wiley, New York).

Wilson, G.T., 1973, Comment following Chatfield and Prothero (1973).

Zarembka, P., 1974, Transformation of variables in econometrics, in: P. Zarembka, ed., Frontiers of econometrics (Academic Press, New York).