A PRIMER ON BOX-COX ESTIMATION

John J. Spitzer*

I. Introduction

THE power transformation introduced by Box and Cox (1964) and given by

$$y^{(\lambda)} = \begin{cases} (y^{\lambda} - 1)/\lambda & \lambda \neq 0 \\ \ln y & \lambda = 0 \end{cases} \tag{1}$$

has been extensively used in recent years. Transformed variables can be included in a "linear" function so that generalized models of the form

$$y^{(\lambda_1)} = \beta_1 + \beta_2 X_2^{(\lambda_2)} + \ldots + \beta_k X_k^{(\lambda_k)} + \epsilon \tag{2}$$

can be specified and estimated. On occasion, neither a priori reasoning nor theory clearly dictate the correct functional form (transformation) which an additive model should assume. With the Box-Cox transformation, the functional form is dictated by the parameters, λ_i, which are themselves estimated. Note that if $\lambda_1 = 1$ in (2), then $y^{(\lambda)}$ enters the equation linearly; also, $y^{(0)}$ enters (2) as $\ln y$, and $y^{(-1)}$ enters (2) as the reciprocal of y. Thus the estimation procedure itself chooses the transformations which best fit the data. Furthermore, hypothesis tests can be made on the estimated λ_i in order to determine if alternative functional forms (transformations) are also consistent with the data. (See the appendix for a note on discriminating between functional forms.)

Estimation of (2) requires the maximization of a nonlinear likelihood function which can be extremely complicated. Since computer programs for maximizing such complex functions may not be readily available, the estimation of generalized functional forms such as (2) may be impeded.

It may not be generally recognized that estimation of the parameters of (2) can be accomplished in at least four different ways. This paper will look at four alternative ways of estimating the parameters β_i, λ_i and σ^2. Each approach can be made to yield identical parameter estimates, and identical estimates of the covariance matrix of the parameter estimates. In section II, the general problem will be addressed and the likelihood function derived. Section III will look at each estimation approach. Section IV will conclude the paper.

Problems of estimation only are dealt with in this paper. Furthermore, the approximate normality of the error terms is assumed throughout. For a discussion of estimation methodology when the error terms are truncated normal, see Poirier (1978). For a discussion of the interpretation of estimated coefficients in Box-Cox models, see Poirier and Melino (1978) or Huang and Kelingos (1979).

Received for publication December 22, 1980. Revision accepted for publication September 15, 1981.

* State University of New York, Brockport.

The author wishes to thank the referees for their helpful suggestions which led to improvement in the final version.

II. The Likelihood Function

A. *Deriving the Likelihood Function and Covariance Matrix*

In this section the likelihood function of the sample and an estimator of the variance-covariance matrix of the estimators will be derived. Initially, it will be assumed that (2) is the model to be estimated. Subsequent parts of the paper will use a simpler model for heuristic purposes.

Under the assumption that there exists some λ for which ϵ in (2) is approximately normally distributed with mean zero and variance, σ^2, the density function of the t^{th} observation of ϵ is given by

$$f(\epsilon_t) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp(-\epsilon_t^2/2\sigma^2). \tag{3}$$

The likelihood function is then

$$L^*(\beta_i, \lambda_i, \sigma^2; X, y) = \prod_{t=1}^{T} f(\epsilon_t) y_t^{\lambda_1 - 1} \tag{4}$$

where the last term is the Jacobian of the transformation from ϵ to y. The logarithm of (4) is given by

$$L = \ln L^* = k_1 - T/2 \ln \sigma^2 - (2\sigma^2)^{-1} \sum_{t=1}^{T} (y^{(\lambda_1)} - \beta_1 - \beta_2 X_2^{(\lambda_2)} - \ldots - \beta_k X_k^{(\lambda_k)})^2 + (\lambda_1 - 1) \sum_{t=1}^{T} \ln y, \tag{5}$$

where k_1 is a constant. Since (5) is a monotonic transformation of (4), both functions will be maximized for the same parameter values. Maximization of (5) under the assumption of approximate normality of the ϵ_t obtains estimators β_i, λ_i and σ^2 which are Best Asymptotically Normal (BAN) under general regularity conditions. Furthermore, the asymptotic covariance matrix of the parameter estimates is given by

$$-E(\partial^2 L/\partial\theta\partial\theta')^{-1} \tag{6}$$

where $\theta' = \{\beta_1 \ldots \beta_k \lambda_1 \ldots \lambda_k \sigma^2\}$. This matrix is the Cramer-Rao lower bound. Irrespective of the approach taken to estimation of the parameters of (2), a consistent estimate of the covariances of the parameter estimates from (6) may be obtained. In practice, the expected value of the second derivative matrix cannot be evaluated for functions which contain the Box-Cox transformation; the expectations in (6) are too complex to evaluate. Instead, the observed (computed) matrix

$$(-\partial^2 L/\partial\theta\partial\theta')^{-1} \tag{7}$$

is used. Goldfeld and Quandt (1972, pp. 63–64) indicate that (6) is consistently estimated by (7) if the estimators are sufficient. Thus, use of (7) is an acceptable practice.

B. *Examining Simpler Models*

Several different versions of (2) are possible. Each representation places certain restrictions on some of the λ making the estimation problem somewhat simpler. For example,

$$y^{(\lambda)} = \beta_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \epsilon \tag{8}$$

$$y^{(\lambda)} = \beta_1 + \beta_2 X_2^{(\lambda)} + \ldots + \beta_k X_k^{(\lambda)} + \epsilon \tag{9}$$

$$y^{(\lambda_1)} = \beta_1 + \beta_2 X_2^{(\lambda_2)} + \ldots + \beta_k X_k^{(\lambda_2)} + \epsilon \tag{10}$$

$$y^{(\lambda_1)} = \beta_1 + \beta_2 X_2^{(\lambda_2)} + \ldots + \beta_k X_k^{(\lambda_k)} + \epsilon \tag{11}$$

are all possible combinations. In (8), only the y values are power transformed. In (9), all X's and the y value are transformed by the same value of λ. In (10), all X's are transformed in the same way, but the y value is subject to a different transformation. Lastly, (11), which is the most general case, allows for different Box-Cox transformations on all the variables in the model. Equation (11) is identical to (2).

In section III, discussion will be focused on estimating the parameters in the simpler model represented by (8). The only power transformation is on the dependent variable y. This specification makes presentation of the estimation problem simpler, but, as will be shown, does not change any of the essential conclusions. Whether (8) or (2) is estimated will basically change only the dimension of the problem; neither the general approach nor the derivation of the covariance matrix is changed.

Using the more compact notation of matrix algebra, equation (8) is

$$y^{(\lambda)} = X\beta + \epsilon \tag{12}$$

where $y^{(\lambda)}$ is a $T \times 1$ vector of transformed observations on the dependent variable; X is a $T \times K$ matrix of observations on the independent variables, where the first column is a vector of ones. Unless this vector is included, the log-likelihood function is not scale invariant (Schlesselman, 1971). The vector β is a $K \times 1$ coefficient vector of the X's, and ϵ is a $T \times 1$ vector of disturbances, assumed to be approximately normally distributed. The likelihood function to be maximized for model (8) is given by

$$L(\beta,\lambda,\sigma^2; X,y) = k_1 - T/2 \ln \sigma^2 + (\lambda - 1) i' \ln y - (y^{(\lambda)} - X\beta)'(y^{(\lambda)} - X\beta)/2\sigma^2. \tag{13}$$

where $i' = [1\ 1\ 1\ 1 \ldots 1]$. The third term on the right-hand side of (13) is the log of the Jacobian.

III. Estimation Approaches

Four approaches to the estimation of (8) are presented below. Each is equivalent to maximum likelihood (ML) estimation.

1. Maximizing the full log-likelihood function—(13).
2. Maximizing the concentrated log-likelihood function.
3. Maximizing a function of the transformed sum of squares function. This method is identical to nonlinear least squares (NLSQ).
4. Minimizing the transformed sum of squares function by repeated use of ordinary least squares (iterated ordinary least squares—IOLS).

A. *Maximizing the Full Log-Likelihood Function*

Equation (13) is to be maximized with respect to β, λ and σ^2. Let $\epsilon_\lambda = \partial\epsilon/\partial\lambda$ and $\epsilon_{\lambda\lambda} = \partial^2\epsilon/\partial\lambda^2$. The first order conditions for a maximum require that

$$\frac{\partial L}{\partial \beta} = \sigma^{-2} X'\epsilon = 0 \tag{14}$$

$$\frac{\partial L}{\partial \lambda} = -\sigma^{-2}\,\epsilon'\epsilon_\lambda + i' \ln y = 0 \tag{15}$$

$$\frac{\partial L}{\partial \sigma^2} = -(T/2)\sigma^{-2} + \epsilon'\epsilon/(2\sigma^4) = 0. \tag{16}$$

The second order conditions for a local maximum require that the matrix of second derivatives be negative definite. Let $\theta' = [\beta'\ \lambda\ \sigma^2]$. Then the second derivative matrix is a $(K+2)\times(K+2)$ matrix given by

$$\frac{\partial^2 L}{\partial\theta\partial\theta'} = -\sigma^{-2} \times \begin{bmatrix} X'X & -X'\epsilon_\lambda & 0 \\ & \epsilon'\epsilon_{\lambda\lambda} + \epsilon'_\lambda\epsilon_\lambda & -i'\ln y \\ \text{symmetric} & & \epsilon'\epsilon/(\sigma^2)^2 - T/(2\sigma^2) \end{bmatrix} \tag{17}$$

Use has been made of the first order computations in the derivation of (17). For example, the $(K\times 1)$ vector of zeros in (17) results from the vector

$$\partial^2 L/\partial\beta\partial\sigma^2 = -(\sigma^{-2})^2\, X'\epsilon = -\sigma^{-2}\frac{\partial L}{\partial\beta} = 0.$$

The substitution of $i'\ln y$ in (17) is obtained from the first order condition given in (15). The inverse of the negative of (17), evaluated at L_{max} is the estimated covariance matrix of the parameter estimates. The maximization problem can be simplified, as will be shown in the next section.

B. *Concentrating the Log-Likelihood Function*

The parameter σ^2 can be solved for directly from equation (16) in terms of the other parameters and the data. Thus, let

$$\hat\sigma^2 = T^{-1}\epsilon'\epsilon = T^{-1}(y^{(\lambda)} - X\beta)'(y^{(\lambda)} - X\beta). \tag{18}$$

It is then not necessary to estimate σ^2 simultaneously with β and λ. The parameter σ^2 is a "nuisance" variable which can be eliminated or "concentrated out" of the likelihood function, thus reducing the dimension of the estimation problem by one parameter. Substituting (18) into (13), the concentrated log-likelihood function becomes

$$L(\beta,\lambda;X,y) = k_2 - T/2 \ln\hat\sigma^2 + (\lambda-1)\, i'\ln y \tag{19}$$

since the last term of (13) is now a constant equal to $-T/2$ and is included in k_2. First order conditions for a maximum of (19) require that

$$\partial L/\partial\beta = -T/(2\hat\sigma^2)\frac{\partial\hat\sigma^2}{\partial\beta} = \hat\sigma^{-2}X'\epsilon = 0 \tag{20}$$

$$\partial L/\partial\lambda = -T/(2\hat\sigma^2)\frac{\partial\hat\sigma^2}{\partial\lambda} + i'\ln y = 0 \tag{21}$$

$$= -\hat\sigma^{-2}\epsilon'\epsilon_\lambda + i'\ln y = 0.$$

Second order conditions require the matrix

$$\partial^2 L/\partial\theta\partial\theta' = -\hat\sigma^{-2}\begin{bmatrix} X'X & -X'\epsilon_\lambda \\ -\epsilon'_\lambda X & (\epsilon'\epsilon_{\lambda\lambda} + \epsilon'_\lambda\epsilon_\lambda - 2T^{-1}\hat\sigma^2(i'\ln y)^2] \end{bmatrix} \tag{22}$$

to be negative definite. The inverse of the negative of (22) is the estimated covariance matrix of $\hat\theta' = [\hat\beta'\ \hat\lambda]$. It may be easily shown that the covariance matrix obtained from (22) is identical to the covariance matrix for $\hat\beta$ and $\hat\lambda$ obtained from (17). Write (17) in partitioned form as

$$Z = \begin{bmatrix} A & B \\ B' & C \end{bmatrix}$$

where A is a $(K+1)\times(K+1)$ matrix; B is a $(K+1)\times 1$ vector, and C is (1×1). The full covariance matrix is, of course,

$$-Z^{-1} = -\begin{bmatrix} A & B \\ B' & C \end{bmatrix}^{-1} = \begin{bmatrix} U & V \\ V' & W \end{bmatrix}.$$

The covariance matrix for the $\hat\beta$s and $\hat\lambda$ is $U = -(A - BB'/C)^{-1}$. But $(A - BB'/C)$ is exactly (22).

C. *Transforming the Problem to Nonlinear Least Squares*

The estimation problem may be further simplified by applying a scaling trick originally noted by Zarembka (1968, note 8). Let

$$y^{(\lambda)} = X\beta + \epsilon \tag{23}$$

be multiplied through by $\dot{y}^{-\lambda}$, where $\dot{y}$ is the geometric mean of the sample y's. That is,

$$\dot{y} = \left(\prod_{t=1}^{T} y_t\right)^{1/T} = \exp\,(T^{-1}\Sigma \ln y).$$

This obtains

$$\begin{aligned}[(y/\dot{y})^{\lambda} - \dot{y}^{-\lambda}]/\lambda &= X\beta\dot{y}^{-\lambda} + \epsilon\dot{y}^{-\lambda}\\ [(y/\dot{y})^{\lambda} - 1]/\lambda &= \lambda^{-1}(\dot{y}^{-\lambda} - 1) + X\beta\dot{y}^{-\lambda}\\ &\quad + \epsilon\dot{y}^{-\lambda}.\end{aligned} \tag{24}$$

The original equation has been transformed into

$$y^{*(\lambda)} = X\beta^{*} + \epsilon^{*} \tag{25}$$

where $y^{*}_{t} = y_t/\dot{y}$ and $\beta^{*} = \dot{y}^{-\lambda}[\beta_1 - \dot{y}^{(\lambda)}\,\beta_2\,\beta_3 \ldots \beta_k]'$. The log-likelihood function is merely

$$\begin{aligned}L(\beta^{*},\lambda;X,y^{*}) &= k_2 - T/2 \ln \hat{\sigma}^{*2}\\ &\quad + (\lambda - 1)\, i' \ln y^{*}\\ &= k_2 - T/2 \ln \hat{\sigma}^{*2}\end{aligned} \tag{26}$$

since $i' \ln y^{*} = 0$. Maximizing L is clearly equivalent to minimizing $\hat{\sigma}^{*2}$; the problem is now a nonlinear least squares problem. Initially, let us focus on maximizing (26). First order conditions require

$$\partial L/\partial\beta^{*} = \hat{\sigma}^{*-2}\, X'\,\epsilon^{*} = 0 \tag{27}$$

$$\partial L/\partial\lambda = -\,\hat{\sigma}^{*-2}\,\epsilon^{*\prime}\epsilon^{*}_{\lambda} = 0 \tag{28}$$

and second order conditions require

$$\partial^2 L/\partial\theta^{*}\partial\theta^{*\prime} = -\,\hat{\sigma}^{*-2}\begin{bmatrix} X'X & -X'\epsilon^{*}_{\lambda} \\ -\epsilon^{*\prime}_{\lambda}X & [\epsilon^{*\prime}\epsilon^{*}_{\lambda\lambda} + \epsilon^{*\prime}_{\lambda}\epsilon^{*}_{\lambda}] \end{bmatrix} \tag{29}$$

to be negative definite. The negative of the inverse of (29) is the estimated covariance matrix of (β^{*},λ). Obtaining the covariance matrix of (β,λ) is a simple process which can be shown using an argument of Bard (1974, p. 205). Let the vector of coefficients from the original model be $\theta' = \{\beta'\lambda\}$ and the vector of coefficients from the scaled model be

$$\theta^{*\prime} = \{\beta^{*\prime}\,\lambda\} = g\,(\theta). \tag{30}$$

Equation (30) is reversible; i.e.,

$$\begin{aligned}\theta = g^{-1}\,(\theta^{*}) = h(\theta^{*}) = \{&\beta^{*}_{1}\dot{y}^{\lambda}\\ &+ \dot{y}^{(\lambda)}\,\beta^{*}_{2}\dot{y}^{\lambda} \ldots \beta^{*}_{k}\dot{y}^{\lambda}\;\lambda\}'.\end{aligned} \tag{31}$$

Expanding θ in (31) around θ^{*} in a Taylor's Series approximation gives

$$\theta \simeq h(\hat{\theta}^{*}) + h'(\hat{\theta}^{*})\,(\theta^{*} - \hat{\theta}^{*}) \tag{32}$$

where $\hat{\theta}^{*}$ is the vector of estimated coefficients in the scaled model, and $h(\hat{\theta}^{*}) = \hat{\theta}$ is the vector of estimated coefficients in the original model obtained from (31). Let $h'(\hat{\theta}^{*}) = J$, a $(K + 1) \times (K + 1)$ upper triangle matrix given by

$$J = \begin{bmatrix} \dot{y}^{\lambda}\;0\;0\;0\;\ldots & \beta_1 \ln \dot{y} + \lambda^{-1} & (\ln \dot{y} - \dot{y}^{(\lambda)}) \\ \dot{y}^{\lambda}\;0\;0\;\ldots & \beta_2 \ln \dot{y} & \\ \dot{y}^{\lambda}\;0\;\ldots & \vdots & \\ \dot{y}^{\lambda}\;\ldots & \beta_k \ln \dot{y} & \\ \vdots & & \\ & 1 & \end{bmatrix} \tag{33}$$

Then from (32), $(\theta - \hat{\theta}) \simeq J\,(\theta^{*} - \hat{\theta}^{*})$ and

$$\begin{aligned}(\theta - \hat{\theta})\,(\theta - \hat{\theta})' &\simeq J\,(\theta^{*} - \hat{\theta}^{*})\,(\theta^{*} - \hat{\theta}^{*})'\,J'\\ &= JV(\hat{\theta}^{*})J'.\end{aligned} \tag{34}$$

Equation (34) provides an estimate of the covariance matrix of the coefficient vector of the original model in terms of the covariance matrix estimated from the scaled model. $V(\hat{\theta}^{*})$ is estimated by the negative of the inverse of (29).

It should be further noted that this procedure for obtaining the correct covariance matrix estimates does *not* increase in complexity for models more general than (8). The β^{*} are functions only of the β and the power transformation on y and are not functions of any *other* power transformation; the matrix J will merely have an addi-

tional column and row for more complex models. The additional column (row) will have a one on the main diagonal and zeros elsewhere.

In sum, estimation of (8) using the scaled variables can be accomplished in four steps: (1) scale each y_i by $\dot{y}$, (2) find the parameter estimates, (θ^*), which maximize (26); (3) substitute the values of θ^* into (31) to obtain parameter estimates of θ for the original model, and (4) estimate the covariance matrix of the parameter estimates from the original model. This last step is carried out by premultiplying the inverse of the negative of (29) by J and then postmultiplying that result by J'. This approach simplifies the estimation problem by eliminating the Jacobian term from the function to be maximized.

D. *Iterated Ordinary Least Squares (IOLS)*

Since maximizing (26) is equivalent to minimizing $\hat{\sigma}^{*2}$, the maximum likelihood solution for the β^*, conditional on λ, is $\beta^* = (X'X)^{-1}X'y^{*(\lambda)}$. An ordinary least squares computer program may be simply modified or controlled to repeatedly estimate β^* for different λ. The error sum of squares, $T\hat{\sigma}^{*2}$, is computed in each case, and the value of λ for which this is minimized will be the same, given the same data, as for any of the preceding three approaches. A systematic grid search can be done, searching say from $\lambda = -2$ to $\lambda = +2$, in steps of 0.1. If additional accuracy is required, the search can be repeated in smaller steps. For either model (8) or (9) the grid search is the same, so both of these models can be easily estimated by iterated OLS. Models given by (10) or (11) can be estimated by IOLS but the grid search must be at least two dimensional; the number of OLS regressions which must be performed becomes very large.

Estimates of the covariances of the parameter estimates can be obtained as well. The matrix $X'X$ used in estimation must be augmented by a single column and row to form the matrix (29). Step (4) from section III C can then be implemented.

IOLS estimation does not provide correct ML covariance estimates directly. Clearly some additional work must be done to get these estimates. However, considering the amount of computer "work" required to obtain the parameter estimates, the additional computation required to obtain correct covariance estimates is trivial by comparison. Failure to obtain ML covariance matrix estimates makes any hypothesis testing on the coefficient vector incorrect, even asymptotically. The proof of this is straightforward.

Let the negative of (22) be partitioned into

$$\sigma^{-2}Z = \sigma^{-2}\begin{bmatrix} A & B \\ B' & C \end{bmatrix} \tag{35}$$

where $A = X'X$ is $(K \times K)$, B is a $(K \times 1)$ vector and C is (1×1). OLS incorrectly computes σ^2A^{-1} as the covariance matrix of β. The correctly estimated covariance matrix of β and λ is

$$\sigma^2Z^{-1} = \sigma^2\begin{bmatrix} U & V \\ V' & W \end{bmatrix} \tag{36}$$

and the correct covariance matrix of β is

$$\begin{aligned}\sigma^2U &= \sigma^2[A - BB'/C]^{-1} \\ &= \sigma^2[A^{-1} + (|A|/|Z|)A^{-1}BB'A^{-1}].\end{aligned} \tag{37}$$

Thus, the ML covariance matrix is bigger by

$$\sigma^2(|A|/|Z|)A^{-1}BB'A^{-1}, \tag{38}$$

a positive semidefinite matrix. OLS clearly underestimates the coefficient variance estimates.

Emphasis is being placed on the correct computation of the variances and covariances of the parameter estimates because several published papers have included standard errors on β obtained not from the information matrix, but directly from OLS estimates. It has been shown that these estimates are biased downward and their use for hypothesis testing is therefore inappropriate. While it is true that none of the authors of these papers performed hypothesis tests, the inclusion of biased statistics is itself a grossly misleading practice.

E. *Ascertaining the Size of OLS Covariance Bias*

Some general conclusions about the size of the OLS covariance bias may be obtained by manipulation of the partitioned matrices in section III D. First note that $(|A|/|Z|) = W$, a scalar, if the model being estimated is either (8) or (9). Thus $\sigma^2(|A|/|Z|) = \sigma^2W = \text{Var}(\hat{\lambda})$. Second, $-\sigma^2A^{-1}WB = -\text{Var}(\hat{\lambda})A^{-1}B$ which are the covariances of λ and the βs. Thus, the OLS biases in the variances of the linear parameters in (38) are equal to the diagonal elements of $\text{Cov}(\hat{\lambda},\hat{\beta})^2/\text{Var}(\hat{\lambda})$.

These observations lead to the following general conclusions about the bias in the OLS variances:

(a) Not surprisingly, the bias increases with Var ($\hat{\lambda}$).
(b) The bias increases the larger the covariance of λ with the βs.

These results are not very helpful in a practical sense, since the size of β depends on the scaling of the data *and the power transformation parameter*. Practical experience dictates that using OLS variance estimates is a poor policy. It is not unusual to find that OLS variance estimates underestimate the ML variance estimates by several thousand percent! That is, the bias is several hundred times larger than the OLS variance estimates themselves.

IV. Concluding Observations

1. There are several approaches to the estimation of the parameters in models which contain the Box-Cox transformation. All methods, given the same data, can be made to obtain the same parameter estimates and the same covariance matrix estimates.

2. The full maximum likelihood method requires the solution of a $(K + 2)$ dimensional problem. Since at each iteration a simultaneous equations system of $(K + 2)$ equations must be solved, the amount of work per iteration required for the full ML method will be larger than for any of the other methods. It is not clear, and one should not mistakenly conclude, that the full ML method is slower to converge than the other approaches. The speed of convergence will depend on the conditioning of the system of simultaneous equations. For the same problem, and the same initial guesses, full ML may converge faster (require less computer time) than any of the other methods. This occurs because the system of equations for matrix (17) may be better conditioned than that of (22) or (29). Thus while inverting the larger matrix (17) requires more work per iteration, it is quite possible that fewer iterations may be needed.

3. The same conclusion holds for NLLSQ versus the concentrated log-likelihood approach; i.e., the speed with which each converges will depend upon the data of a particular problem. Sometimes one will converge much faster than the other; other times, the reverse may hold.

4. IOLS is a viable approach to the problem, particularly for models (8) and (9) where only one transformation parameter is specified. There are two advantages to this approach: (1) existing OLS programs may be easily used, and (2) "convergence" is guaranteed, since all realistic values of λ are evaluated. Disadvantages of using IOLS include: (1) the manipulation at the conclusion of estimation to obtain the correct covariance matrix estimate, and (2) the amount of computer time required for estimation may easily exceed that required for the nonlinear estimation methods. With good initial guesses, the nonlinear approaches may converge in relatively few iterations. To obtain accurate, say 4 decimal place, estimates of λ by IOLS, over 100 "regressions" may be required.

5. For the nonlinear methods, second derivatives are important, and rapid convergence necessitates their accurate computation. Failure to use second derivatives will slow convergence and may give larger estimated variances.

6. It is probably a good idea to use double precision arithmetic on all Box-Cox estimation problems. The matrices (17), (22) and (29) are all generally very ill-conditioned. Spectral condition numbers in excess of 10^6 are not unusual.

7. Where one approach to the estimation problem fails after a large number of iterations, try another method.

REFERENCES

Bard, Yonathan, *Nonlinear Parameter Estimation* (New York: Academic Press, 1974).

Box, G. E. P., and D. R. Cox, "An Analysis of Transformations," *Journal of the Royal Statistical Society*, Ser. B, 26 (Apr. 1964), 211–243.

Goldfeld, Stephen M., and Richard E. Quandt, *Nonlinear Methods in Econometrics* (Amsterdam: North-Holland Publishing Company, 1972).

Huang, Cliff J., and John A. Kellogos, "Conditional Mean Function and a General Specification of the Disturbance in Regression Analysis," *Southern Economic Journal* 45 (Jan. 1979), 710–717.

Mallela, Parthasaradhi, "Discrimination between Linear and Logarithmic Forms—A Note," this REVIEW 62 (Feb. 1980), 142–144.

Poirier, Dale J., "The Use of the Box-Cox Transformation in Limited Dependent Variable Models," *Journal of the American Statistical Association* 73 (June 1978), 284–287.

Poirier, Dale J., and Angelo Melino, "A Note on the Interpretation of Regression Coefficients within a Class of Truncated Distributions," *Econometrica* 46 (Sept. 1978), 1207–1209.

Schlesselman, J., "Power Families: A Note on the Box and Cox Transformation," *Journal of the Royal Statistical Society* Ser. B, 33 (2)(1971) 307–311.

Zarembka, Paul, "Functional Form in the Demand for Money," *Journal of the American Statistical Association* 63 (June 1968), 502–511.

APPENDIX

Discrimination between Linear and Logarithmic Forms

In a recent note in this REVIEW, Mallela (1980) proposed that the Box-Cox transformation should not be used to distinguish between different functional forms. Mallela argued that Zarembka's (1968) generalization of the Box-Cox transformation cannot be used to test H: $\lambda = 0$. We question the validity of Mallela's argument.

Mallela criticized Zarembka's derivation of the money demand function as follows. The generalized functional form is given by

$$Y^{\lambda} = \beta_0 + \beta_1 X_1^{\lambda} + \beta_2 X_2^{\lambda}. \tag{A.1}$$

If (a) unity is subtracted from both sides of (A.1), (b) β_1 and β_2 are added and subtracted from the right-hand side, and (c) both sides of (A.1) are divided by λ, one obtains

$$(Y^{\lambda} - 1)/\lambda = (\beta_0 + \beta_1 + \beta_2 - 1)/\lambda + \beta_1(X_1^{\lambda} - 1)/\lambda + \beta_2(X_2^{\lambda} - 1)/\lambda. \tag{A.2}$$

Furthermore,

$$\lim_{\lambda \to 0} (\beta_0 + \beta_1 + \beta_2 - 1)/\lambda = \pm\infty \tag{A.3}$$

unless

$$\beta_0 + \beta_1 + \beta_2 = 1. \tag{A.4}$$

Lastly, Mallela argued that Zarembka's transformation of (A.1) into

$$\ln Y = \beta^*_0 + \beta_1 \ln X_1 + \beta_2 \ln X_2 \tag{A.5}$$

(a) is of limited use since (A.4) cannot in general be assumed to hold, and (b) is not estimable because if (A.4) does hold, (A.2) is no longer scale invariant (see Schlesselman (1971)).

Mallela's critique is in error. *With complete generality*, let

$$\beta_0 = \lambda\alpha - \beta_1 - \beta_2 + 1. \tag{A.6}$$

Equation (A.3) is now finite at $\lambda = 0$ since, substituting (A.6) into (A.4), the limit in (A.4) is now $\alpha \neq \infty$. Zarembka's derivation is thus both legitimate and estimable. Hypothesis tests for H: $\lambda = 0$ are asymptotically valid so long as the error terms are approximately normally distributed.