A Note on the Box-Cox Transformation Under Heteroskedasticity

James Blaylock, Larry Salathe and Richard Green

The Box-Cox transformation (BCT) has been frequently used as both a flexible functional form and as a decision device to distinguish among alternative model specifications. Most researchers have failed to recognize that the BCT when applied to the dependent variable can compensate for heteroskedasticity. This paper investigates a new procedure which estimates both the BCT parameters and the analytic form of heteroskedasticity. Results from the new procedure are compared to estimates obtained from the traditional method of estimating BCT models. Comparisons indicate that proper specification of the error variance can influence the magnitude of BCT parameters and alter the results of hypothesis testing.

The monotonic transformation introduced by Box and Cox has been employed in a number of econometric applications, including those in monetary economics [Zarembka, 1968], production theory [Appelbaum], and demand analysis [Chang, Kulshreshtha], for added flexibility in model specification. For example, application of the Box-Cox transformation (BCT) to both the dependent and independent variables of a regression model defines a general class of functional forms which includes the linear and double-logarithmic functions as special cases. This feature of the BCT model allows the data added flexibility in determining the degree of nonlinearity in a relationship and provides a unified structure for statistically distinguishing among alternative functional specifications. However, Box and Cox have also pointed out that transformation of the dependent variable is a convenient device to render the density of an equation error term more normal-like. Consequently, application of the BCT to the dependent variable of an equation affects the distributional properties of the residual errors as well as the functional form of the relationship.

Most empirical applications have ignored the influence of the BCT on the structure of the error term. In particular, the relationship between the BCT and heteroskedasticity of the error variance is neglected. This is despite Zarembka's [1974] contention that the estimation process will bias the BCT parameter on the dependent variable to compensate for any heteroskedasticity. Thus, for any particular application, it is important to ascertain the robustness of the BCT parameters to heteroskedasticity to achieve unbiased parameter estimates and for proper statistical testing.

In light of the foregoing, the objective of this paper is twofold. The first is to investigate the impact of heteroskedasticity on the estimated values of the BCT parameters and on the conclusions of subsequent hypothesis tests. The second objective is to analyze the

James Blaylock and Larry Salathe are agricultural economists with the National Economics Division, ESCS, and Richard Green is assistant professor of agricultural economics at the University of California, Davis.

The authors thank David Smallwood and the referees for many constructive comments. The views expressed in this paper are not necessarily those of the United States Department of Agriculture.

analytical form of heteroskedasticity as it pertains to BCT models.

To examine the question of potential bias, a method proposed by Gaudry and Dagenais is employed. This technique simultaneously estimates the BCT parameters and the analytical form of heteroskedasticity. In addition, this procedure allows separate statistical testing to be performed on the nonstochastic and stochastic parts of the model. Zarembka's procedure [1974] for obtaining "approximately consistent" estimates of the BCT parameters under heteroskedasticity is also examined. These methods are compared to the traditionally estimated BCT model which postulates a homoskedastic error variance.

The above methods are employed to estimate models, using cross-section data, which express per capita expenditure on a commodity as a function of per capita income. The typical Engel relationship was chosen for its simplicity and widespread use. The various methods are compared and analyzed to delineate the role proper heteroskedastic specification of the error variance plays in the estimation of BCT parameters and in hypothesis testing.

Basic Models

The Box-Cox transformation for any positive non-Boolean variable W is defined as

$$(1)\qquad W^{(\lambda_w)} = (W^{\lambda_w} - 1)/\lambda_w,\ \lambda \neq 0$$

$$= \ln(W) \qquad \lambda = 0$$

with the corresponding inverse function

$$(2)\ W = [\,\lambda_w W^{(\lambda_w)} + 1\,]^{1/\lambda_w},\ \lambda_w \neq 0$$

$$= \exp(W^{(\lambda_w)}) \qquad ,\ \lambda_w = 0$$

where λ_w is a parameter to be estimated.

The Box-Cox model written in the usual notation for each observation is

$$(3)\ Y_i^{(\lambda_y)} = \sum_{k=1}^{K} \beta_k X_{ki}^{(\lambda_{xk})} + u_i,\ i = 1,\ldots,Q$$

where

$$(4)\ u_i \sim N(0,\sigma^2),\ \text{for the traditional model}$$

and,

$$(5)\qquad u_i = [f(Z_{li},\ldots,Z_{Mi})]^{1/2} v_i$$

for the heteroskedastic case. The variables Z_M explain heteroskedasticity and may be different from the explanatory variables X_k. In addition, using matrix notation and assuming Z and v to be independent yields

$$(6)\qquad E(v) = 0 \text{ and } E(vv') = \Psi^2 I$$

where 0 represents the null vector, I is a (Q by Q) identity matrix, E denotes the expectation operator, and Ψ is a constant. Also, in matrix notation

$$(7)\qquad E(u) = 0 \text{ and } E(uu') = \Omega$$

where u is the (Q by 1) vector of u_i elements and Ω is a (Q by Q) matrix. A representative diagonal element of Ω will be denoted by

$$(8)\qquad \omega_{ii} = \Psi^2 f(Z_{li},\ldots,Z_{Mi}).$$

All off-diagonal elements of Ω are zero.

Following Gaudry and Dagenais, it is assumed that in equation (8) any given diagonal element of Ω can be expressed as

$$(9)\ \omega_{ii} = \Psi^2\{\lambda_\nu[\delta_0 + \sum_{m=1}^{M} \delta_m Z_{mi}^{(\lambda_{am})}] + 1\}^{1/\lambda_\nu}$$

in which a Box-Cox transformation (λ_{am}) has been applied to the Z_m variables and an inverse transformation (λ_ν) is applied to the expression in squared parenthesis. The constant δ_0 is necessary to preserve the invariance of the transformation parameters to the units of measurement of Z [Schlesselmann].

Several interesting specifications of the analytic form of the heteroskedasticity are possible by first setting λ_ν equal to zero. Other values of λ_ν are permissible but they can produce negative elements in Ω.

By setting $\lambda_\nu = 0$ in equation (9) and using the definition of the inverse transformation given in equation (2) yields

$$(10) \quad \omega_{ii} = \Psi^2 \exp(\delta_0 + \sum_{m=1}^{M} \delta_m Z_{mi}^{(\lambda_{am})})$$

$$\equiv \sigma^2[\exp(\sum_{m=1}^{M} \delta_m Z_{mi}^{(\lambda_{am})})]$$

where $\sigma^2 = \Psi^2 \exp(\delta_0)$. Equation (10) reduces to the multivariate model

$$(11) \qquad \omega_{ii} = \sigma^2 \prod_{m=1}^{M} Z_{mi}^{\delta_m}$$

when each $\lambda_{am} = 0$. It reduces to the classical multivariable form

$$(12) \qquad \omega_{ii} = \sigma^2 \prod_{m=1}^{M} Z_{mi}^{2}$$

when each δ_m is set equal to 2 and each λ_{am} is set equal to zero. The homoskedastic case is obtained from equation (10) by setting each δ_m equal to zero. Furthermore, Park's specification is derived by setting all δ_m but one equal to zero in equation (10) and by setting the λ_{am} of the remaining variable equal to zero. Hence, a number of traditional specifications are obtainable depending on the values of λ_{am} and δ_m.

The likelihood function corresponding to the traditional Box-Cox model (i.e. expressions (3) and (4)) is

$$(13) \qquad L = \prod_{i=1}^{Q} [1/(2\pi\sigma^2)^{1/2}]^Q$$

$$\exp\{-(1/2\sigma^2)(Y_i^{(\lambda_y)} - \sum_{k=1}^{K} \beta_k X_{ki}^{(\lambda_{xk})})^2\}|J|$$

where $|J|$ denotes the Jacobian of the transformation from $Y_i^{(\lambda_y)}$ to the observed Y_i

$$(14) \ |J| = |\det(\partial Y_i^{(\lambda_y)}/\partial Y_i)| = \prod_{i=1}^{Q} Y_i^{\lambda_y - 1}$$

Likewise, the likelihood function corresponding to the heteroskedastic case (i.e. expressions (3), (7), and (10)) is

$$(15) \qquad L = \prod_{i=1}^{Q} [Y_i^{\lambda_y - 1}/(2\pi\sigma^2)^{1/2}$$

$$\exp\{\sum_m \delta_m Z_{mi}^{(\lambda_{am})}\}]\cdot$$

$$\exp[-(Y_i^{(\lambda_y)} - \sum_k \beta_k X_{ki}^{(\lambda_{xk})})^2/2\sigma^2$$

$$\exp\{\sum_m \delta_m Z_{mi}^{(\lambda_{am})}\}]$$

from which estimates of δ_m, λ_{am}, β_k, λ_{xk}, λ_y, and σ^2 can be obtained.

Zarembka's [1974] procedure for obtaining "approximately consistent" estimates of the transformation parameters under heteroskedasticity involves finding those values of the parameters (λ_y and λ_{xk}) such that the following equation holds

$$(16) \quad \frac{\partial LL}{\partial \lambda_y} = \frac{(1-\lambda_y-h)(Q)[\mathrm{var}(\ln Y)]}{[1/2+(1-\lambda_y-h)^2\ \mathrm{var}(\ln Y)]}$$

where $\partial LL/\partial\lambda_y$ is the first derivative of the log of the likelihood function (equation 13) with respect to λ_y and h is chosen *a priori* to reflect the nature of the heteroskedasticity.

The crucial role that error specification brings to bear on the estimates of the parameters is indicated by the following. If h is set equal to one, relation (16) implies that the transformation parameters will be estimated under the assumption the variance of Y_i increases with the square of its expected value. If the values of λ_y and λ_{xk} obtained by maximizing equation (13) are significantly different from those values derived from relation (16), given h is equal to one, the estimates from (13) are not reflecting nonlinearities but rather λ_y is being biased in the estimation process to compensate for heteroskedasticity. On the other hand, if the values of λ_y are essentially equal for the two procedures then either heteroskedasticity is unimportant and λ_y is reflecting nonlinearities or the estimated value of λ_y while now reflecting nonlinearities is also the value of λ_y which stabilizes the error variance.

Three points should be clarified at this time. First, Zarembka's procedure requires *a priori* specification of the analytic form of the heteroskedasticity and, therefore, is a restrictive method. Secondly, a consistent estimate of λ_y is a sufficient condition for consistent estimation of the other parameters and thirdly, since Zarembka's method provides only "approximately consistent" estimates the use of likelihood ratio tests is not valid.

Empirical Results

The above methods, represented by equations (13), (15), and (16), are applied to a model which postulates that per capita expenditure on a commodity is a function of per capita income. Separate analyses are performed for poultry, pork, and eggs. The data base used is a random sample of 100 households drawn from the urban section of the Northeast region contained in the USDA Household Food Consumption Survey of 1965. The sample was uniformly selected to be mean and standard deviation preserving.

The traditional BCT model (equation 13) will assume there is one explanatory variable (income), an error term with a constant variance, and that the same transformation parameter is applied to both the dependent and independent variables, i.e.,

$$(17) \qquad Y_i^{(\lambda_y)} = \sum_k \beta_k X_{ki}^{(\lambda_{xk})} + u_i$$

$$= \alpha + \beta X_{li}^{(\lambda_y)} + u_i, \quad u_i \sim (0,\sigma^2)$$

where α is a constant.

The heteroskedastic model assumes:

$$(18) \qquad Y_i^{(\lambda_y)} = \sum_k \beta_k X_{ki}^{(\lambda_{xk})} + u_i$$

$$= \alpha + \beta X_{li}^{(\lambda_y)} + u_i, \quad u_i \sim N(0,\Omega)$$

where a typical diagonal element of Ω is given by

$$\omega_{ii} = \sigma^2[\exp(\sum_m \delta_m Z_{mi}^{(\lambda_{am})})] = \sigma^2[\exp(\delta X_{li}^{(\lambda_a)})]$$

and all off-diagonal elements of Ω are zero. In other words, the same BCT parameter is applied to both dependent and independent variables and the variable which explains heteroskedasticity (Z_m) will be income (X).

The Zarembka procedure is predicated on the assumption that h is equal to one in equation (16).

The regression equations given in models (17) and (18) are Box-Cox representations of typical Engel curves. Other variables, e.g. race, could be included in the model but for the purpose of this paper the specifications are kept as simple as possible. The dependent and independent variables are subject to the same transformation parameter to simplify estimation and permit the use of simple t-ratios. The assumption that λ_v is equal to zero in the heteroskedastic model is not very restrictive as the analytic form of the heteroskedasticity is still quite general. While, the above assumptions do not affect the objectives of this analysis, the estimated elasticities should be interpreted with these assumptions in mind.

The Fletcher-Powell algorithm is used for the maximization of the log-likelihood functions, concentrated on σ^2 and β, corresponding to equations (13) and (15). The Newton-Raphson routine [Henrici] was used to find the value of λ_y for which equation (16) obtains. Consistent estimates of the asymptotic standard errors of the parameters were obtained from the negative of the inverse matrix of second partial derivatives of the appropriate likelihood function. Asymptotic tests can then be performed via the asymptotic normality of the parameters or by conventional likelihood ratio methods.

Estimation of the heteroskedastic BCT model is accomplished in the following manner. First, a starting value of λ_y for use in estimating the homoskedastic BCT model is found by employing a search routine over the parameter interval (7, -7). The maximum likelihood value of λ_y, estimated from the homoskedastic model, was used as a starting value for the heteroskedastic model with δ started at zero (to begin with the homoskedastic specification). Convergence of the likelihood functions was accomplished within 15-20 iterations. Other starting values for λ_y and δ were tried with the algorithm always converging to the same parameter estimates. Both expenditures and income were divided by 1,000 to eliminate the possibility of numerical problems in the estimation process.

The parameter estimates for the traditional homoskedastic model (equation 17), Zarembka's procedure (equation 16), and the heteroskedastic model (equation 18) are presented in Table 1.

The asymptotic standard error for the transformation parameter (λ_y) in the eggs model, as estimated from the traditional specification, indicates that λ_y is not different from zero at the 0.05 level of significance. Zarembka [1974] indicates that the BCT parameter on the dependent is biased toward zero if the error variance increases with the expected value of the dependent variable. Thus in the egg model an indication is given that heteroskedasticity could be exercising influence in the estimation of λ_y.

The standard approach has been to ignore this potential problem and instead indicate that a double-logarithmic function appears to be the best fitting of the "classical" functional forms. That the double-log function fits the data best under heteroskedasticity is not surprising as it is well-known that logarithmically transforming the dependent variable of a model will compensate for heteroskedasticity when the error variance is proportional to the square of the expected value of the dependent variable.

The transformation parameters associated with the poultry and pork models appear to be significantly different from both zero and one. In summary, using the traditional method for estimating the BCT model would lead to the acceptance of the double-log functional form for eggs. Both the linear and double-log specifications are rejected for poultry and pork.

Zarembka's method, conditioned on the assumption that h is equal to one in equation (16), shows a major difference in the magnitude of the transformation parameters for the poultry and egg models vis-a-vis the traditional estimates. Although t-tests indicate that λ_y is significantly different from zero for poultry in the traditional model the parameter does not appear to be robust to heteroskedasticity, i.e. the unstable error variance is adversely influencing the estimation of λ_y. Heteroskedasticity has also influenced the estimated value of the BCT parameter in the egg model. On the other hand, the transformation parameter associated with the pork equation does appear to be robust under the nature of the heteroskedasticity as outlined above. The BCT parameters associated with the three commodities are significantly different from both zero and one. Hence, the linear and double-log functions are rejected for the three commodities. This is in contrast to the homoskedastic BCT model which accepted the double-log form for eggs.

The generalized method, where the analytical form of the heteroskedasticity and the BCT parameters are estimated simultaneously, produces results that are substantially different from those obtained using the traditional and Zarembka methods, except for pork. For example in the case of eggs, λ_y ranged from -0.0241 using the traditional estimation process to -0.1148 with the Zarembka procedure to -0.2443 using the heteroskedastic specification. The asymptotic

TABLE 1. Estimated Parameter Values

Method	Poultry						
	α	β	λ_y	δ	λ_a	ε[1]	LL[2]
Traditional	0.6785	0.3207 (0.0946)[3]	0.0043 (0.0020)	0.0	-	0.3214	216.83
Zarembka's	0.6661	0.3087 (0.0872)	0.1407 (0.0530)	-	-	0.3318	209.41[4]
Heteroskedastic	0.6541	0.3097 (0.1001)	0.1987 (0.0826)	0.1374 (0.0584)	0.9656 (0.2963)	0.3429	229.88
Method	**Pork**						
	α	β	λ_y	δ	λ_a	ε	LL
Traditional	0.6948	0.2375 (0.0595)	0.1818 (0.0353)	0.0	-	0.2735	294.57
Zarembka's	0.6908	0.2327 (0.0628)	0.1919 (0.0444)	-	-	0.2701	293.24
Heteroskedastic	0.6775	0.2446 (0.0695)	0.1924 (0.0656)	0.0027 (0.0131)	0.0026 (0.0010)	0.2841	295.08
Method	**Eggs**						
	α	β	λ_y	δ	λ_a	ε	LL
Traditional	0.8524	0.1816 (0.0454)	−0.0241 (0.0290)	0.0	-	0.1739	324.65
Zarembka's	0.9512	0.2224 (0.0702)	−0.1148 (0.0415)	-	-	0.1810	318.79
Heteroskedastic	1.1085	0.3004 (0.0538)	−0.2443 (0.0975)	0.4378 (0.2182)	0.0942 (0.3603)	0.1938	341.40

[1] Refers to the expenditure elasticity evaluated at the sample means.

[2] Value of the Log-likelihood function.

[3] Estimated standard error.

[4] Value of the log-likelihood function declines using the Zarembka method because of the restricted assumption concerning the form of the heteroskedasticity.

standard errors associated with the BCT parameters for all three commodities indicate rejection of the linear and double-log hypotheses. This is in agreement with the results obtained from the Zarembka method but different from the result for eggs in the traditional BCT model.

The analytic form of the heteroskedasticity conformed to Park's specification in the egg model as the value of λ_a is not significantly different from zero and the value of δ is significantly different from zero. The form of the heteroskedasticity did not conform to any of the traditional specifications in the poultry model. The pork model is homoskedastic as δ is not significantly different from zero. The importance of proper heteroskedastic specification is illustrated by the change in the BCT parameters for the poultry and egg models going from Zarembka's restrictive assumption on the error variance to the general method.

The values of the log-likelihood functions for the heteroskedastic egg and poultry models are a significant improvement over the values of the log-likelihoods obtained from the traditional method. Likelihood ratio tests can be used to compare the heteroskedastic and traditional BCT models as the traditional model is a special case of the general model when δ is restricted to zero. The values of the log-likelihood functions for pork were virtually identical. This is expected as the pork model was shown to be homoskedastic. Zarembka's procedure actually produced a decline in the values of the log-likelihood functions vis-a-vis the traditional method. This is also expected as the Zarembka method uses a very restricted analytic form of heteroskedasticity.

The elasticities estimated from the various procedures are almost identical. Hence, for policy considerations the traditional model may suffice due to the small change in elasticities when evaluated at the sample means. But, for structural issues, hypothesis testing, and for elasticities evaluated at points other than the means, one must correct for heteroskedasticity to achieve reliable results.

Conclusions

This note has attempted to demonstrate that the proper specification of both the nature of the error term and the analytical form of heteroskedasticity is of critical importance for the correct estimation and, hence, interpretation of results generated from BCT models. Whether the data base is of a cross-sectional or time series nature the problem of heteroskedasticity must be broached.

Empirical analyses which utilize the transformation of variables technique should be regarded with skepticism unless homoskedasticity of the error variance is established through procedures as outlined in this paper.

In conclusion, it appears that while the BCT is a powerful device for selecting among alternative functional forms and as a technique to introduce flexibility into a model specification, its random application without regard to heteroskedasticity renders ineffectual conclusions.

References

Appelbaum, E. "On the Choice of Functional Forms." *International Economic Review*, 2(1979): 449-458.

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

Chang, H. S. "Functional Forms and the Demand for Meat in the United States." *Review of Economics and Statistics*, 59(1977): 355-359.

Gaudry, M. and M. Dagenais. "Heteroscedasticity and the Use of Box-Cox Transformations." *Economics Letters*, 2(1979): 225-229.

Henrici, P. *Elements of Numerical Analysis*. John Wiley & Sons, New York, 1967.

Kulshreshtha, S. "Functional Form Specification in the Quarterly Demand for Red Meats in Canada." *Western Journal of Agricultural Economics*, 4(December, 1979): 89-97.

Park, R. E. "Estimation with Heteroskedastic Error Terms." *Econometrica*, 4(1966): 888.

Schlesselmann, J. "Power Families: A Note on the Box-Cox Transformation." *Journal of the Royal Statistical Society*, 33(1971): 307-311.

Zarembka, P. "Functional Form in the Demand for Money." *Journal of the American Statistical Association*, 63(1968): 502-511.

Zarembka, P. "Transformation of Variables in Econometrics." in P. Zarembka (Ed.), *Frontiers of Econometrics*, Academic Press, New York, 1974: 81-104.