Title
Global Identification of the Semiparametric Box-Cox Model

Author
Komunjer, Ivana

Publication Date
2008-04-01

GLOBAL IDENTIFICATION OF THE SEMIPARAMETRIC BOX-COX MODEL

IVANA KOMUNJER

ABSTRACT. This paper establishes the identifiability of the parameters of the Box-Cox model under restrictions that do not require the disturbance in the model to be independent of the explanatory variables. The proposed restrictions are semiparametric in nature: they restrict the support of the conditional distribution of the disturbance but do not require the latter to be known.

JEL Codes: C13, C14

Keywords: identification; Box-Cox regression; structure

Affiliation and Contact Information: Department of Economics, University of California San Diego, 9500 Gilman Drive, La Jolla CA 92093-0508, United States. Phone: (+1 858) 822-0667. Fax: (+1 858) 534-0172. E-mail: komunjer@ucsd.edu.

1. Introduction

We consider a transformed regression model in which the dependent variable is transformed using the Box-Cox transformation (Box and Cox, 1964). Letting $Y \in (0, +\infty)$ denote an observed dependent variable, $X \in \mathbb{R}^k$ an observed explanatory variable, and $U \in \mathbb{R}$ a latent disturbance, the Box-Cox model is given by:

$$\frac{Y^\lambda - 1}{\lambda} = \beta' X + U, \quad (\beta', \lambda)' \in \mathbb{R}^{k+1}. \tag{1}$$

Assuming the model to be correctly specified, the object of interest is the true value $\theta_0 \equiv (\beta_0', \lambda_0)'$ of the parameter $\theta \equiv (\beta', \lambda)' \in \mathbb{R}^{k+1}$. To be defined for all values of λ, the dependent variable needs to be positive. Note that when $\lambda \neq 1$ the transformed model is nonlinear in variables. When $\lambda = 0$ the transformation reduces to the familiar log linear regression model since for any $y \in (0, +\infty)$ we have $\lim_{\lambda \to 0} \frac{y^\lambda - 1}{\lambda} = \ln y$.

The goal of this paper is to derive sufficient conditions under which any true value θ_0 of the parameter θ in the Box-Cox model is identifiable. In particular, we shall focus on the identification conditions that do not require the disturbance U to be independent of the explanatory variable X. The proposed conditions are semiparametric in nature, meaning that they do not require the probability distribution of U (and X) to be known.

This paper has several important antecedents. Foster, Tian, and Wei (2001) show the identifiability of θ_0 in cases in which the disturbance U is known to be independent of the explanatory variable X. Apart from independence, the remaining requirements of Foster, Tian, and Wei's (2001) result are fairly weak: $\beta_0' X$ is assumed to take at least two possible values (which excludes the case $\beta_0 = 0$), and the unknown distribution of U is assumed to have mean 0. However, in applications in which the endogeneity of U is suspected, the independence assumptions between U and X typically fail. They are often replaced by a weaker requirement that the disturbance U is mean independent of the explanatory variable X. This approach is taken, for

example, in Amemiya and Powell (1981) and Powell (1996) who consider moment restrictions on the joint distribution of U and X.

In this paper, we consider the restrictions on the support of the conditional distribution of U given X, and show that they are sufficient to identify the true value λ_0 of the exponent parameter in the Box-Cox model. The identification of β_0 is then straightforward. Like in Foster, Tian, and Wei (2001) our identification results are global. Our restrictions on U and X are, however, weaker. Like in Amemiya and Powell (1981) and Powell (1996) we do not require U to be independent of X.

Due to the nature of the Box-Cox transformation, the identifiability of θ_0 is non trivial to establish. Khazzoom (1989), Powell (1996) and Savin and Würtz (2005), for example, discuss this point. Indeed, the Box-Cox regression in Equation (1) is highly nonlinear in the parameter θ, and no transformation reduces it to a linear-in-parameters model. Thus, the well-known rank conditions for identification established in Koopmans (1950) and Fisher (1961, 1965) do not apply here. The identification results for nonlinear models (see, e.g., Fisher, 1966; Rothenberg, 1971; Komunjer, 2008) typically require θ_0 to satisfy a number of unconditional moment restrictions. Hence, they do not directly apply to our setup in which the restrictions on the support of the conditional distribution of U given X are used to identify the true value λ_0 of the exponent in the Box-Cox model. To the best of our knowledge, such identification conditions have not yet been studied by the literature.

The remainder of the paper is organized as follows: Section 2 discusses the structure of the Box-Cox transformed regression model and studies its properties. Section 3 contains our main result—the set of conditions that are sufficient to identify θ_0.

2. Structure of the Box-Cox Transformation Model

We start our analysis with a discussion of the structure relevant in the context of the Box-Cox regression in Equation (1). Let F_{XU} be the joint probability distribution

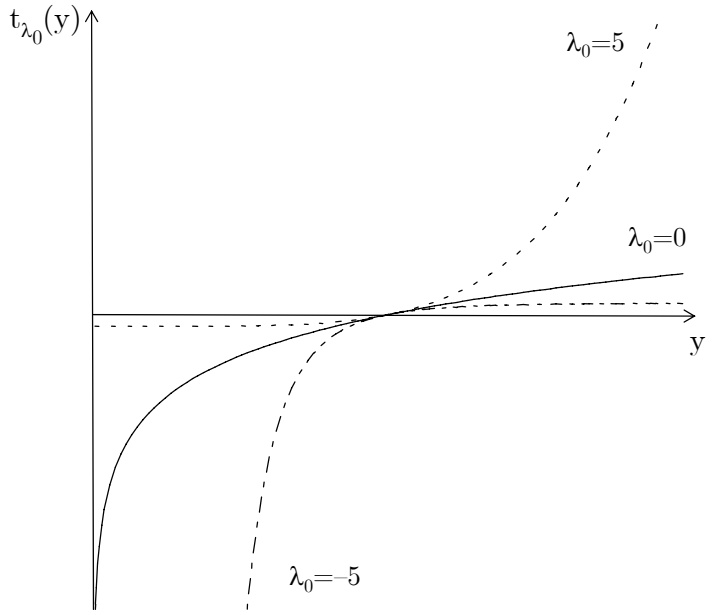

FIGURE 1. Plot of $y \mapsto t_{\lambda_0}(y)$ with $\lambda_0 = 5$ (dashed line), $\lambda_0 = 0$ (solid line) and $\lambda_0 = -5$ (dot-dashed line).

(measure) of X and U defined on $\mathbb{R}^{k+1}$. The support of F_{XU} is denoted D_{XU}.[1] The realizations of the random variables X and U are denoted x and u, respectively, with $(x', u)' \in D_{XU}$.

Fix $\theta_0 = (\beta_0', \lambda_0)' \in \mathbb{R}^{k+1}$ and $(x', u)' \in D_{XU}$, and consider the transformation equation: $t_{\lambda_0}(y) = \beta_0' x + u$, in which we have defined $t_{\lambda_0}(y) \equiv (y^{\lambda_0} - 1)/\lambda_0$ for all $y \in (0, +\infty)$. Note that the mapping t_{λ_0} is continuous and strictly increasing on $(0, +\infty)$. Hence, it is a homeomorphism from $(0, +\infty)$ onto its image $t_{\lambda_0}((0, +\infty))$, with $t_{\lambda_0}((0, +\infty)) = (-\infty, -1/\lambda_0)$ if $\lambda_0 < 0$, $t_0((0, +\infty)) = \mathbb{R}$, and $t_{\lambda_0}((0, +\infty)) = (-1/\lambda_0, +\infty)$ if $\lambda_0 > 0$.[2] The inverse of the map t_{λ_0} is defined for any $z \in t_{\lambda_0}((0, +\infty))$ and equals: $t_{\lambda_0}^{-1}(z) = (1 + \lambda_0 z)^{1/\lambda_0}$, which reduces to $t_0^{-1}(z) = \exp z$ when $\lambda_0 = 0$.

[1]Recall that the support of F_{XU} is defined to be the set of all points $(x', u)'$ in $\mathbb{R}^{k+1}$ for which every open neighborhood of $(x', u)'$ has positive measure. By construction, D_{XU} is closed in $\mathbb{R}^{k+1}$.

[2]A mapping is a homeomorphism if it is continuous, one-to-one, onto, and has a continuous inverse.

In order for the structure $\mathcal{S} = (\theta_0, F_{XU})$ to generate the joint probability distribution (measure) F_{XY} on $\mathbb{R}^k \times (0, +\infty)$ of the observables X and Y it is necessary that the transformation equation $t_{\lambda_0}(y) = \beta_0' x + u$ can always be solved for $y \in (0, +\infty)$ in terms of x, u and θ_0. This gives the following useful property of the Box-Cox model in Equation (1):

Lemma 1. *If the structure $\mathcal{S} = (\theta_0, F_{XU})$ generates F_{XY} then it must hold that: $D_Z \equiv \{z \in \mathbb{R} : z = \beta_0' x + u, (x', u)' \in D_{XU}\} \subseteq t_{\lambda_0}((0, +\infty))$.*

Lemma 1 is a simple implication of the correct specification of the Box-Cox regression model. Indeed, saying that the latter is correctly specified with true parameter value θ_0 is equivalent to saying that the probability distribution F_{XY} is generated from Equation (1) by the structure $\mathcal{S} = (\theta_0, F_{XU})$. This property can only hold if for every $(x', u)' \in D_{XU}$ the structural equation $t_{\lambda_0}(y) = \beta_0' x + u$ has a unique solution in y. The transformation T_{λ_0} which to each $(x', u)'$ associates $(x', y)' \equiv (x', t_{\lambda_0}^{-1}(\beta_0' x + u))'$ is then a single-valued mapping (or function) that is continuous on D_{XU}, which leads to the usual definition of the image measure F_{XY} of the observables X and Y, $F_{XY} = F_{XU} \circ T_{\lambda_0}^{-1}$. In particular, the support of F_{XY} is given by:

$$D_{XY} \equiv \{(x', y)' \in \mathbb{R}^k \times (0, +\infty) : (x', y)' = (x', t_{\lambda_0}^{-1}(\beta_0' x + u))', (x', u)' \in D_{XU}\} \tag{2}$$

and map T_{λ_0} is a homeomorphism from D_{XU} onto D_{XY}.[3] Hence, the two probability distributions (measures) F_{XU} and F_{XY} are isomorphic.

In what follows we shall focus on the sets $D_{U|X=0} \equiv \{u \in \mathbb{R} : (0', u)' \in D_{XU}, 0 \in \mathbb{R}^k\}$ and $D_{Y|X=0} \equiv \{y \in \mathbb{R} : (0', y)' \in D_{XY}, 0 \in \mathbb{R}^k\}$. If $0 \in \mathbb{R}^k$ is a possible realization of X, then $D_{U|X=0} \neq \emptyset$ and $D_{Y|X=0} \neq \emptyset$ are the supports of the conditional probability distributions given $X = 0$ of U and Y, respectively. In particular, we have $D_{U|X=0} \subseteq D_Z$ with D_Z as defined in Lemma 1 and so:

$$D_{U|X=0} \subseteq t_{\lambda_0}((0, +\infty)) \quad \text{and} \quad D_{Y|X=0} = t_{\lambda_0}^{-1}(D_{U|X=0}) \tag{3}$$

[3]Similar to previously, the support of F_{XY} is defined to be the set of all points $(x', y)'$ in $\mathbb{R}^k \times (0, +\infty)$ for which every open neighborhood of $(x', y)'$ has positive measure. By construction, D_{XY} is closed in $\mathbb{R}^k \times (0, +\infty)$.

A number of interesting properties result from Equation (3). For example, consider the case in which conditional on X being equal to $0 \in \mathbb{R}^k$, U is a continuous random variable with full support, i.e. $D_{U|X=0} = \mathbb{R}$. Now, under the first property in Equation (3), $D_{U|X=0} = \mathbb{R}$ can hold if and only if $\lambda_0 = 0$. The resulting support for the distribution of Y conditional on $X = 0$ then equals $D_{Y|X=0} = t_0^{-1}(\mathbb{R}) = (0, +\infty)$. Moreover, having $D_{Y|X=0} = (0, +\infty)$ is only possible if $\lambda_0 = 0$. So having a probability distribution F_{XY} of the observables such that the support of $F_{Y|X=0}$ equals $(0, +\infty)$ is sufficient to identify $\lambda_0 = 0$. The question which we now explore is: what information on $D_{U|X=0}$ is sufficient to identify any true value $\lambda_0 \in \mathbb{R}$?

3. Identification Condition

Following Koopmans and Reiersøl (1950) and Roehrig (1988), the true value θ_0 of the parameter θ in the Box-Cox model (1) is said to be identifiable if every structure $\mathcal{S}^* = (\theta_0^*, F_{XU}^*)$ whose characteristics are known to apply to $\mathcal{S} = (\theta_0, F_{XU})$ and which is observationally equivalent to $\mathcal{S}$—i.e. generates the same distribution of the observables F_{XY} as $\mathcal{S}$—satisfies $\theta_0^* = \theta_0$.

We are now ready to state our main result—Theorem 1—which provides sufficient conditions for θ_0 to be identifiable:

Theorem 1. *Let $\mathcal{S} = (\theta_0, F_{XU})$ with $\theta_0 \in \mathbb{R}^{k+1}$ be a structure that generates F_{XY}. If F_{XU} satisfies Assumptions A and B (stated below), then θ_0 is identifiable.*

In what follows, we give Assumptions A and B, and show that Theorem 1 holds.

3.1. Identifiability of λ_0.

Hereafter, we shall assume the following:

Assumption A. *The probability distribution F_{XU} is such that: (i) $D_{U|X=0} \neq \emptyset$ and $D_{U|X=0} \neq \{0\}$; (ii) $\inf D_{U|X=0} = \underline{u}_0$ and $\sup D_{U|X=0} = \overline{u}_0$ where $\underline{u}_0 \in \mathbb{R} \cup \{-\infty\}$ and $\overline{u}_0 \in \mathbb{R} \cup \{+\infty\}$ are known to satisfy: $\underline{u}_0 \neq 0$ if $\overline{u}_0 = +\infty$ and $\overline{u}_0 \neq 0$ if $\underline{u}_0 = -\infty$.*

As previously, $D_{U|X=0}$ is the support of the conditional distribution of U given $X = 0$. Underlying the first property in item (i) is the condition that $0 \in \mathbb{R}^k$

is a possible realization of X; the second property on the other hand states that conditional on $X = 0$, U can take more than just a value 0. According to the item (ii), the boundaries of the support of U given $X = 0$ equal $\underline{u}_0$ and $\overline{u}_0$, with the two constants $\underline{u}_0$ and $\overline{u}_0$ being fixed and independent of F_{XU}, but not necessarily known. What is known about the couple $(\underline{u}_0, \overline{u}_0)$, however, is that it is not equal to either $(0, +\infty)$ or $(-\infty, 0)$. It is worth pointing out that Assumption A does not make any statements regarding the discreteness or the continuity of $D_{U|X=0}$, nor does it require the support of U given $X = 0$ to be known.

The nature of the restrictions in Assumption A is semiparametric: the distribution F_{XU} is allowed to remain unknown, provided the conditional distribution of U given $X = 0$ satisfies the conditions (i) and (ii). These conditions fix the boundaries of the support of $F_{U|X=0}$ rather than the mean of $F_{U|X}$; thus, they are different from the commonly employed mean independence conditions which fix $E(U|X)$ to be equal to zero. Note that both $\underline{u}_0$ and $\overline{u}_0$ are allowed to depend on the realization $x = 0$ of X. In particular, Assumption A(ii) does not restrict the boundaries of $D_{U|X=0}$ to be constant as x changes. Thus, unlike Foster, Tian, and Wei (2001), we do not require U and X to be independent. We now show that the restrictions in Assumption A are sufficient to identify λ_0.

For this, consider a structure $\mathcal{S}^* = (\theta_0^*, F_{XU}^*)$ with $\theta_0^* \equiv (\beta_0^{*\prime}, \lambda_0^*)'$ that generates the same probability distribution F_{XY} as $\mathcal{S}$—so Lemma 1 applies to $\mathcal{S}^*$—and satisfies the same conditions as $\mathcal{S}$—so the properties (i) and (ii) in Assumption A hold under $\mathcal{S}^*$. In particular, $\underline{u}_0$ is the common value of the inf of the supports $D_{U|X=0}$ and $D_{U|X=0}^*$ of the conditional distributions of U given $X = 0$ under F_{XU} and F_{XU}^*, respectively. Similarly, $\overline{u}_0$ is the common value of the sup of the two supports.

If $\mathcal{S}^*$ is observationally equivalent to $\mathcal{S}$, then by Equation (3) it must hold that: $\inf D_{Y|X=0} = (1 + \lambda_0 \underline{u}_0)^{1/\lambda_0} = (1 + \lambda_0^* \underline{u}_0)^{1/\lambda_0^*}$ and $\sup D_{Y|X=0} = (1 + \lambda_0 \overline{u}_0)^{1/\lambda_0} = (1 + \lambda_0^* \overline{u}_0)^{1/\lambda_0^*}$. We now show that these equations imply $\lambda_0 = \lambda_0^*$.

We first consider the case in which $\underline{u}_0 = -\infty$ and $\overline{u}_0 = +\infty$, or equivalently, $\inf D_{Y|X=0} = 0$ and $\sup D_{Y|X=0} = +\infty$. From our previous discussion, this is only

possible if $\lambda_0 = \lambda_0^* = 0$. In other words, if conditional on $X = 0$ the disturbance U has support with infinite boundaries on $\mathbb{R}$, or equivalently, if conditional on $X = 0$ the dependent variable Y has support with boundaries 0 and $+\infty$, then $\lambda_0 = 0$ is identified.

Now, consider the case in which at least one of $\underline{u}_0 > -\infty$ and $\overline{u}_0 < +\infty$ holds, so $\lambda_0 \neq 0$. This case is equivalent to the one in which at least one of $\inf D_{Y|X=0} > 0$ and $\sup D_{Y|X=0} < +\infty$ holds. Without loss of generality, assume that $\underline{u}_0 > -\infty$ with $\underline{u}_0 \neq 0$, so $\inf D_{Y|X=0} = (1 + \lambda_0 \underline{u}_0)^{1/\lambda_0} > 0$ with $\inf D_{Y|X=0} \neq 1$. A necessary and sufficient condition for λ_0 to be identified by the conditions in Assumption A is that $\lambda = \lambda_0$ be the unique solution to the equation:

$$(1 + \lambda \underline{u}_0)^{1/\lambda} = (1 + \lambda_0 \underline{u}_0)^{1/\lambda_0} \tag{4}$$

Consider then the mapping $g : D_0 \to (0, +\infty)$ defined on $D_0 \equiv \{\lambda \in \mathbb{R} : 1 + \lambda \underline{u}_0 > 0\}$, and which to each $\lambda \in D_0$ associates $g(\lambda) \equiv (1 + \lambda \underline{u}_0)^{1/\lambda}$. Recall that $\underline{u}_0 \neq 0$ so g is not a constant map on D_0. For any $\lambda \in D_0 \backslash \{0\}$ we have $g'(\lambda) = -\frac{1}{\lambda^2}\big[\ln(1 + \underline{u}_0\lambda) - \frac{\underline{u}_0\lambda}{1+\underline{u}_0\lambda}\big] g(\lambda)$ and $\lim_{\lambda \to 0} g'(\lambda) = -\frac{\underline{u}_0^2}{2} \exp(\underline{u}_0) = g'(0)$, so the map g is continuously differentiable on D_0. Now, note that for any $x > 0$ it holds that $\ln(x) - \frac{x-1}{x} \geqslant 0$ with equality only if $x = 1$. Hence, for any $\lambda \in D_0 \backslash \{0\}$ we have $g'(\lambda) < 0$. Moreover, $g'(0) < 0$, so g is strictly decreasing on D_0. It then follows that Equation (4) has a unique solution $\lambda = \lambda_0$.

In other words, if given $X = 0$, the support of U is bounded on $\mathbb{R}$ either from below or above or both, and if at least one of the (finite) boundaries is different from 0, then the domain conditions in Assumption A identify any $\lambda_0 \neq 0$.

3.2. **Identifiability of β_0.** It remains to derive sufficient conditions for the regression parameter β_0 in Equation (1) to be identified as well. For this, we shall assume that—in addition to Assumption A—we have the following:

Assumption B. *The probability distribution F_{XU} is such that:*

$$\mathrm{Cov}(U, X) = 0 \text{ and } \det \mathrm{Var}(X) \neq 0$$

When the domain condition in Equation (1) holds, the moment conditions in Assumption B induce the following moment restrictions on the probability distribution F_{XY} of the observables X and Y:

$$\text{Cov}\left(\frac{Y^{\lambda_0}-1}{\lambda_0}, X\right) - \text{Var}(X)\beta_0 = 0 \text{ and } \det \text{Var}(X) \neq 0 \tag{5}$$

The expectation in Equation (5) is taken with respect to F_{XY} obtained under the true parameter value θ_0. If the structure $\mathcal{S}^* = (\theta_0^*, F_{XU}^*)$ is observationally equivalent to $\mathcal{S}$ then $\theta_0^* = (\beta_0^{*\prime}, \lambda_0^*)'$ satisfies the same set of moment restrictions. Since under Assumption A we have $\lambda_0 = \lambda_0^*$, and given that the moment conditions in Equation (5) are linear (and non-constant) in β_0, it follows that necessarily $\beta_0^* = \beta_0$.

To resume, under Assumption A, any true value λ_0 of the exponent parameter λ in the Box-Cox model in Equation (1) is identifiable. The conditions used for identification restrict the boundaries of the support of $F_{U|X=0}$ rather than the mean of $F_{U|X}$. Once λ_0 identified, the usual orthogonality condition $\text{Cov}(U, X) = 0$ between the disturbance U and the explanatory variable X and the full rank condition $\det \text{Var}(X) \neq 0$ on the regressor given in Assumption B are sufficient to identify any $\beta_0 \in \mathbb{R}^k$.

References

Amemiya, T., and J. L. Powell (1981): "A Comparison of the Box-Cox Maximum Likelihood Estimator and the Non-Linear Two-Stage Least-Suqares Estimator," *Journal of Econometrics*, 17, 351–381.

Box, G. E. P., and D. R. Cox (1964): "An Analysis of Transformations," *Journal of the Royal Statistical Society. Series B (Methodological)*, 26, 211–252.

Fisher, F. M. (1961): "Identifiability Criteria in Nonlinear Systems," *Econometrica*, 29, 574–590.

——— (1965): "Identifiability Criteria in Nonlinear Systems: A Further Note," *Econometrica*, 33, 197–205.

——— (1966): *The Identification Problem in Economics*. McGraw-Hill, New York.

Foster, A. M., L. Tian, and J. L. Wei (2001): "Estimation of the Box-Cox Transformation Model Without Assuming Parametric Error Distribution," *Journal of the American Statistical Association*, 96, 1097–1101.

Khazzoom, D. J. (1989): "A Note on the Application of the Nonlinear Two-Stage Least-Squares Estimator to a Box-Cox Transformed Model," *Journal of Econometrics*, 42, 377–379.

Komunjer, I. (2008): "Global Identification in Nonlinear Semiparametric Models," Mimeo: University of California, San Diego.

Koopmans, T. C. (ed.) (1950): *Statistical Inference in Dynamic Economic Models*, vol. 10 of *Cowles Commission Monograph*. John Wiley & Sons, Inc.

Koopmans, T. C., and O. Reiersøl (1950): "The Identification of Structural Characteristics," *The Annals of Mathematical Statistics*, 21, 165–181.

Powell, J. L. (1996): "Rescaled Method-of-Moments Estimation for the Box-Cox Regression Model," *Economics Letters*, 51, 259–265.

Roehrig, C. S. (1988): "Conditions for Identification in Nonparametric and Parametric Models," *Econometrica*, 56, 433–447.

Rothenberg, T. J. (1971): "Identification in Parametric Models," *Econometrica*, 39, 577–591.

Savin, N. E., and A. H. Würtz (2005): "Testing the Semiparametric Box-Cox Model with the Bootstrap," in *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, ed. by D. W. K. Andrews, and J. H. Stock, pp. 322–354.